Computers in Biology and Medicine 106 (2019) 97–105

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Medical expenditure estimation by Bayesian network for lung cancer patients at different severity stages

Kung-Jeng Wang[a,*], Jyun-Lin Chen[a], Kung-Min Wang[b]

[a] Department of Industrial Management, National Taiwan University of Science and Technology, No.43, Sec. 4, Keelung Rd., Da'an Dist., Taipei, 106, Taiwan, ROC
[b] Department of Surgery, Shin-Kong Wu Ho-Su Memorial Hospital, Shilin District, Taipei, 111, Taiwan, ROC

ARTICLE INFO

Keywords:
Conditional Gaussian Bayesian network
Lung cancer
Medical expenditure
Medical information analytics

ABSTRACT

Lung cancer is one of the leading causes of mortality, and its medical expenditure has increased dramatically. Estimating the expenditure for this disease has become an urgent concern of the supporting families, medial institutes, and government. In this study, a conditional Gaussian Bayesian network (CGBN) model was developed to incorporate the comprehensive risk factors to estimate the medical expenditure of a lung cancer patient at different stages. A total of 961 patients were surveyed by the four severity stages of lung cancer. The proposed CGBN model identified the correlation and association of 15 risk factors to the medical expenditure of different severity stages of lung cancer patients. The relationships among the demographic, diagnosed-based, and prior-utilization variables are constructed. The model predicted the lung cancer-related medical expenditure with high accuracy of 32.63%, 50.30%, 50.36%, and 66.58%, respectively for stages 1–4, as compared with the reported models. A greedy search was also applied to find the upper threshold of R^2, while our model also shows that it approached the upper threshold.

1. Introduction

Lung cancer is the leading cause of cancer deaths in the world [1]. This malignancy is responsible for nearly one out of five cancer deaths. In Taiwan, lung cancer is the leading cause of mortality, accounting for 8854 cases or 19.8% of all cancer deaths per year [2]. Moreover, compared with other leading death cancers (e.g., breast, cervical, liver, and colorectal), lung cancer has the highest percentage of patients with survival of less than 1 year [3].

Lung cancer contributes to a significant medical expenditure. The expenditure has also considerably increased over time because of the evolution of cancer treatment [4]. The mean expenditure for lung cancer in the initial phase of treatment has increased by 93.4% from 1996 to 2007 [5]. This issue has largely raised concerns by the supporting families, medial institutes, and government [6,7]. Therefore, the investigation of the financial burden has been placed on a high priority.

Taiwan has established a National Health Insurance (NHI) system since 1995, and the NHI covers more than 99.9% of the population. According to the NHI records, critically ill patients only account for 3.9% of the population, but they consume 27.6% of the overall medical expenditure [8]. The estimation and control of the medical expenditure has become an urgent problem for the government. The situation is worse for those diagnosed with lung cancer. Moreover, the medical resource planning for lung cancer patients can be enhanced if the medical expenditure is properly estimated. However, the comprehensive factors for the medical expenditure of lung cancer patients are difficult to explore because of the uncertainty of this malignancy and the complex interaction among those factors. A systematic way to evaluate medical expenditure and form an executable estimation model is needed.

A risk adjustment model is promising in modeling medical causality. This model is composed of the risk factors, and a precise function of these risk factors can increase the predictive power of medical expenditure [9,10]. Exploring the risk factors to increase the predictability is a critical step in implementing the risk adjustment model. Further, previous studies on risk adjustment employed basic statistical methods to predict personal medical expenditure and survivability (e.g., Refs. [11,12]. However, such approach is not always committed to the general situation that deals with missing and uncertainty data. Moreover, nonlinear property has a sustainable and important impact on the prognosis of expenditure and survivability as well as potential medical decision support applications. Therefore, a more effective predictive modeling formalism to handle domains with mixed discrete and continuous variables, nonlinear characters among risk factors, as well as with missing and uncertainty data, is required.

[*] Corresponding author.
E-mail addresses: kjwang@mail.ntust.edu.tw (K.-J. Wang), jlchen.ian0411@gmail.com (J.-L. Chen), albert.hua@msa.hinet.net (K.-M. Wang).

https://doi.org/10.1016/j.compbiomed.2019.01.015
Received 21 July 2018; Received in revised form 7 January 2019; Accepted 19 January 2019

There are two main types of lung cancer: non-small cell lung cancer and small cell lung cancer. Each has its own system of staging, a process that determines the extent to which a cancer has spread. Non-small cell lung cancer has four major stages. In stage 1, cancer is found in the lung, but it has not spread outside the lung. In stage 2, the cancer exists both in the lung and nearby lymph nodes. The cancer spreads in the lung and lymph nodes in the middle of the chest is defined as stage 3. In stage 4, the cancer has spread to both lungs, into the area around the lungs, or to distant organs. Moreover, small-cell lung cancer has two major stages. In the limited stage, cancer is found in only one lung or nearby lymph nodes on the same side of the chest. In the extensive stage, the cancer has spread throughout one lung, to the opposite lung, lymph nodes on the opposite side, fluid around the lung, bone marrow, and/or distant organs. Unfortunately, few studies have categorized the severity of lung cancer to extend their specification and utilization in terms of target treatment plans. Lung cancer stages indicate how severe the cancer has spread and help determine treatment and medicine. Based on the cancer stage, a treatment plan can be developed in accordance. Treatment may include a combination of methods, depending on the stage and type of lung cancer. Precise models should be explored to estimate the medical expenditure for lung cancer patients at different stages.

This study was performed to identify the lung cancer risk factors for medical expenditure and develop an estimation model to predict medical expenditure against data uncertainty by Bayesian network (BN). The result could provide valuable information for further studies on cancer expenditure estimation and healthcare insurance management. The risk factors that influence the lung cancer medical expenditure were identified. The lung cancer patients were classified according to the different stages of severity, and the medical expenditure for these patients was estimated.

2. Literature review

2.1. Risk adjustment factors for medical expenditure

Risk adjustment is a technique that can be applied to estimate insurance costs [9,13,14]. Different combinations of risk factors will lead to distinct prediction accuracies. Risk factors can be classified into six groups, namely, demographic, prior utilization, diagnostic-based, prescription, physiological risk factors, and self-reported [13,15]. The demographic risk factors include demographic data, such as age, gender, education, and residential area. However, these factors are weak predictors of individual medical expenditure. Prior utilization factors use personal past medical utilization information to predict future medical expenditure. Several common factors in this group are past medical expenditure, hospitalized record, and frequency of seeking medical service. Diagnosis-based risk factors, such as outpatient and inpatient data, indicate the use of diagnosis record in a specific period to predict future medical expenditure. Based on valid International Classification of Diseases (ICD-9) codes, patients could be classified into different groups. These codes would provide more detailed information on the personal pathogen. Prescription factor can be used to evaluate specific patients for their medical expenditure based on the differences in the treatments and drugs. Physiological factors are used to evaluate the health status of individuals and measure the risk to future medical demand, especially for chronic diseases. Personal health status mainly includes blood pressure, cholesterol, and blood sugar. This type of factor requires examination and private medical record, but the data are hard to retrieve. Meanwhile, self-reported measures include perceived health status and instrumental activities of daily living to evaluate functional health status of patients and the needed medical expenditure. Questionnaires and interviews are necessary to obtain the related information to understand the self-health, functional status, and medical utilization tendency.

