Using machine learning approaches to predict high-cost chronic obstructive pulmonary disease patients in China

Li Luo
Jialing Li
Shuhao Lian
Sichuan University, China

Xiaoxi Zeng
Lin Sun
Chunyang Li
West China Hospital of Sichuan University, China

Debin Huang
Chengdu Medical Insurance Administration, China

Wei Zhang
Sichuan University, China

Abstract
The accurate identification and prediction of high-cost Chronic obstructive pulmonary disease (COPD) patients is important for addressing the economic burden of COPD. The objectives of this study were to use machine learning approaches to identify and predict potential high-cost patients and explore the key variables of the forecasting model, by comparing differences in the predictive performance of different variable sets. Machine learning approaches were used to estimate the medical costs of COPD patients using the Medical Insurance Data of a large city in western China. The prediction models used were logistic regression, random forest (RF), and extreme gradient boosting (XGBoost). All three models had good
predictive performance. The XGBoost model outperformed the others. The areas under the ROC curve for Logistic Regression, RF and XGBoost were 0.787, 0.792 and 0.801. The precision and accuracy metrics indicated that the methods achieved correct and reliable results. The results of this study can be used by healthcare data analysts, policy makers, insurers, and healthcare planners to improve the delivery of health services.

**Keywords**
high-cost chronic obstructive pulmonary disease patients, machine learning approaches, prediction models

**Introduction**

In recent years, with overall improvements in living standards, health care is becoming increasingly important worldwide. The total cost of health care is also gradually increasing. In 2006, the total cost of medical and health care in China was ¥984.3 billion, while in 2016 China’s total health expenditure rose to ¥4634.5 billion, an increase of nearly four times. Other countries are showing similar trends in healthcare spending. In the United States, health spending increased from US$680.8 billion in 2006 to US$989.23 billion in 2016.\(^1\)

Research into medical expenses is often concentrated on medical management in foreign countries, especially in countries with advanced levels of medical care. Since the 1980s, the potential use of medical insurance administrative claims data for predicting demand has become the subject of considerable research. Many studies have proposed prediction models based on administrative claims data for medical expenses.\(^2\)–\(^5\)

These predictive models can be used for individual and group cost predictions, and as a basis for patient intervention in health management. Several algorithms have been used for review of financial reimbursement and insurance pricing. In China, due to the public welfare and nonprofit nature of medical and health services in the past, domestic research into the prediction of medical expenses is rare. There are similar situations in developing countries. Therefore, developing an expenditure-forecasting model that is applicable to countries such as China should provide effective methodological support for controlling the excessive growth of medical expenses, and for the review of medical insurance costs.

Chronic obstructive pulmonary disease (COPD) is a severe chronic disease that is a major cause of morbidity and mortality. In the developed world, COPD is the fastest growing cause of death.\(^6\) In China, air pollution has worsened over the past few years, due to rapid industrialization and urbanization. COPD has become a major leading cause of death in China. There are more than 38 million patients with COPD in China, and the prevalence rate of COPD in patients over 40 years old reached 9.9 percent in 2015.\(^7\)

Accurately identifying and predicting high-cost COPD patients are significant steps in addressing the potential economic burden of COPD. Research has been conducted into the predictive modeling of medical costs or expenditures based on claims data, using statistical approaches such as heuristic rules and regression methods.\(^8\),\(^9\) Most of this research has focused on high-cost profiles among patients who have been diagnosed with specific medical conditions including asthma,\(^10\),\(^11\) chronic disease,\(^12\),\(^13\) and cancers.\(^14\) Despite the prevalence of COPD and its debilitating effects on the patient quality of life, most research has focused on the prevention of exacerbation and the management of COPD.\(^15\)–\(^17\) There have been a few studies evaluating the economic burden of this disease.\(^18\) A significant shortcoming of those studies is the predictive approach. Traditional statistical methods are still the most popular. There are few forecasting models that use China’s medical insurance data or claims data. Increasingly, many people have identified machine learning
approaches as being valuable for the analysis of healthcare data.\textsuperscript{19,20} However, a drawback of these algorithms is that they often require many variables with which to build a predictive model. For example, Lee et al.\textsuperscript{21} applied neural network and decision tree models to predict the hospital charges of colorectal cancer patients, using 154 variables. Some variables used in research using data from the Medical Expenditure Panel Survey (MEPS), such as body mass index (BMI), smoking status, household income, and family population are unavailable in China,\textsuperscript{22–24} and the differences in the datasets may affect the performance of predictive models. Therefore, we aimed to develop a model for predicting high-cost patients using China’s existing database, and we tried to identify key predictors among a large number of variables obtained from admissions information, in order to construct a model that still generalizes well.

We ranked the variables, used machine learning approaches to identify high-medical-cost COPD patients, and predicted high-expenditure patients. The model uses a minimal number of high-impact variables, including some that are unique to China. Three algorithms were used: logistic regression, random forests (RF), and extreme gradient boosting (XGBoost). The main contribution of this article is that we implement common machine learning approaches which can be used by medical workers in other regions of China. Similar results can be obtained using key variables, which are easy to collect because Medical Insurance Data in other parts of China includes similar fields. This study follows the recent trend of using machine learning approaches to forecast healthcare-related issues. For example, Liu and Shi\textsuperscript{25} developed a predictive model and applied an ensemble of binary classification and regression techniques to predict how many days patients would spend in the hospital in the following year. Araújo et al.\textsuperscript{26} applied machine learning classifiers (random trees, naive Bayes, support vector machines, and nearest neighbor) to support healthcare professionals in making pre-authorization decisions. The findings of this study can be of significant value for healthcare data analysts, policymakers, insurers, and healthcare planners for improving the delivery of health services, especially in cases in which the medical data are imperfect.