Previous studies on risk adjustment are commonly based on the six factors to construct their models. However, various combinations of risk factors will lead to diverse results for estimating medical expenditure. Kapur et al. [16] investigated five risk-adjustment models to estimate the future medical expenditure for high-cost mental-health service utilizers. Their models included demographic variables, such as age, gender, ethnicity, education level, marital status, residence, living arrangement, and language, insurance indicator, homelessness, indicators for diagnoses, and inpatient and outpatient costs.

Pope et al. [17] developed the Principal Inpatient Diagnostic Cost Group Model (PIPDCG), which included age, sex, disabled status, and inpatient diagnoses derived from prior year inpatient hospital records. Lin et al. [18] developed a Taiwan version of the PIPDCG (TPIPDCGs) to predict an individual's medical expenditure. The TPIPDCG is a revision of the PIPDCG from the Health Care Financing Administration in the US. They built a weighted least squares regression model and used predictive R2 to evaluate the performance of the risk adjustment model. They compared eight different models with various risk factors. Chang and Lai [19] employed three kinds of risk adjustment factors, namely, demographic, diagnosis-based, and prior utilization. Demographic factors included age and gender. Diagnosis-based factors included inpatient and outpatient information. Prior utilization factors included individual's inpatient and outpatient expenditures.

In addition, Omachi et al. [20] verified that measuring the severity of an illness can increase the predictive power of a risk adjustment model. They used the common risk factors, such as age, gender, and diagnosis codes, and further considered the severity of chronic obstructive pulmonary disease.

Previous studies and their factors and R^2 measures are summarized in Table 1.

2.2. Lung cancer risk factor

Age is an important factor to medical expenditure, because the elderly may have higher demands to seek medical service than others [21,22]. Epidemiology studies have reported that gender and environmental factors are risk factors that increase a person's probability of developing lung cancer [23,24]. The lung cancer site and comorbidity also influence medical expenditure. The frequency of seeking medical

Table 1
Summary of related risk adjustment researches and their performance.

Researches	Factors	Best R^2
Kapur et al. [16]	age, gender, marital status, ethnicity, education level, type of residence, and language, insurance indicators, homelessness, indicators for diagnoses, global assessment of functioning scores, inpatient and outpatient costs for the previous year, inpatient and outpatient costs for the two previous years.	16%
Pope et al. [17]	age, sex, disabled status, and inpatient diagnoses, PIPDCG.	6.2%
Lin et al. [18]	age, gender, TPIPDCG and prior outpatient	36.5%
Chang & Lai [19]	age, gender, TPIPDCGs, BTL, inpatient and outpatient expenditures.	35.2%
Omachi et al. [20]	age, gender, diagnosis codes, severity of COPD.	21%

Note: R^2 statistic is used as a measure of variability, ranging from 0 to 1 (0 means no correlation and 1 indicates perfect correlation). It is used to evaluate the explanatory power of risk factor to medical expenditure.

Table 2
Defined treatment for stage classification.

Surgery	Chemotherapy		
	Injection	Drugs	Targeted therapy
● Segmental resection of lung ● Lobectomy of lung ● Complete pneumonectomy	Injection or infusion of cancer chemotherapeutic substance	● Docetaxel ● Doxorubicin ● Etoposide ● Pemetrexed ● Vinorelbine ● Gemcitabine ● Paclitaxel	● Erlotinib ● Gefitinib ● Afatinib ● Bevacizumab

service and hospitalization, as well as the duration of hospitalization, will directly affect the expenditure [19]. These factors have also been used in the literature (e.g., Refs. [18,25,26].

A study has indicated that the risk of getting lung cancer increased with age and was greater in men than in women, and especially for people who were greater than 60 years old [27]. A high risk of lung cancer was reported in populations exposed to high levels of air pollution. Therefore, the environmental factors, such as residential area, should be considered when choosing the risk factor [23,24]. Wang et al. [28] conducted a population-based study in Taiwan. The factors used to evaluate survivability included age, sex, tumor location, cell type, histologic grading, surgical resection methods, clinical stage, treatment modality, survival time, and cause of death. The results indicated that survival rate had significant differences between stages and tumor locations, and women with lung cancer had higher prognosis than men.

From the literature survey, we determined that age, gender, tumor location, and comorbidities are potentially important factors for the medical expenditure for lung cancer. Different severity stages of lung cancer may also present distinct characteristics. These factors were further investigated to evaluate the medical expenditure for lung cancer patients.

2.3. Bayesian network

The BN performs probabilistic inference of relationships among variables. This process is beneficial to elucidate the causal relationship in a complex structure. The BN can handle missing data and uncertain knowledge [29,30]. The applications of BN include diagnoses of diseases, optimal treatment alternatives, and prediction of treatment outcomes [31–33].

Although the variables in a BN are usually assumed to be discrete, the variables in the real world are often continuous [34,35]. Studies have usually discretized a continuous variable into categorical type [36], while discretization may lead to bias in prediction. Lauritzen and Wermuth [37] presented a model that can simultaneously handle discrete and continuous variables. The model known as conditionally Gaussian distributions (CG-distributions) assumed that a continuous variable should follow Gaussian distribution. The CG-distributions have been extended by Cobb et al. [34] and Lauritzen & Jensen [38]. In the CG-distributions, discrete variables are analyzed similarly to the traditional BN, while the probability of continuous variables is determined by parent nodes.

3. Materials and model

3.1. Data source and study population

A population-based retrospective cohort and the corresponding expenditure of illness and medical record were all derived from the NHI research database. The study population based on the lung cancer patients were identified from 1996 to 2010 Ambulatory Care Expenditures by Visits (CD) file and Inpatient Expenditures by Admissions (DD) file; on the other hand, this study also considered Details of Ambulatory Care Orders (OO) file to define the treatment type, then used Registry for Contracted Medical Facilities (HOSB) file and Registry for Beneficiaries (ID) file to obtain detailed medical information. Those data files are all in the Original Claim Data for Reimbursement (OCDR) dataset.

The diseases were investigated and distinguished by their ICD-9-CM diagnostic code. The ambulatory visit file provided up to three ICD-9-CM diagnostic codes which allow the diagnosed disease for each visit to be identified. Further, the CD and DD files both provided medial expenditure for each hospitalization and outpatient visit, including various diagnostic procedures and established treatment or other medical services. Based on those medical expenditure information, the total expenditure can be calculated.

The selection procedure of the study population in this study is chronologically described as follows. First, we used ICD-9-CM code which starts with 162.x to identify cohort representing malignant neoplasm of trachea bronchus and lung from 1996 to 2010 outpatient and inpatient files of the OCDR dataset. The cohort is regarded as the patients diagnosed with lung cancer.

Since there is no severity stage information of lung cancer patients in the NHI research database, we needed to confirm it by the treatment they had taken. Different stages of lung cancer patients should take different treatments [39]. By discussing with 2 domain experts in thoracic surgery and oncology departments of medical-center level hospitals in Taipei, we summarize the types of surgery and chemotherapy as Table 2. The types of surgery are segmental resection of lung, lobectomy of lung, and complete pneumonectomy. For lung cancer, the purpose of operating on surgery is the full removal of the lung tumor and the nearby lymph nodes in the chest. The tumor must be removed with a surrounding border of normal lung tissue to make sure no cancer was found in the healthy tissue surrounding the tumor. The type of chemotherapy can be classified as injection, drugs, and targeted therapy. Chemotherapy is to use drugs to destroy cancer cells, stopping cancer cells to grow [40].

Accordingly, the four stages of lung cancer patients were as the follows:

Stage I: Stages I and II lung cancer patients are generally treated with surgery. Patients who only received surgery and no chemotherapy were classified as stage I.

Stage II: Patients in this stage usually receive both surgery and chemotherapy. Several patients with large tumors were treated by surgery before chemotherapy to increase the benefit of treatment.

Stage III: Patients in this stage suffer by cancer metastasis to other organs. Thus, surgery alone is usually not sufficient to cure the disease for most patients. Stage III patients have a high risk of cancer recrudescence, either in the same place or distantly, even after successful surgery. Thus, doctors generally do not recommend immediate surgery. Instead, chemotherapy before surgery is suggested. Hence, patients who received chemotherapy before surgery were classified to stage III.

Stage IV: Patients in stage IV do not receive surgery on the whole. Most patients at this stage receive only chemotherapy. However, stage IV patients are not considered “cured.” Chemotherapy can only be continued as long as it can control the growth of cancer. If the cancer worsens or causes too many severe side effects, the treatment will be stopped, and patients would continue to receive palliative care. In this study, stage IV included patients who only received defined chemotherapy and who did not receive any treatment shown in Table 2 but died in 4 months since the day they were diagnosed with lung cancer.

3.2. Variables

Four comorbidities were chosen as variables. According to the Ministry of Health and Welfare, 100 diseases were defined as chronic and separated into different biological systems [41]. In this study, 22 chronic diseases, which are common for Taiwan's population, were chosen after discussion with domain experts. Eventually, four key chronic diseases, namely, hypertension, bronchiectasis, emphysema, and tuberculosis, were selected on the basis of odds ratio.

The lung cancer site indicates the location with lung cancer. This site will influence the treatment type and has been confirmed by the American Cancer Society [42]. The relationship was also adopted by Wang et al. [43]. The lung cancer sites were classified into eight types based on ICD-9-CM. These categories included bronchus and lung (162), trachea (1620), bronchus (1622), upper lobe, bronchus or lung (1623), middle lobe, bronchus or lung (1624), lower lobe, bronchus or lung (1625), other parts of bronchus or lung (1628), bronchus and lung, and unspecified (1629).

To highlight the importance of a medical resource, the Taiwan's NHI Administration, Ministry of Health and Welfare in Taiwan, classified the hospital types into 4 levels, and each level has different charges or fees. These hospital levels are district hospitals, regional hospitals, physician clinics, and medical centers. People who seek medical service at different hospital levels will be charged distinctly. This service mainly influences the outpatient visit expenditure [8]. Thus, the hospital level will affect the medical expenditure.

The geographical region is an important factor of lung cancer, and this factor has been confirmed by Samet et al. [24]. Environmental impacts, e.g., tobacco smoking, alcoholic beverages, dietary intake of cholesterol and/or fat, radiation, asbestos exposure, air pollution, and hazard occupational exposure, are also correlated to the development of a city as well as lung cancer occurrences. Six hospital regions exist in Taiwan. These regions included the central, northern, Taipei, eastern, southern, and Kaohsiung branches.

In addition, 7 kinds of urbanization, including general township, moderate urbanization, highly urbanization, aging town, remote town, new town, and agricultural town, have been identified. The definition of urbanization is not included in the NHI databank. The criteria of classifying different areas to the seven types of urbanization were demographic characteristics, industrialization, and medical resource distribution. More details on the township and downtownship separated into urbanization can be found in Liu et al. [44].

Accordingly, in the present study, the 15 variables used to predict the medical expenditure are listed in Table 3. The factors were chosen based on the literature of risk adjustment factors. Age, gender, level, region, and urbanization could be classified as demographic factors. Site, hypertension, bronchiectasis, emphysema, and tuberculosis could be classified as diagnosed-based factors [45,46]. Outpatient visit, outpatient expenditure, inpatient visit, inpatient days, and inpatient expenditure could be classified as prior utilization factors. Moreover, the treatment defined in Table 2 can be used as prescription factors. The physiological and self-reported factors were not used in this study, because these two factors were not included in NHI databank. Finally, the risk factors used in this study were demographic, diagnosed-based, prior utilization, and prescription factors.

For the prediction target, expenditure is the future unit average cost of individuals. Except for the future expenditure, the other factors were all based on the information of the year when the patient was diagnosed with lung cancer.

A list of key variables to build a BN for expenditure prediction is depicted in Table 3.

3.3. Modeling

Assuming a hybrid cause-effect model with discrete and continuous variables, the set of variables X can be divided into two, $X = Y \cup Z$, where Y represents the discrete variable, and Z presents the continuous variable. $|Y| = d$, and $|Z| = c$ are the numbers of continuous and discrete variables, respectively. The joint state space can be denoted as $x = (y, z) = (y_1, ..., y_d, z_1, ..., z_c)$, where $y_1, ..., y_d$ are qualitative data, and $z_1, ..., z_c$ are numeric numbers. The BN graph defines the independent relations among the variables and presents a set of joint probability distributions of the variables in the topology as the combination of the conditional distribution of each variable given its parents in the network. Supposing a continuous variable X, with discrete parent Y and continuous parent Z, the Gaussian distribution conditional on the values of the parents can be $p(X|Y = y, Z = z) = N[\alpha(y) + \beta(y)^T z, \ \gamma(y))]$, where N represents the normal distribution, $(y, \ z)$ are the states of the parents, $\gamma(y) > 0$, $\alpha(y)$ is a real number, and $\beta(y)$ is a vector of the same size as the continuous part of the parents. For each configuration of the discrete parents, a different linear function of the continuous parents is defined as the normal distribution.

After the risk factors for medical expenditure of lung cancer have been delineated, a relationship among these factors should be determined. This study constructed a conditional Gaussian BN (CGBN) to estimate the medical expenditure of lung cancer patients. A model topology is needed to create an appropriate CGBN. Such topology in the study was derived from the literature and reviewed by domain experts. The continuous variables in our conditional CGBN followed the normal distribution, while those variables that did not pass the Kolmogorov-

Table 3
Variables to build a Bayesian network for expenditure prediction.