**Materials and methodology**

We aimed to use patients’ previous year’s hospitalization information to predict potential high-cost patients for the following year. The research methodology is shown in Figure 1. It is divided into the following steps: (1) we introduce the data of this study; (2) preprocessing the data: sample selection, attribute selection, and definition of the high-cost patient; (3) we use three machine learning methods and their final attributes to build predictive models, and select the best predictive
model based on three performance metrics—area under the receiver operating characteristic curve (AUC), sensitivity (SEN), and specificity (SPE); (4) for the best model we rank the impact of the variables, and then use the best model and the key variables to repeat the predictions, and compare the results with the original model.

**Data source**

This study used Medical Insurance Data from a large city in western China for the period from 1 January 2011 to 31 December 2013. We extracted data from the admission records of COPD patients, whose ICD-10 codes for any of the first three diagnoses were J44.1 or J44.901, with 780,295 records. Each admission record contains basic demographic information, diagnostic information (using ICD-10-CM code, International Classification of Diseases, 10th Revision, Clinical Modification), clinical and medical information regarding hospitalization, together with cost information. Based on an extensive review of all the variables in the database files, we selected 54 basic variables. We grouped the original 54 variables into four separate categories: Demographics (5 attributes), Disease information (8 attributes), Hospitalization information (30 attributes), and Medication information (11 attributes) (Table 1).

**Data preprocessing**

High-quality data are important for accurate predictive modeling. Data preprocessing was divided into two phases: sample selection and variable selection. All data preprocessing was performed using the R software.

**Sample selection.** The primary dataset included 780,295 records, with each record representing one COPD patient’s admission record, and 54 attributes. After removing records with more than 60 percent null values for variables, removing records with negative values or abnormal sex attributes, and discarding records of patients who died during their hospital stay; the final dataset included 650,839 rows. We aimed to predict whether these patients would become high-cost patients in the next year, using the patient’s first-year admission data. Therefore, we needed to merge the records of each patient’s hospital admission by year. After the data were combined, we identified three kinds of COPD patients: COPD patients in 2011 (n=62,490), COPD patients in 2012 (n=93,760), and COPD patients in 2013 (n=106765).

**Attribute merger and adjustment.** Variables were then merged and adjusted. The detailed conversion of variables is shown in Table 2.

**Demographics.** In the demographics, after the variables were combined Gender, Age, Region and Insurance were not changed. We removed PID from the demographics group.

**Disease information.** For the diagnosis information, we summarized the comorbidities of these patients by ICD-10 code (Table 9 in Appendix 1), and calculated the number and proportion of patients with comorbidities (Table 10 in Appendix 1). We then added the top 10 ranked comorbidities (comorbidities with proportions greater than 11 percent, shown in bold in Table 10 in Appendix 1), as new attributes. Among those comorbidities, there was high correlation between CPHD, which stands for chronic pulmonary heart disease, and Heart Disease (Pearson’s r=0.774, significant, p value < 0.001); the correlation between Digestive System Disease and No. Comorbidities (the number of complications) is also highly significant (Pearson’s r=0.486, significant,
Table 1. The original 54 attributes.

| Attribute                           | Description                                                                 |
|-------------------------------------|-----------------------------------------------------------------------------|
| **Demographics (5)**                |                                                                             |
| PID                                 | Personal information code (unique)                                         |
| Gender                              | Gender                                                                      |
| Age                                 | Age                                                                         |
| Region                              | The area in which the patient lives (21 regions)                           |
| Insurance                           | Health insurance type (urban or rural)                                     |
| **Disease information (8)**         |                                                                             |
| DX1–DX8                             | Primary diagnosis, 2nd diagnosis, 3rd diagnosis, . . . , 8th diagnosis       |
| **Hospitalization information (30)**|                                                                             |
| Inpatient date                      | Inpatient date                                                              |
| Sequence of admissions              | Sequence of admissions this year                                            |
| HL                                  | Hospital level                                                              |
| HC                                  | Hospital code                                                               |
| Discharge department                | Discharge department of hospital                                            |
| LOS                                 | Length of stay                                                              |
| LFLA                                | Length from last admission                                                  |
| Total cost                          | Total cost                                                                  |
| Drug cost                           | Drug cost                                                                   |
| Check Cost                          | Check cost                                                                   |
| Equipment cost                      | Service equipment cost                                                      |
| Diagnosis cost                      | Diagnosis cost                                                               |
| Other cost                          | Other cost                                                                   |
| Self-paid cost                      | All self-paid cost                                                          |
| Reimbursement cost                  | Reimbursement cost                                                          |
| Daily cost                          | Mean daily cost                                                             |
| DODC                                | Days over daily cost                                                        |
| Highest cost                        | Highest daily cost                                                          |
| Surgery                             | Whether surgery was performed                                               |
| Surgery code                        | Surgery code                                                                |
| ICU times                           | ICU times                                                                   |
| Died                                | Died during hospitalization                                                  |
| Times of ECG                        | Times of ECG monitoring                                                     |
| Cost of ECG                         | Cost of ECG monitoring                                                      |
| Ambulance                           | Whether an ambulance was used                                               |
| Times of oxygen                     | Times of oxygen service                                                     |
| Oxygen cost                         | Oxygen service cost                                                         |
| Times of nursing                    | Times of nursing                                                            |
| Nursing cost                         | Nursing cost                                                                |
| HSDC                                | Highest Single-Day Cost                                                     |
| **Medication information (11)**     |                                                                             |
| TCM                                 | Traditional Chinese medicine cost                                          |
| Aspirin                             | Whether aspirin was used                                                    |
| Aspirin cost                        | Aspirin cost                                                                |
| Antibiotics                         | Whether antibiotics were used                                               |
| β2                                  | Whether β2 agonists were used                                               |
| Anticholinergic                     | Whether anticholinergics were used                                         |
| Combined Drug1                      | Whether short-acting β2 agonists and anticholinergic drugs were used        |
| Glucocorticoid                      | Whether glucocorticoids were used                                          |
| Combined Drug2                      | Whether long-acting β2 agonists and inhaled corticosteroids combined        |
|                                     | preparations were used                                                      |
| GWB                                 | Whether glucocorticoid-whole-body was used                                 |
| Methyl xanthine                     | Whether methyl xanthine was used                                            |

ICU: intensive care unit; ECG: electrocardiogram.
p value < 0.001), and the latter are more relevant to the target. Therefore, we removed CPHD and Digestive System Disease from the Disease information group.

We added four new attributes, including Cancer, and Hypertension, because chronic diseases have a large influence on costs for the following year. We also added Which dia. COPD (which diagnosis is COPD) and No. Comorbidities to reflect the severity of the patient’s illness.

Comorbidity is the presence of one or more disorders or diseases co-occurring with the primary disease or disorder. The Charlson Comorbidity Index (CCI) is a weighted index that takes into account the number and seriousness of the comorbid diseases. In our study, we built CCI score sheets for this study using 14 modified CCI conditions, scoring standards and scores (Table 3), as in research published by Sundararajan et al. The scores represent the objective risk of dying associated with each condition. Finally, we calculated the CCI score of each patient based on the modified CCI and discharge diagnosis, and used it as a predictor in the model.
We replaced the readmission date with the Sequence of Admissions in each quarter. We removed Died, Discharge department, and Surgery code because (1) we excluded the dead patients, so there is no reason for us to retain the attribute Died, and (2) the attributes Discharge department and Surgery code had too many missing values, due to improper recording, and so were removed from our data.

We converted the expenditure-related variables to percentages with the total cost of a specific item as the numerator and the total hospitalization cost as the denominator. For example, Drug Cost Proportion (DCP) is the proportion of the total drug cost in the given year to the total hospitalization cost, and reimbursement cost proportion (RCP) is the proportion of the reimbursement cost. LOS, CCP (Check Cost Proportion), and OCR (Other Cost Proportion) were also discarded, as they were strongly correlated with Total Cost, DCP (Diagnosis Cost Proportion), and ASPCP (All self-pay proportion), respectively (Pearson’s $r = 0.777$, Pearson’s $r = -0.640$, Pearson $r = 0.517$, significant, $p$ value $< 0.001$). For similar reasons, we removed Oxygen Cost, and Nursing Cost.

We converted HC to a new variable, HC average cost, the average cost of hospitalized COPD patients in the hospital in the current year. In addition, based on the day of the patient’s maximum daily hospitalization cost, we created a new variable, MAXGAP, which is equal to the date of discharge minus the date when the highest daily hospitalization cost occurred. Finally, we considered fluctuations in the daily cost of two patients with the same average daily cost and the same length of stay. For example, looking at two records (Figure 2), it is apparent that although the average daily cost is the same, fluctuations in the daily cost of sample 1 are more extreme than those of sample 2. We created a new variable, Sharpness, which reflects fluctuations in the cost of hospitalization for patients. Sharpness equals the number of hospital days in excess of the average daily cost divided by the total number of length of days in hospital (days over daily cost/LOS).

**Medication information.** Anticholinergics and Combined Drug1 (Pearson’s $r = 0.930$, significant, $p$ value $< 0.001$); B2 LABA and Glucocorticoid (Pearson’s $r = 0.548$, significant, $p$ value $< 0.001$);

### Table 3. Modified CCI.

| No. | Modified CCI conditions | Scores |
|-----|-------------------------|--------|
| 1   | Patient’s age           | 0/1/2/3/4 |
| 2   | Myocardial Infarction   | 0/1     |
| 3   | Congestive heart failure| 0/1     |
| 4   | Peripheral vascular disease| 0/1     |
| 5   | Cerebrovascular disease | 0/1     |
| 6   | Dementia                | 0/1     |
| 7   | Connective tissue disease| 0/1     |
| 8   | Peptic ulcer disease    | 0/1     |
| 9   | Moderate to severe chronic kidney disease| 0/1     |
| 10  | Hemiplegia              | 0/2     |
| 11  | Leukemia                | 0/2     |
| 12  | Malignant lymphoma      | 0/2     |
| 13  | Liver disease           | 0/3     |
| 14  | AIDS                    | 0/6     |

CCI: Charlson Comorbidity Index.

For Nos 2–14 modified CCI conditions, if a patient did not have the disease in the condition, the score of the condition was 0; if the patient had the disease, then its score was added. For instance, for a 48-year-old COPD patient with peripheral vascular disease and malignant lymphoma, the CCI score is 4 points, consisting of 1(age) + 1 (peripheral vascular disease) + 2 (malignant lymphoma), to produce 4.
and Total hospitalization cost and Combined Drug2 (Pearson’s $r = 0.363$, significant, $p$ value $< 0.001$) were significantly correlated, so we discarded Combined Drug1, Glucocorticoid, and Combined Drug2.

Through the variable selection process, we identified 50 predictors that could be used for modeling. Table 4 presents descriptive statistics for the 62,490 members in the combined model building and testing subsamples, as well as for the subset of children ($n = 12,498$) with cost in the top 20 percent. The proportion of high-cost patients with urban health insurance is much higher than that of the whole sample, which means that urban health insurance patients are more likely to become high cost. The number of comorbidities in high-cost patients was significantly higher than the number of comorbidities in the full-sample patients.

Before building the predictive model, we needed to discretize the continuous variables. We used the equation $X_i = (X_i - \text{mean}(X))/\text{SD}(X)$ to discretize Age, No. Comorbidities, Which Dia. COPD, HC Average Cost, Mean Gap, Total Cost, Daily Cost, Highest Cost, MAXGAP, TCM Cost, Times of ECG, Oxygen Times.