Factor	Class description	Data type	Risk adjustment type
Age	The age when patient be diagnosed with lung cancer	Continuous	demographic
Gender	Male or Female	Binary	demographic
Level	4 kind of hospital level, including District Hospitals, Regional Hospitals, Physician Clinics, and Medical Centers	Discrete	demographic
Region	6 kind of hospital region, including Central branch (CB), Northern branch (NB), Taipei branch (TB), Eastern branch (EB), Southern branch (SB), and Kaohsiung branch (KB)	Discrete	demographic
Urban	7 kind of urbanization, including General Township (GT), Moderate urbanization (MU), Highly urbanization (HU), Aging town (AGT), Remote town (RT), New town (NT), Agricultural town (AT)	Discrete	demographic
Site	8 type of location of lung cancer, including trachea, bronchus and lung (162), trachea (1620), bronchus (1622), upper lobe, bronchus or lung (1623), middle lobe, bronchus or lung (1624), lower lobe, bronchus or lung (1625), other parts of bronchus or lung (1628), bronchus and lung, unspecified (1629)	Discrete	diagnosed-based
Hypertension	A Comorbidity of hypertension which diagnosed before lung cancer or not	Binary	diagnosed-based
Bronchiectasis	A Comorbidity of bronchiectasis which diagnosed before lung cancer or not	Binary	diagnosed-based
Emphysema	A Comorbidity of emphysema which diagnosed before lung cancer or not	Binary	diagnosed-based
Tuberculosis	A Comorbidity of tuberculosis which diagnosed before lung cancer or not	Binary	diagnosed-based
Outpatient visits	Total frequencies of outpatient visits of the patient in the year when lung cancer is diagnosed	Continuous	prior-utilization
Outpatient expenditure	Total cost for outpatient of the patient in the year when lung cancer is diagnosed	Continuous	prior-utilization
Inpatient visits	Total frequencies of hospitalization of the patient in the year when lung cancer is diagnosed	Continuous	prior-utilization
Inpatient days	Total days of hospitalization of the patient in the year when lung cancer is diagnosed	Continuous	prior-utilization
Inpatient expenditure	Total cost of inpatient for individual patient in the year when lung cancer is diagnosed	Continuous	prior-utilization
Medical expenditure estimation	Average annual cost of the patient future traceability years. By using parent nodes in the year when the patient is diagnosed with lung cancer to predict future average annual cost.	Continuous	

Table 4
Theoretical links to build CGBN topology.

Links	Reference
Region→Urban, Urban→Site	Samet et al. [24]
Gender→Site	Ko et al. [23]; Samet et al. [24]
Hypertension→Site, Bronchiectasis→Site, Emphysema→Site, Tuberculosis→Site	Domain doctor; odds ratio calculated in this study
Site→Inpatient visits	Cucciare & O'Donohue [46]; Corral et al. [45]
Age→Outpatient visit	Fuch [22]; Cheng [21],
Level→Outpatient expenditure	Ministry of Health and Welfare [41]
Outpatient visits→Outpatient expenditure	Stearns et al. [26]; Pannarunothai & Phanthunane [25]
Inpatient visits→Inpatient expenditure	
Inpatient days→Inpatient expenditure	
Outpatient expenditure→future expenditure	Chang & Lai [19]; Lin et al. [18]
Inpatient expenditure→future expenditure	

Smirnov test of normality were considered a natural logarithm.

The relationships among the demographic, diagnosed-based, and prior-utilization variables are constructed. Finally, outpatient and inpatient expenditures were used to predict medical expenditure. All links among factors followed evidence from the literature and domain experts as summarized in Table 4. The proposed CGBN composed of 15 variables, which represented the influence of risk factor to expenditure. The CGBN for expenditure prediction is depicted in Fig. 1.

Previous studies did not consider the severity of lung cancer patients, and many patients are diagnosed in their terminal stage. The study population was classified into four stages which defined the severity of lung cancer patients. Notably, the same model topology was applied to all stages, but the data profiles were different.

The conditional probabilities could explain the changes of each variable affected by evidence. The conditional probability between conjunctive nodes of the CGBN was estimated using our database. The resulting conditional relationship from the four stages listed as Table 5.

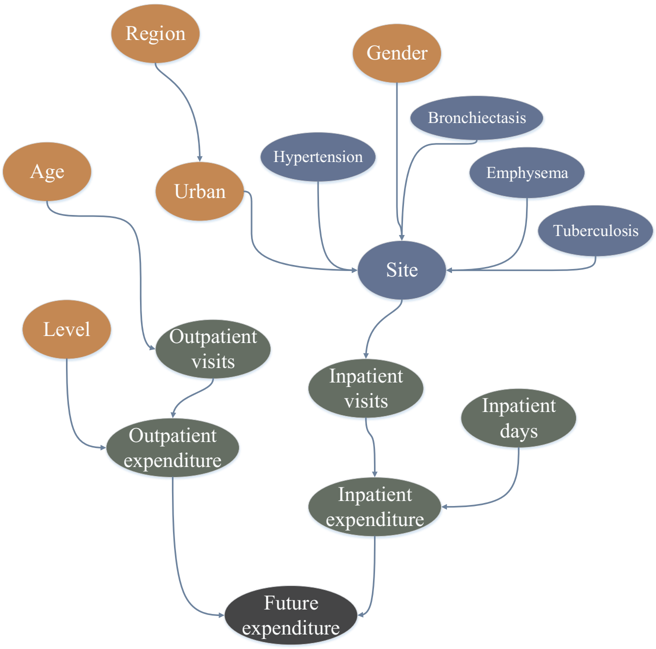

Fig. 1. The proposed CGBN model to estimate medical expenditure of lung cancer.

4. Results and discussions

In this section, a demographic data profile is presented. Four key chronic diseases, namely, hypertension, bronchiectasis, emphysema, and tuberculosis, were identified by high odds ratio. The Bayesian network estimation on medical expenditure with distinct severity stages is conducted and verified by a 10-fold validation to demonstrate a high performance of the proposed model. We illustrate and conclude the medical expenditure estimate for each severity stage of lung cancer that can facilitate an accurate estimation for medical expenditure.

4.1. Data profile of the key variables by severity stage

The data profile of the key variables by severity stage is presented in Table 6. A total of 961 lung cancer patients were included in the study for expenditure prediction modeling. The male has 57.5%. The sit at bronchus and lung, and unspecified (1629) accounts for 75.3%. The proportion of lung cancer patients at severity stages from I to IV is 28.2%, 12.8%, 2.2%, and 56.8%, respectively. In terms of outpatient expenditure, the patient at stage IV ranks the highest, whereas the patient at stage III has the highest inpatient expenditure. We further calculated the odds ratio of each chronic disease to the lung cancer site. Four key chronic diseases, namely, hypertension, bronchiectasis, emphysema, and tuberculosis, were selected owing to their high odds ratio. The details refer to Appendix. The statistics in Table 6 also indicates that hypertension is highly related to lung cancer.

4.2. Bayesian network estimation on medical expenditure

The R tool with two packages “deal” and “bnlearn” was employed to construct the proposed CGBN and for model learning [47–49]. The performance evaluation criterion was adjusted R^2 to evaluate the explanatory power of the risk factor to medical expenditure estimation [46,50].

Cross-validation is a technique to evaluate the stability and accuracy of predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. A 10-fold validation was performed in the study. The original sample is randomly partitioned into 10 equal size subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing our model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times, with each of the 10 subsamples used once as the validation data. The 10 results from the folds are averaged to produce a single estimation. The advantage of doing so is that all observations are used for both training and validation, and each observation is used for validation exactly once. The results are recorded in Table 7. Generally, a prospective risk model should achieve 20% of R^2, and the standard increases to 30% by the evolution of computation approaches [51,52]. Compared with previous studies, the proposed CGBN model had high performance, especially in stage IV. A greedy search was also applied to find the structure to gain the highest R^2, which was constructed arbitrarily for expenditure estimation. Such topology assumed that all types of links among variables were applicable. Our model shows that it approached the upper threshold of the adjusted R^2. The adjusted-R2 by arbitrary CGNB built using a greedy search showed a ceiling of 34.96%, 54.92%, 59.23%, and 67.87% for stages 1–4, while the proposed CGBN model approached to 32.63%, 50.30%, 50.36%, and 66.58%, respectively. Note that the existing best adjusted-R2 ranges from 6.2 to 36.5% [16–20].