**Definition of future high-cost patient**

We aimed to use information from 1 year to identify potential high-cost COPD patients and predict the potential high-cost COPD patients in the next year. The way in which to define thresholds for high-cost patients is an open question. Various thresholds (top 5%, 10%, 20%, and 50%) for defining “high cost” have been used.$^{19,29,31}$ Due to the lack of standard high-cost patient definitions, we explored the high-cost thresholds through data analysis. The cumulative cost in 2011 to 2012 COPD patients was ¥7 million, with the top 5, 10, and 20 percentiles accounting for 35.2 percent, 51.8 percent, and 72.1 percent of the total cost, respectively. Similar cost figures are obtained for patients in 2012 to 2013 COPD patients (Table 5).

The well-known 80/20 Pareto principle (also known as the 80/20 rule) was proposed by the 19th-century Italian professor of political economy Vilfredo Pareto. He calculated that the top 20 percent of any country’s population accounts for approximately 80 percent of its total income.
Table 4. Fifty attributes of the samples (2011).

| Attribute (predictor) | Type       | Proportion or mean (full sample N = 62,490) | Proportion or mean (high cost (20%) N = 12,498) |
|-----------------------|------------|---------------------------------------------|-----------------------------------------------|
| **Demographics (4)**  |            |                                             |                                               |
| Gender (male, %)      | Binary     | 54.0                                        | 59.5                                          |
| Age (mean (SD))       | Continuous | 71.5 (9.9)                                  | 72.63 (9.3)                                   |
| Region (region 0, %)  | Continuous | 15.9                                        | 29.2                                          |
| Insurance (rural, %)  | Binary     | 50.9                                        | 29.7                                          |
| **Disease information (18)** |        |                                             |                                               |
| Which Dia. COPD (mean (SD)) | Integer | 1.1 (0.4)                                    | 1.1 (0.4)                                     |
| No. comorbidities (mean (SD)) | Integer | 2.8 (3.2)                                    | 4.3 (4.1)                                     |
| AECOPD (%)            | Binary     | 79.8                                        | 84.1                                          |
| Heart disease (%)     | Binary     | 35.7                                        | 47.2                                          |
| Hypertension (%)      | Binary     | 18.6                                        | 27.6                                          |
| Emphysema (%)         | Binary     | 19.8                                        | 28.0                                          |
| Chronic bronchitis (%)| Binary     | 14.2                                        | 19.9                                          |
| ORSD (%)              | Binary     | 13.0                                        | 17.6                                          |
| Urogenital disease (%)| Binary     | 11.7                                        | 17.9                                          |
| Pneumonia (%)         | Binary     | 9.0                                         | 11.7                                          |
| CAD (%)               | Binary     | 8.8                                         | 14.6                                          |
| Diabetes (%)          | Binary     | 6.1                                         | 10.4                                          |
| ALRI (%)              | Binary     | 3.0                                         | 3.0                                           |
| Heart failure (%)     | Binary     | 7.4                                         | 10.0                                          |
| Arthritis (%)         | Binary     | 2.7                                         | 4.3                                           |
| Asthma (%)            | Binary     | 2.9                                         | 4.8                                           |
| Cancer (%)            | Binary     | 2.40                                        | 3.4                                           |
| CCI score (mean (SD)) | Integer    | 3.5 (1.0)                                   | 3.7 (1.0)                                     |
| **Hospitalization information (21)** |        |                                             |                                               |
| HosSeason1 (mean (SD)) | Integer   | 0.7 (0.8)                                   | 0.9 (0.9)                                     |
| HosSeason2 (mean (SD)) | Integer   | 0.5 (0.8)                                   | 0.8 (0.9)                                     |
| HosSeason3 (mean (SD)) | Integer   | 0.5 (0.8)                                   | 0.7 (0.9)                                     |
| HosSeason4 (mean (SD)) | Integer   | 0.6 (0.8)                                   | 1.0 (1.0)                                     |
| HC average cost (mean (SD)) | Continuous | 6080.9 (4334.9)                           | 8229.9 (4457.3)                              |
| Mean gap (mean (SD))  | Continuous | 33.0 (40.0)                                 | 43.6 (37.2)                                  |
| Total cost (mean (SD)) | Continuous | 13,248.4 (195,178)                        | 25,507.3 (27,285)                            |
| DCP (mean (SD))       | Continuous | 54.4% (14.4%)                               | 56.4% (13.5%)                                |
| ECP (mean (SD))       | Continuous | 6% (3.7%)                                   | 5% (2.9%)                                     |
| DIACP (mean (SD))     | Continuous | 14.6% (7.5%)                                | 14.2% (6.9%)                                  |
| ASPCP (mean (SD))     | Continuous | 7.2% (6.1%)                                 | 6.3% (5.6%)                                   |
| RCP (mean (SD))       | Continuous | 78.9% (13.1%)                               | 84.0% (11.2%)                                 |
| Daily cost (mean (SD)) | Continuous | 468.4 (347.3)                               | 542.8 (332.1)                                 |
| Sharpness (mean (SD)) | Continuous | 0.21 (0.15)                                 | 0.15 (0.12)                                   |
| Highest cost (mean (SD)) | Continuous | 1132.3 (3099.1)                           | 1498.2 (2111.6)                              |
| Surgery (%)           | Binary     | 7.4                                         | 8.9                                           |
| Times of ECG (mean (SD)) | Integer   | 70.8 (206.0)                                | 126.6 (304.7)                                 |
| Times of oxygen (mean (SD)) | Integer   | 17.46 (28.0)                               | 34.90 (44.2)                                  |
| Times of nursing (mean (SD)) | Integer   | 46.70 (129.5)                              | 79.97 (171.3)                                 |

(Continued)
Table 4. (Continued)

| Attribute (predictor)                | Type          | Proportion or mean (full sample N = 62,490) | Proportion or mean (high cost (20%) N = 12,498) |
|--------------------------------------|---------------|---------------------------------------------|-------------------------------------------------|
| Ambulance (%)                        | Binary        | 4.3                                         | 5.2                                             |
| MAXGAP (mean (SD))                   | Continuous    | 9.4 (7.1)                                   | 12.0 (9.2)                                      |
| Medication information (7)           |               |                                             |                                                 |
| TCM Cost (mean (SD))                 | Continuous    | 1026.4 (2141.1)                             | 2124.6 (3578.9)                                 |
| Aspirin (%)                          | Binary        | 12.0                                        | 21.1                                            |
| Antibiotics (%)                      | Binary        | 93.9                                        | 96.2                                            |
| β2 (%)                               | Binary        | 33.0                                        | 47.7                                            |
| Anticholinergic (%)                  | Binary        | 8.3                                         | 15.5                                            |
| Systemic corticosteroids (%)         | Binary        | 11.5                                        | 17.6                                            |
| Methyl xanthine (%)                  | Binary        | 84.0                                        | 87.2                                            |

SD: standard deviation; COPD: chronic obstructive pulmonary disease; CCI: Charlson Comorbidity Index; HC: hospital code; DCP: drug cost proportion; ASPCP: all self-pay proportion; RCP: reimbursement cost proportion; TCM: Traditional Chinese medicine AECOPD: acute exacerbation of chronic obstructive pulmonary disease.

ORSID represents Other Respiratory System Disease; HosSeason1 represents hospital times in the first season, and HosSeason2, HosSeason3, and HosSeason4 follow the same rules.

Table 5. COPD patients’ next-year-cost in 2011 and 2012.

| Percentile      | Cost (¥/billion) | Cost (%) |
|-----------------|------------------|----------|
| 2011–2012 COPD patients |
| Top 5%          | 0.27             | 35.2     |
| Top 10%         | 0.41             | 51.8     |
| Top 20%         | 0.57             | 72.1     |
| All records (100%) | 0.79         | 100.0   |
| 2012–2013 COPD patients |
| Top 5%          | 0.40             | 36.9     |
| Top 10%         | 0.59             | 53.7     |
| Top 20%         | 0.81             | 73.6     |
| All records (100%) | 1.10         | 100.0   |

COPD: chronic obstructive pulmonary disease.

“2011 to 2012 COPD patients” refers to the hospitalization cost in 2012 of those patients who were admitted in 2011.

“2012 to 2013 COPD patients” refers to the hospitalization cost in 2013 of those patients who were admitted in 2012.

The 80/20 rule is widely used in economics and management.32,33 Inspired by the 80/20 rule, Gondra34 used two machine learning algorithms—artificial neural networks and support vector machines—to predict the potential faultiness of software components at an early stage. He proposed that testing effort can be focused on a subset of software components, the potentially “most troublesome” 20 percent. In recent years, more and more medical cost/expenses-related research has found that the 80/20 rule holds, and have used it to identify high-cost thresholds.35,36

In this study, we found that 20 percent of patients with COPD consumed nearly 80 percent of the medical resources. Figure 3 shows the proportion of patients with high and low costs in 2011 and 2012 when 20 percent is the high-cost threshold. About 19.4 percent of COPD patients in 2011 were high-cost patients, and 81 percent were low-cost patients. Among these high-cost patients, about 47 percent of those in the next year (2012) were still high-cost patients, and the remaining
53 percent were low-cost patients; about 14 percent of low-cost patients in 2011 were converted to high-cost patients in 2012, and 86 percent of patients continued to be low-cost patients.

**Methods**

We applied three machine learning approaches to predict high-cost COPD patients: logistic regression, RF, and XGBoost. Logistic regression has the advantages of low computational cost, ease of understanding, and ease of implementation. The term “ensemble learning” refers to an algorithm that combines base learners such as decision trees and linear classifiers; the RF and XGBoost approaches are examples of ensemble learning methods. RF is a relatively new predictive method. Its superior performance on large covariate lists and complex interaction structures are the reasons we considered it to be applicable to our study. We adopted XGBoost for the following reasons: (1) Gradient boosting is a well-known machine learning technique for classification and regression problems, and produces a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. The algorithm builds the model in a stepwise fashion, similar to other boosting methods, but is more general than many approaches because it allows the optimization of an arbitrary differentiable loss function. (2) The XGBoost approach utilizes a form of gradient boosting to create an appropriate model for the data without overfitting. It has been successfully used in machine learning applications and performs efficiently. All models were implemented in Python. We used the Sklearn package for logistic regression, and random forest classifiers, and the AdaBoost classifier for RF and XGBoost.

We divided the data from 2011 into a training set (70%) and a validation set (30%). In order to avoid accidental errors due to random sampling, we performed 100 random samplings, so each model was repeated 100 times. The data from 2012 were used as test set I (2012 all tests) and the data from 2013 was used as test set II (2013 all tests).

We assessed the predictive performances of our models using three main measures: AUC, SEN, and SPE, all commonly used performance measures.

The classifier’s sensitivity, also known as true positive rate (TPR), is defined as the ability to correctly identify the true value (TP) of the cases. In this study the metric measures the probability of correctly identifying high-cost patients. The classifier’s specificity, also referred to as true negative (TN), is defined as the ability to correctly identify negative cases. In this study it measured the
probability of the model correctly identifying low-cost patients. The AUC is a measure comparing different classification models, which combines sensitivity and specificity. Sensitivity and specificity are calculated as follows

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \quad \text{Specificity} = \frac{TN}{TN + FP}
\]

**Results**

We initially used thresholds of 5 percent, 10 percent, and 20 percent for high-cost patients to test the predictive performances of all three models with different thresholds. We then selected 20 percent as the threshold for further investigation.