4.3. Inference and findings

This study concludes the medical expenditure estimate for each severity stage of lung cancer. The medical expenditure formulas can facilitate an accurate estimation for medical expenditure. The statistical distribution of medical expenditure for each stage is shown in Table 8.

Table 5
Theoretical links to build CGBN topology.

Stage I	Stage II
P(Outpatient visit $\mid$ age) = $LogN(1.654 + 0.0067(age),\ 1.0548^2)$P(Outpatient expenditure $\mid$ outpatient visit (OV), level)= $LogN(6.5323 + 1.4472(ln(OV)),\ 0.7725^2)$, when level = District Hospitals P(Outpatient expenditure $\mid$ outpatient visit (OV), level)= $LogN(6.1396 + 1.6377(ln(OV)),\ 1.4594^2)$, when level = Regional Hospitals P(Outpatient expenditure $\mid$ outpatient visit (OV), level)= $LogN(5.9242 + 1.9374(ln(OV)),\ 1.0599^2)$, when level = Physician Clinics P(Outpatient expenditure $\mid$ outpatient visit (OV), level)= $LogN(4.5687 + 2.2911(ln(OV)),\ 1.8047^2)$, when level = Medical Centers P(Inpatient expenditure $\mid$ inpatient visit (IV), inpatient days (ID))= $LogN(4.2541 - 0.8529(ln(IV)) + 2.8436(ln(ID)),\ 1.7083^2)$ P(Inpatient visits $\mid$ site) = $LogN(0.4889,\ 0.4271^2)$, when site = 162 P(Inpatient visits $\mid$ site) = $LogN(0.0990,\ 0.2620^2)$, when site = 1620 P(Inpatient visits $\mid$ site) = $LogN(0.1383,\ 0.3099^2)$, when site = 1623 P(Inpatient visits $\mid$ site) = $LogN(0.2270,\ 0.3850^2)$, when site = 1625 P(Inpatient visits $\mid$ site) = $LogN(0.3500,\ 0.3943^2)$, when site = 1628 P(Inpatient visits $\mid$ site) = $LogN(0.3493,\ 0.4519^2)$, when site = 1629	P(Outpatient visit $\mid$ age)= $LogN(2.6428 - 0.0056(age),\ 1.1549^2)$, P(Outpatient expenditure $\mid$ outpatient visit (OV), level) = $LogN(6.2730 + 1.6453(ln(OV)),\ 0.8030^2)$, when level = District Hospitals P(Outpatient expenditure $\mid$ outpatient visit (OV), level) = $LogN(6.9644 + 1.3670(ln(OV)),\ 0.9234^2)$, when level = Regional Hospitals P(Outpatient expenditure $\mid$ outpatient visit (OV), level) = $LogN(4.8845 + 2.2116(ln(OV)),\ 0.5528^2)$, when level = Physician Clinics P(Outpatient expenditure $\mid$ outpatient visit (OV), level) = $LogN(5.2566 + 1.9458(ln(OV)),\ 1.8030^2)$, when level = Medical Centers P(Inpatient expenditure $\mid$ inpatient visit (IV), inpatient days (ID)) = $N(-66979.94 + 60938.27(ln(IV)) + 82861.48(ln(ID)),\ 92953.47^2)$ P(Inpatient visits $\mid$ site)= $LogN(1.0397,\ 0.8948^2)$, when site = 162 P(Inpatient visits $\mid$ site)= $LogN(1.0594,\ 0.9396^2)$, when site = 1620 P(Inpatient visits $\mid$ site)= $LogN(0.9792,\ 0.6855^2)$, when site = 1623 P(Inpatient visits $\mid$ site)= $LogN(0.7226,\ 0.8811^2)$, when site = 1625 P(Inpatient visits $\mid$ site)= $LogN(1.1513,\ 0.9102^2)$, when site = 1628 P(Inpatient visits $\mid$ site)= $LogN(0.9882,\ 0.7708^2)$, when site = 1629
Stage III	**Stage IV**
P(Outpatient visit $\mid$ age)= $N(17.0340 - 0.0141(age),\ 11.5957^2)$ P(Outpatient expenditure $\mid$ outpatient visit (OV), level)= $N(-43838.52 + 7265(OV),\ 96437^2)$, when level = Regional Hospitals P(Outpatient expenditure $\mid$ outpatient visit (OV), level)= $N(15524.998 + 4749(OV),\ 49362.63^2)$, when level = Medical Centers P(Inpatient expenditure $\mid$ inpatient visit (IV), inpatient days (ID))= $N(32033.905 + 15165.316(IV) + 7523.074(ID),\ 66529.04^2)$P(Inpatient visits $\mid$ site) = $N(8,\ 1\text{ e}- 16^2)$, when site = 162 P(Inpatient visits $\mid$ site)= $N(8,\ 9.8995^2)$, when site = 1620 P(Inpatient visits $\mid$ site)= $N(6,\ 4.0445e - 16^2)$, when site = 1623 P(Inpatient visits $\mid$ site)= $N(3{,}1\text{ e}- 16^2)$, when site = 1625 P(Inpatient visits $\mid$ site)= $N(2,\ 1\text{ e}- 16^2)$, when site = 1628 P(Inpatient visits $\mid$ site)= $N(4.4286,\ 3.4130^2)$, when site = 1629	P(Outpatient visit $\mid$ age)= $LogN(3.0574 - 0.0145(age),\ 1.2685^2)$, P(Outpatient expenditure $\mid$ outpatient visit (OV), level) = $LogN(5.7180 + 1.9781(ln(OV)),\ 1.8864^2)$, when level = District Hospitals P(Outpatient expenditure $\mid$ outpatient visit (OV), level) = $LogN(5.5646 + 2.0471(ln(OV)),\ 1.8309^2)$, when level = Regional Hospitals P(Outpatient expenditure $\mid$ outpatient visit (OV), level) = $LogN(5.5637 + 2.0183(ln(OV)),\ 2.0204^2)$, when level = Physician Clinics P(Outpatient expenditure $\mid$ outpatient visit (OV), level) = $LogN(5.6260 + 1.9643(ln(OV)),\ 1.9107^2)$, when level = Medical Centers P(Inpatient expenditure $\mid$ inpatient visit (IV), inpatient days (ID)) = $LogN(3.1681 - 0.2584(ln(IV)) + 2.9367(ln(ID)),\ 2.1386^2)$ P(Inpatient visits $\mid$ site)= $LogN(0.6698,\ 0.9570^2)$, when site = 162 P(Inpatient visits $\mid$ site)= $LogN(0.7129,\ 0.8347^2)$, when site = 1620 P(Inpatient visits $\mid$ site)= $LogN(0.7645,\ 0.6857^2)$, when site = 1623 P(Inpatient visits $\mid$ site)= $LogN(0.6212,\ 0.7269^2)$, when site = 1624 P(Inpatient visits $\mid$ site)= $LogN(0.7067,\ 0.6862^2)$, when site = 1625 P(Inpatient visits $\mid$ site)= $LogN(0.3531,\ 0.6623^2)$, when site = 1628 P(Inpatient visits $\mid$ site)= $LogN(0.8163,\ 0.7912^2)$, when site = 1629

Note: demographic variables are in brown color; diagnosed-based variables are in blue color; and prior-utilization variables are in green color.