**Performance of the algorithms**

We ran the three algorithms using the training set, in order to build better classifiers by debugging the parameters of each algorithm, and calibrated the classifiers using the validation set. Table 6 shows the performance metrics. The predictive performances of the three different classifiers are significantly higher than that of random guessing on the validation set, test sets I and II, indicating that the model selection and data preprocessing approaches are appropriate. As shown in Table 6, as the threshold is increased, the AUC of each model gradually decreases, for instance, the AUC of XGBoost decreased from 0.842 when using a threshold of top 5 percent to 0.80 when using a threshold of top 20 percent by 0.04 in the validation set. The AUC of RF decreased significantly. When the top 5 percent was used as the high-cost criterion, the AUC of RF in the validation set was 0.828, whereas in test set I, the AUC decreased to 0.752. Compared to logistic regression, it only decreased from 0.826 to 0.817, indicating that the prediction model constructed by RF is less stable than the other two prediction models.

Although the predictive performances using the 20 percent threshold were not optimal, we selected the top 20 percent as the cost target because it is commonly used as a classification standard. Then, we used one classical machine learning algorithm, logistic regression, with all variables, to build predictive Model I, and logistic regression with the base variables to build predictive

| Target      | AUC       | Logistic regression | RF  | XGBoost |
|-------------|-----------|---------------------|-----|---------|
| Validation set | Top5%     | 0.826               | 0.828 | 0.842   |
|              | Top10%    | 0.805               | 0.815 | 0.824   |
|              | Top20%    | 0.787               | 0.792 | 0.800   |
| Test set I   | Top5%     | 0.817               | 0.752 | 0.830   |
|              | Top10%    | 0.803               | 0.769 | 0.810   |
|              | Top20%    | 0.776               | 0.761 | 0.784   |
| Test set II  | Top5%     | 0.806               | 0.737 | 0.812   |
|              | Top10%    | 0.784               | 0.749 | 0.792   |
|              | Top20%    | 0.748               | 0.734 | 0.759   |

AUC: area under the receiver operating characteristic curve; RF: random forest.
Model II. The base variables include Age, Gender, Region, Insurance, and No. Comorbidities. We used the other two algorithms, RF and XGBoost, to build predictive Models III and IV. Our predictive model was constructed using the final 50 variables. The details of the model’s performance are shown in Table 7. The optimal values of the performance metrics sensitivity, specificity, and AUC of the four models are indicated in bold font. For the four models, we used the validation set from the 2011 data with the 20 percent threshold.

(A) Model IV (XGBoost) produced the highest AUC (0.801) [95% confidence interval (CI), 0.8–0.802] compared to Model III (RF, 0.792 [95% CI, 0.792–0.793]) and Model II (logistic regression, 0.787 [95% CI, 0.787–0.788]). Among Models II, III, and IV, each model had markedly improved performance compared to the AUC (0.71 [95% CI, 0.71–0.711]) of Model I. In addition, XGBoost provides the highest sensitivity (71.3%), which is far greater than those of the other four models.

(B) Model III produced the highest specificity (83.3%) compared to Model II (81.4%), Model I (76.9%), and Model 4 (72.9%).

(C) The performances of Model II are not the best under any of the metrics.

The performance metrics of Models II and III are similar, while Model IV shows excellent predictive performance on test sets I and II. Compared to the validation set, AUC and sensitivity decrease on test set I, while Model IV performs the best. The specificities for the four models increased on test sets I and II. Since the AUC of Models II, III, and IV are not significantly different on the verification set, we performed an independent sample T test on the AUC values of 100 experiments in Models II–IV. Before the T test, we first tested the normality of the AUC values of Models II–IV. The three sets of results passed the normality test, based on the p value of Model II (p = 0.568, asymptotic significance), Model III (p = 0.939, asymptotic significance), and Model IV (p = 0.748, asymptotic significance). Subsequently, an independent sample test was performed on

### Table 7. Performance evaluation of the classifiers using 50 attributes.

| Dataset       | Model algorithm           | AUC      | SEN      | SPE      |
|---------------|---------------------------|----------|----------|----------|
| Validation set| Model I Logistic regression (+base attributes) | 0.710 (0.71, 0.711) | 0.517 | 0.769 |
|               | Model II Logistic regression | 0.787 (0.787, 0.788) | 0.584 | 0.814 |
|               | Model III RF              | 0.792 (0.792, 0.793) | 0.564 | **0.833** |
|               | Model IV XGBoost          | **0.801** (0.8, 0.802) | **0.713** | 0.729 |
| Test set I    | Model I Logistic regression (+base attributes) | 0.697 | 0.483 | 0.778 |
|               | Model II Logistic regression | 0.775 | 0.570 | 0.818 |
|               | Model III RF              | 0.765 | 0.545 | 0.835 |
|               | Model IV XGBoost          | **0.783** | | 0.763 |
| Test set II   | Model I Logistic regression (+base attributes) | 0.667 | 0.436 | 0.780 |
|               | Model II Logistic regression | 0.747 | 0.517 | 0.823 |
|               | Model III RF              | 0.737 | 0.499 | 0.836 |
|               | Model IV XGBoost          | **0.758** | | 0.769 |

AUC: area under the receiver operating characteristic curve; RF: random forest. SEN: sensitivity, SPE: specificity. Top 20 percent is defined as high cost. 95 percent confidence intervals are in parentheses.
the AUC values of the three models. The results showed that there was a significant difference in the AUC values of the three models ($p < 0.05$).