Table 6
Data profile of key variables by severity stages(a) Discrete variables.

Variable	Class	Stage I	Stage II	Stage III	Stage IV	Total	%
Gender	Female	130	66	5	207	408	42.5%
	Male	141	57	16	339	553	57.5%
Site	162	13	4	1	11	29	3.0%
	1620	7	3	2	20	32	3.3%
	1622	3	1	0	5	9	0.9%
	1623	28	9	2	35	74	7.7%
	1624	3	1	0	4	8	0.8%
	1625	14	4	1	24	43	4.5%
	1628	17	6	1	18	42	4.4%
	1629	186	95	14	429	724	75.3%
Level	Medical Centers	175	78	16	265	534	55.6%
	Regional Hospitals	65	28	5	189	287	29.9%
	District Hospitals	25	14	0	79	118	12.3%
	Physician Clinics	6	3	0	13	22	2.3%
Region	Taipei branch	115	49	7	190	361	37.6%
	Northern branch	24	18	6	74	122	12.7%
	Central branch	49	25	3	110	187	19.5%
	Southern branch	41	18	3	85	147	15.3%
	Kaohsiung branch	39	11	2	77	129	13.4%
	Eastern branch	3	2	0	10	15	1.6%
Urbanization	Highly urbanization	118	34	7	154	313	32.6%
	Moderate urbanization	128	69	10	288	495	51.5%
	General Township	21	12	3	61	97	10.1%
	Aging town	0	0	0	2	2	0.2%
	Remote town	2	4	0	20	26	2.7%
	New town	2	3	1	20	26	2.7%
	Agricultural town	0	1	0	1	2	0.2%
Hypertension	Yes	140	60	12	292	504	52.4%
	No	131	63	9	254	457	47.6%
Bronchiectasis	Yes	18	8	2	32	60	6.2%
	No	253	115	19	514	901	93.8%
Emphysema	Yes	13	4	0	25	42	4.4%
	No	258	119	21	521	919	95.6%
Tuberculosis	Yes	24	4	2	35	65	6.8%
	No	247	119	19	511	896	93.2%
Grand total		271	123	21	546	961	
%		28.2%	12.8%	2.2%	56.8%		

(b) Continuous variables

Variable		Stage I	Stage II	Stage III	Stage IV
Age	Median	63	61	61	67
	Mean	62.6	60.2	59.9	65.5
Outpatient visit	Median	9	13	16	10.5
	Mean	12.4	16.7	16.2	15.6
Outpatient expenditure	Median	15,632	29,159	56,575	44,630
	Mean	51,324	76,660	89,912	111,693
Inpatient visit	Median	1	3	4	2
	Mean	1.46	3.42	4.90	2.84
Inpatient days	Median	16	18	18	15
	Mean	17.89	21.19	23.48	20.41
Inpatient expenditure	Median	169,481	192,117	293,971	113,488
	Mean	183,698	218,885	283,029	158,992

Table 7
Prediction performance of medical expenditure (in NTD dollar).

	Stage I		Stage II		Stage III		Stage IV	
	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual
Mean	49,210	10,6730	200,052	318,354	256,028	342,680	310,378	444,665
Standard deviation	44,060	156,827	46,618	270,061	143,061	245,363	50,071	316,656
Adjusted-R^2 by the proposed CGBN model	32.63%		50.30%		50.36%		66.58%	
Existing best Adjusted-R^2	6.2–36.5% [16–20]							
Ceiling Adjusted-R^2 by arbitrary CGNB built using a greedy search	34.96%		54.92%		59.23%		67.87%	

1. $R^2 = \frac{SS_{reg}}{SS_{total}}$, where SS_{reg} represents the regression sum of squares, and SS_{total} represents the total sum of squares. Adjusted $R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$, where p is the total number of regressors in the model, and n is the sample size.
2. This study applied a greedy search to find the 'best' structure which is constructed arbitrarily for expenditure estimation.

Table 8
The distribution of future medical expenditure by stages.

Stage	Medical expenditure estimation in normal distribution
I	P(Expenditure estimation$\mid$ Outpatient expenditure(OE), Inpatient expenditure(IE))= $LogN(11.6061 + 0.0894\,(\ln(OE)) - 0.1635\,(ln(IE)),\ 1.367^2)$
II	P(Expenditure estimation $\mid$ Outpatient expenditure(OE), Inpatient expenditure(IE))= $N(370{,}232.8 - 3014.357(ln(OE)) - 0.031(IE),\ 278,776.7^2)$
III	P(Expenditure estimation $\mid$ Outpatient expenditure(OE), Inpatient expenditure(IE))= $N(516{,}583.6 - 0.1795(OE) - 0.4486(IE),\ 300,224.3^2)$
IV	P(Expenditure estimation $\mid$ Outpatient expenditure(OE), Inpatient expenditure(IE))= $LogN(11.9765 + 0.0526(ln(OE)) + 0.0131(ln(IE)),\ 1.0833^2)$

Given that the expenditure estimation of stages I and IV does not follow the normal distribution, the natural logarithm was considered as a log-normal distribution for this node to transform the expenditure to a normal distribution. Compared with existing modeling approaches, such as logistic regression, which suffer by predicting expenditure as a categorical variable, the proposed model could predict expenditure as a continuous variable to be more accurate.

To illustrate the use of the model, the outpatient and inpatient expenditures were both assumed to be 10,000 NTD. The medical expenditure estimation for each stage were obtained as follows: stage I: $log\ N(10.9231,\ 1.367^2)$; stage II: $N\ (369924.0475,\ 278776.7^2)$; stage III: $N\ (510303.184,\ 300224.3^2)$; and stage IV: $log\ N(12.5817,\ 1.0833^2)$.

5. Conclusion

A set of CGBN models was presented to evaluate the medical expenditure for lung cancer patients at different severity stages in this study. The proposed CGBNs could handle discrete and continuous variables. This process extended the ability of a regular BN and conventional statistical models, specifically for categorical response estimation. The proposed CGBN model provided an accurate model to measure the impact between multiple-layer influencing factors and their response in medical expenditure. The experimental results showed that our model outperformed previous models and approached the ceiling of adjusted R^2.

Adopting the proposed model will help the patient, hospital, and the government to predict the expenditure of a patient at a specific severity stage. Therefore, finance and resource planning for lung cancer treatment can be managed precisely.

This study has several limitations. The data in this study does not include clinical records, such as actual lung cancer severity. In addition, the data only included registration data from 1996 to 2010. The factors chosen were mainly based on literature review and domain expert recommendation, while other potentially factors, such as drugs, could be further surveyed. It is worthy to note that the high correlation and association identified in the study facilitate the estimate of the medical expenditure for lung cancer patients at different severity stages, but such association does not necessarily imply causation. The causation of those factors and corresponding responses require further investigation and clinical validation. In addition, a high standard deviation of the medical expenditure estimates varies widely within the stages (I to IV) of the lung cancer and could be further addressed for root causes. It indicates that more potential risk factors could be include the cost estimation model to improve its accuracy, such as government insurance policies, facial year effect, and treatment technology evolution etc.

Appendix. Odds ratio of chronic to lung cancer site

SYSTEM/Chronic	162	1620	1622	1623	1624	1625	1628	1629	Overall
ENDOCRINE									
Thyroid Dysfunction	0.69	1.35	1.27	1.61	1.19	1.55	1.16	1.23	1.26
Hyperlipidemia	2.24	2.81	2.38	2.98	3.29	3.07	2.74	2.77	2.75
Arthrolithiasis	1.87	2.44	3.09	2.60	2.69	2.42	2.62	2.43	2.36
Diabetes	1.91	2.73	3.78	3.59	3.04	2.98	2.77	2.98	2.90
NERVOUS									
Trigeminal Sensory Neuropathy	1.21	2.90	4.66	1.25	1.28	0.88	1.14	1.70	1.61
Migraine	0.57	1.33	0.78	1.30	2.08	1.99	1.16	1.17	1.16
CIRCULATORY									
Heart Disease	3.03	4.92	4.97	4.23	5.55	5.11	4.31	4.59	4.38
Hypertension	3.98	5.61	6.02	5.91	5.90	5.62	5.71	5.53	**5.17**
Arterial Embolism and Thrombosis	1.57	3.20	2.68	3.96	4.39	1.53	3.00	3.51	3.11
Atherosclerosis	2.34	3.62	3.05	3.46	3.86	4.11	2.92	3.05	3.02
RESPIRATORY									
Bronchiectasis	6.70	7.80	4.92	6.90	8.49	6.66	6.99	6.39	**6.03**

Pneumonia	1.88	2.82	2.12	2.49	3.31	3.07	2.62	2.32	2.14
Emphysema	7.36	9.02	3.90	5.26	9.04	7.05	5.79	5.56	**5.30**
Asthma	1.32	1.78	1.44	1.52	2.37	1.68	2.01	1.71	1.67
Allergic Rhinitis	0.57	1.02	0.66	0.87	0.83	1.14	1.05	0.83	0.82
Chronic Bronchitis	4.45	4.58	5.17	3.74	6.56	4.75	5.51	4.72	4.50
Chronic Sinusitis	0.89	1.32	1.76	1.35	0.99	1.35	1.12	1.08	1.07
DIGESTIVE									
Cirrhosis	1.95	2.47	1.86	2.54	3.78	2.58	2.49	2.29	2.31
Peptic Ulcer	1.53	2.84	2.61	2.89	3.33	3.31	2.55	2.86	2.70
MUSCULOSKELETAL									
Osteoporosis	2.62	4.12	3.71	3.21	3.04	4.11	2.84	3.56	3.40
Arthritis	3.03	4.19	5.77	3.79	5.43	3.66	5.37	4.08	3.93
CONTAGIOUS									
Tuberculosis	7.86	8.10	9.75	7.16	7.96	6.44	6.41	7.58	**7.11**

References

[1] U. Sulkowska, M. Mańczuk, J. Łobaszewski, W.A. Zatoński, Lung cancer, the leading cause of cancer deaths among women in Europe, Nowotwory. Journal of Oncology 65 (5) (2015) 395–403.

[2] Health Promotion Administration, Ministry of Health and Welfare, 2014 health promotion administration annual report, Health Promotion Administration, Ministry of Health and Welfare, Taipei, 2014.

[3] H.-C. Lang, S.-L. Wu, Lifetime costs of the top five cancers in taiwan, Eur. J. Health Econ. 13 (3) (2012) 347–353.

[4] J.J. Carlson, K. Suh, P. Orfanos, W. Wong, Cost effectiveness of alectinib vs. Crizotinib in first-line anaplastic lymphoma kinase-positive advanced non-small-cell lung cancer, Pharmacoeconomics 36 (4) (2018) 495–504.

[5] T.-Y. Li, J.-S. Hsieh, K.-T. Lee, M.-F. Hou, C.-L. Wu, H.-Y. Kao, H.-Y. Shi, Cost trend analysis of initial cancer treatment in taiwan, PLoS One 9 (10) (2014) 1–11.

[6] T. Blakely, J. Atkinson, G. Kvizhinadze, N. Wilson, A. Davies, P. Clarke, Patterns of cancer care costs in a country with detailed individual data, Med. Care 53 (4) (2015) 302.

[7] W.P. Smith, P.J. Richard, J. Zeng, S. Apisarnthanarax, R. Rengan, M.H. Phillips, Decision analytic modeling for the economic analysis of proton radiotherapy for non-small cell lung cancer, Transl. Lung Cancer Res. 7 (2) (2018) 122.

[8] Ministry of Health and Welfare, National Health Insurance Annual Report vol. 12, Ministry of Health and Welfare, Taipei, 2014.

[9] R.-E. Chang, T.-L. Chiang, Risk adjustment: a key to efficiency and equity in the health insurance market, Chin. J. Public Health 17 (5) (1998) 373–380.

[10] R.-E. Chang, W. Lin, C.-J. Hsieh, T.-L. Chiang, Healthcare utilization patterns and risk adjustment under taiwan's national health insurance system, J. Formos. Med. Assoc. 101 (1) (2002) 52–59.

[11] N. Mantel, W. Haenszel, Statistical aspects of the analysis of data from retrospective studies of disease, J. Natl. Cancer Inst. 22 (4) (1959) 719–748.

[12] H.M. Zolbanin, D. Delen, A.H. Zadeh, Predicting overall survivability in comorbidity of cancers: a data mining approach, Decis. Support Syst. 74 (2015) 150–161.

[13] R.-E. Chang, C.-L. Lai, Risk adjuster: the basis for capitation payment, Chin. J. Publ. Health 23 (2) (2004) 91–99.

[14] N. Rice, P.C. Smith, Capitation and risk adjustment in health care financing: an international progress report, Milbank Q. 79 (1) (2001) 81–113.

[15] W.P.M.M. van de Ven, R.P. Ellis, Risk adjustment in competitive health plan markets, in: A.J. Culyer (Ed.), Handbool of Health Economics, Elsevier Science, Amsterdam, 2000, pp. 755–845.

[16] K. Kapur, A.S. Young, D. Murata, Risk adjustment for high utilizers of public mental health care, J. Ment. Health Pol. Econ. 3 (3) (2000) 129–137.

[17] G.C. Pope, R.P. Ellis, A.S. Ash, C.-F. Liu, J.Z. Ayanian, D.W. Bates, Principle inpatient diagnostic cost group model for medicare risk adjustment, Health Care Financ. Rev. 21 (3) (2000) 93–118.

[18] W. Lin, R.-E. Chang, C.-J. Hsieh, C.-L. Yaung, T.-L. Chiang, Development of a risk-adjusted capitation model based on principal inpatient diagnoses in taiwan, J. Formos. Med. Assoc. 102 (9) (2003) 637–643.