Overall, Model IV (XGBoost) yielded the best sensitivity and AUC values, key metrics of predictive performance. For these reasons, we conclude that Model IV outperformed the other three models.

**Attribute importance and attribute group comparison**

XGBoost can score the importance of variables via the Gini index. By sorting all the variables, we gradually stripped out the unimportant variables and got the final 30 variables (see Figure 4). The top 30 variables, as ranked by the XGBoost model, are as follows:

1. Demographics (4): *Age, Region, Insurance, Gender*;
2. Hospitalization information (19): *Total Cost, RCP, DCP, MAXGAP, ASPCP, Sharpness, ECP, HC Average Cost, Highest Cost, DIACP, Daily Cost, Times of Nursing, Mean Gap, Times of Oxygen, Times of ECG, HosSeason4, HosSeason1, HosSeason3, Surgery*;
3. Disease information (6): *No. Comorbidities, CCI Score, Emphysema, Hypertension, Heart Disease, Which Dia. COPD*;
4. Medication information (1): *TCM Cost*.

We use XGBoost with the top 30 variables to build predictive models. We compared two AUC values, one from a prediction model that contains all 50 variables, and the other from the top 30 ranked variables. The average AUC of the latter was 0.798, and was subjected to an independent sample T test, which found a statistically significant difference ($p < 0.001$).

*Total Cost* ranks first among all variables, followed by *Age* and *MAXGAP*. Most cost-related variables are ranked from first to tenth. The reason why the expenditure-related attributes occupy most of the positions is readily apparent: this year’s cost will directly affect the cost of the next year. However, we found that the new variables we derived are also important. For instance, *Sharpness* is ranked seventh. In order to verify the predictive importance of the new attributes (such as Sharpness and CCI Score), and other non-cost variables, we recombined different attributes to get three attribute sets.

![Figure 4. Primary variable importance scores in the XGBoost model.](image-url)
1. The first 30 attributes are set 1.
2. We combined eight attributes associated with direct hospitalization costs from the Hospitalization information group and four attributes from Demographics group as set 3.
3. The remaining 11 attributes of the hospitalization information group and the attributes of the other groups became set 2.

We built a predictive model in the validation set using the XGBoost algorithm with three attributes set, and then compared the AUC values obtained using the different attribute sets (Table 8).

The performance of the prediction models constructed using the three attribute sets declined to different extents. The AUC value using set 1 only decreased by about 0.004. The difference was small compared with the original model. The AUC value using set 3 decreased by about 0.029. However, it increased by 0.06, compared with Model I (logistic regression + base attributes) in Table 7.

**Discussion**

Several studies have investigated the prediction of high-cost/high-medical cost patients using statistical analysis techniques, such as regression and analysis of variance.\textsuperscript{41–43} For example, Fleishman and Cohen\textsuperscript{42} used logistic regression methods to predict the upper 10 percent of expenditures using data from the MEPS. However, with the development of information technology in recent years, the medical and health industry has become an important field for big data applications. It is important to investigate how in-depth analysis can be carried out via data mining approaches to determine the cost trends and variation for patients with a specific disease, based on the big data that are collected by healthcare institutions and insurance companies. Our study attempted to compare the performances of different models and identify high-cost COPD patients by ranking the key variables that can help predict whether a patient will be in the upper 20 percent of the distribution of the subsequent year’s expenditures. We first used 50 attributes to construct COPD high-cost forecasting models based on three data mining algorithms: logistic regression, RF, and XGBoost. Then, we identified the key indicator group, consisting of 30 variables, by scoring their importance using the best forecasting model, Model IV (XGBoost). Finally, in order to investigate the importance of the new variables we created (Sharpness, CCI, etc.) and other non-cost variables for prediction, we re-split the top 30 attributes to get three different combinations of attributes (sets 1, 2, and 3). We compared the differences in AUC of the three attribute sets. Although the predictive performance of the model with sets 1, 2, and 3 decreased slightly, the AUC value still reached 0.798, 0.794, and

---

**Table 8. Comparison of the classifiers using three different attribute sets.**

| Attributes | AUC       | Difference         |
|------------|-----------|--------------------|
| All_50: all 50 attributes | 0.801 (0.8, 0.802) | –                  |
| Set 1: top-30 attributes | 0.798 (0.797,0.798) | 0.004 (0.003,0.004)** |
| Set 2: Demographics (4) + Hospitalization information -A (11) + Disease information(6) + Medication information(1) | 0.794 (0.794,0.794) | 0.007 (0.006,0.007)** |
| Set 3: Demographics (4) + Hospitalization information-B (8) | 0.772 (0.768,0.775) | 0.029 (0.026,0.032)** |

AUC: area under the receiver operating characteristic curve.

n = 18,747, for validation set. 95% confidence intervals in parentheses.

**Significant difference from base model at p < 0.01.

Hospitalization information -A (11): MAXGAP, Sharpness, HC Average Cost, Times of Nursing, Mean Gap, Times of Oxygen, Times of ECG, HosSeason4, HosSeason1, HosSeason3, Surgery.

Hospitalization information -B (8): Total Cost, RCP, DCP, ASPCP, ECP, Highest Cost, DIACP, Daily Cost.
0.772, indicating that these 30 variables can also be used to construct the prediction model, and these variables are easy to obtain.

Our research and that of Fleishman and Cohen differ in terms of data sources, the definition of high-cost patients, and experimental designs. Using machine learning methods to predict the cost of medical care has become a trend in recent years. However, no study has determined whether these models are suitable for Chinese medical data. The latter uses only logistic regression, and we chose three machine-learning approaches, logistic regression, RF, and XGBoost, to construct four models. Among the four models, the predictive performance of Model IV (XGBoost) was optimal, and met the application requirement. On the validation set, the AUC of Model 4 of 0.801 was the best, followed by Model 3 (RF, 0.792) and Model 2 (logistic regression, 0.787). Logistic regression, neural networks, and decision trees are currently very popular predictive methods and have achieved good predictive performances. We used conventional machine learning methods (logistic regression and RF) based on previous studies and also the latest ensemble algorithms (XGBoost). The ensemble model, XGBoost, was the optimal model with the highest AUC value. This study shows that the performance of our predictive models is similar to those using data from other countries and studies based on machine learning methods and Chinese Medical Insurance Data.