[19] R.-E. Chang, C.-L. Lai, Use of diagnosis-based risk adjustment models to predict individual health care expenditure under the national health insurance system in taiwan, J. Formos. Med. Assoc. 104 (12) (2005) 883–890.

[20] T.A. Omachi, S.E. Gregorich, M.D. Eisner, R.A. Penaloza, I.V. Tolstykh, E.H. Yelin, C. Iribarren, R.A. Dudley, P.D. Blanc, Risk adjustment for health care financing in chronic disease: what are we missing by failing to account for disease severity? Med. Care 51 (8) (2013) 740–747.

[21] T.-M. Cheng, Taiwan's Health Care System: the Next 20 Years, (2015) Taiwan-U.S. Quarterly Analysis(17) http://www.brookings.edu/research/opinions/2015/05/14-taiwan-national-healthcare-cheng.

[22] V.R. Fuch, The growing demand for medical care, N. Engl. J. Med. 279 (1968) 190–195.

[23] Y.-C. Ko, C.-H. Lee, M.-J. Chen, C.-C. Huang, W.-Y. Chang, H.-J. Lin, H.-Z. Wang, P.-Y. Chang, Risk factors for primary lung cancer among non-smoking women in Taiwan, Int. J. Epidemiol. 26 (1) (1997) 24–31.

[24] J.M. Samet, E.A. Tang, P. Boffetta, L.M. Hannan, S.O. Marston, M.J. Thun, C.M. Ruding, Lung cancer in never smokers: clinical epidemiology and environmental risk factors, Clin. Canc. Res. 15 (18) (2009) 5626–5645.

[25] S. Pannarunothai, P. Phanthunane, Using utilisation data to estimate future demand for medical internists: the impact of demographic demand driver in Thailand, Stud. Health Technol. Inf. 178 (2012) 169–174.

[26] S.C. Stearns, M.G. Kovar, K. Hayes, G.G. Koch, Risk indicators for hospitalization during the last year of life, Health Serv. Res. 31 (1) (1996) 49–69.

[27] N. Howlader, A.M. Noone, M. Krapcho, J. Garshell, D. Miller, S.F. Altekruse, C.L. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D.R. Lewis, H.S. Chen, E.J. Feuer, K.A. Cronin (Eds.), SEER Cancer Statistics Review, 1975-2011, Retrieved from National Cancer Institute, Bethesda, MD, 2014, http://seer.cancer.gov/csr/1975_2011/ (D).

[28] B.-Y. Wang, J.-Y. Huang, C.-Y. Cheng, C.-H. Lin, J.-L. Ko, Y.-P. Liaw, Lung cancer and prognosis in taiwan: a population-based cancer Registry, J. Thorac. Oncol. 8 (9) (2013) 1128–1135.

[29] F.V. Jensen, Introduction to Bayeisan Network, Springer, Berlin, 1996.

[30] L. Uusitalo, Advantages and challenges of bayesian networks in environmental modelling, Ecological Modeling 203 (2007) 312–318.

[31] P.J. Lucas, L.C. van der Gaaq, A. Abu-Hanna, Bayesian networks in biomedicine and health-care, Artif. Intell. Med. 30 (3) (2004) 201–214.

[32] J.H. Oh, J. Craft, R. Al Lozi, M. Vaidya, Y. Meng, J.O. Deasy, et al., A Bayesian network approach for modeling local failure in lung cancer, Phys. Med. Biol. 56 (6) (2011) 1635.

[33] P. Petousis, S.X. Han, D. Aberle, A.A. Bui, 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. 72 (2016) 42–55.

[34] B.R. Cobb, R. Rumi, A. Salmeron, Bayesian network models with discrete and continuous variables, Advances in Probabilistic Graphical Models, Studies in Fuzziness and Soft Computing, vol. 214, 2007, pp. 81–102.

[35] S.L. Lauritzen, Propagation of probabilities, means, and variances in mixed graphical association models, J. Am. Stat. Assoc. 87 (420) (1992) 1098–1108.

[36] P. Lucena-Moya, R. Brawata, J. Kath, E. Harrison, S. ElSawah, F. Dyer, Discretization of continuous predictor variables in bayesian networks: an ecological threshold approach, Environ. Model. Software 66 (2015) 36–45.

[37] S.L. Lauritzen, N. Wermuth, Graphical models for associations between variables, some of which are qualitative and some quantitative, Ann. Stat. 17 (1) (1989) 31–57.

[38] S.L. Lauritzen, F. Jensen, Stable local computation with conditional Gaussian distributions, Stat. Comput. 11 (2001) 191–203.

[39] Medical News (2018) https://www.medicalnewstoday.com.

[40] Cancer.Net, Cancer.Net. Retrieved from Cancer.Net, (2015) http://www.cancer.net/cancer-types/lung-cancer.

[41] Ministy of Health and Welfare, Ministy of health and Welfare, national health insurance administration, Retrieved from Ministy of Health and Welfare, National Health Insurance Administration: http://www.nhi.gov.tw/webdata/webdata.aspx?menu=18&menu_id=683&webdata_id=444.

[42] American Cancer Society, Global Cancer Facts and Figures, second ed., American Cancer Society, Atlanta, 2011.

[43] K.-J. Wang, B. Makond, K.-M. Wang, Modeling and predicting the occurrence of brain metastasis from lung cancer by bayesian network: a case study of taiwan, Comput. Biol. Med. 47 (2014) 147–160.

[44] C.-Y. Liu, Y.-T. Hung, Y.-L. Chuang, Y.-J. Chen, W.-S. Weng, J.-S. Liu, K.-Y. Liang, Incorporating development stratification of taiwan townships into sampling design of large scale health interview survey, J. Health Manag. 4 (1) (2006) 1–22.

[45] J. Corral, J.A. Espinàs, F. Cots, L. Pareja, J. Solà, R. Font, J.M. Borràs, Estimation of lung cancer diagnosis and treatment costs based on a patient-level analysis in catalonia (Spain), BMC Health Serv. Res. 15 (70) (2015) 1–10.

[46] M.A. Cucciare, W. O'Donohue, Predicting future healthcare cost: how well does risk-adjustment work? J. Health Organisat. Manag. 20 (2) (2006) 150–162.

[47] S.G. Bøttcher, C. Dethlefsen, Deal: a package for learning bayesian networks, J. Stat. Software 8 (20) (2003) 1–40 (Retrieved from Deal: A package for learning bayesian networks).

[48] M. Scutari, Learning bayesian networks with the bnlearn R package, J. Stat. Software 35 (3) (2010) 1–22.

[49] M. Scutari, Retrieved from Bnlearn: Bayesian Network Structure Learning, Parameter Learning and Inference, (2015) http://www.bnlearn.com/.

[50] H.-Y. Chang, Evaluation of Alternative Diagnosis-Based Risk Adjustment Models and Morbidity Trajectories for Application in Taiwan, The Johns Hopkins University, Baltimore, 2009.

[51] S.M. Mehmud, T.G. Sawhney, R. Yi, Applications of Risk Adjustment, Society of Actuaries, Baltimore, 2013.

[52] R. Winkelman, S. Mehmud, L. Wachenheim, A Comparative Analysis of Claims-Based Tools for Health Risk Assessment, Society of Actuaries, 2007.