We reduced the number of attributes from 50 to 30 by scoring the variables in terms of importance using XGBoost. We found the following: (1) Prior studies suggest that the categories of Demographics (Age, Region), Hospitalization information (Total Cost, ECP), and Disease information (Emphysema, Hypertension) contribute significantly to the prediction of expenditure, similar to the results that were obtained in our study. Our variables contain extensive diagnostic information that is related to COPD, as diagnosis-based models are better at identifying people with “manageable” conditions and those with increasing expenditure trends. (2) We also found that the new variables created in this study, HC Average Cost, MAXGAP, Sharpness, No. Comorbidities, and CCI Score, also played positive roles in the forecasting model. (3) Two doctors suggested that TCM Cost, Aspirin, Antibiotics, and other four drugs are highly related to COPD. These seven variables consist of medication information. We found that TCM Cost is the key predictor variable in the medication information. Our research supports related research, and demonstrated that some new variables make a significant contribution to the prediction. These variables have not been mentioned in other publications.

Our findings have several implications. Research based on predictive models using big data can be applied to disease management, nursing management and resource utilization review. The results of the model can be used as a reference by medical insurance policymakers. Medical policymakers can use the forecasting model with key variables to design local or national surveys, develop a reimbursement policy for COPD patients, and formulate more targeted medical budgets for the high-cost patients. A small proportion of people consume most of the medical resources; such people have become the focus of cost control. Commercial insurance companies can accurately target the potential pool. Based on the attributes of the high-cost patients, they can identify high-spending groups and track them. Machine learning approaches can be used to identify target customers. Before they seek medical help, insurance companies can develop plans for appropriate prevention and response measures, which will help to resolve the conflict between supply and demand of medical and health resources.

Our work has several limitations. The utility of a predictive model depends on the specific context in which it will be used, the cost of data collection, and the ease of interpreting the results. There is no BMI, smoking status, household income, or family population information in the Medical Insurance Data set in our study. However, these predictors can be found in other research, for example, studies and used data from the Veterans Health Administration (VHA), MEPS,
and claims data, respectively. Due to the unique nature of Medical Insurance Data, the primary variables in our study are different from those in other studies, and our original dataset may limit the generalizability of our results and the ease of comparison of our results with those of other research. Approaches used in this study are static models constructed with static variables. In reality, costs can change dynamically over time. The best way in which to construct a dynamic predictive model that adapts to the patient’s state over time is a question worthy of further exploration.

Conclusion

In sum, we applied four machine learning models to predict high-cost COPD individual in the top 20 percentile among COPD patients using Medical Insurance Data. We found that Model IV (XGBoost) performed best, and was an accurate method with which to predict high-cost COPD patients. We ranked the variables using XGBoost and found a set of variables key to the forecasting model. The variables we created were also important contributors to the prediction. We then split and recombined the top 30 variables, and found that set 2, which included the combination of non-cost variables and new variables, still had good predictive performance. Our results provide a basis for medical insurance departments, commercial insurance companies, and hospitals to assist them in formulating management control strategies to reduce medical expenses.

Acknowledgements

The authors gratefully acknowledged the support from the West China Biomedical Big Data Research Center of Sichuan University and Chengdu Medical Insurance Administration.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval

The study does not involve human subjects and adheres to all current laws of China.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was sponsored by the Nature Science Foundation of China (Nos 71532007 and 71131006).

ORCID iD

Jialing Li https://orcid.org/0000-0003-1827-9313

References

1. OECD. Health spending (indicator), 2018. DOI: 10.1787/8643de7e-en.
2. Zhao Y, Ash AS, Ellis RP, et al. Predicting pharmacy costs and other medical costs using diagnoses and drug claims. Med Care 2005; 43(1): 34–43.
3. Bertsimas D, Bjarnadóttir MV, Kane MA, et al. Algorithmic prediction of health-care costs. Oper Res 2008; 56(6): 1382–1392.
4. Song H, Rajan D, Thiagarajan JJ, et al. Attend and diagnose: clinical time series analysis using attention models. In: Thirty-second AAAI conference on artificial intelligence, 2018, https://e-reports-ext.llnl.gov/pdf/891974.pdf
5. Lipton ZC, Kale DC, Elkan C, et al. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

6. Schayck OCV, Rabe KF and Rudolf M. COPD: the role of primary care in effective diagnosis, treatment and management. *Prim Care Resp J* 2003; 12: 16–20.

7. Fang X, Wang X and Bai C. COPD in China: the burden and importance of proper management. *Chest* 2011; 139: 920–929.

8. Tarraf W, Miranda PY and Gonzalez HM. Medical expenditures among immigrant and nonimmigrant groups in the United States: findings from the medical expenditures panel survey (2000-2008). *Med Care* 2012; 50(3): 233–242.

9. Fleishman JA, Cohen JW, Manning WG, et al. Using the SF-12 health status measure to improve predictions of medical expenditures. *Med Care* 2006; 44: i-54–i-63.

10. Tanihara S and Kobayashi Y. Sequential evaluation of the national medical expenditures for asthma care in Japan. *J Epidemiol* 2004; 14(3): 100–103.

11. Sullivan PW, Ghushchyan VH, Slejko JF, et al. The burden of adult asthma in the United States: evidence from the medical expenditure panel survey. *J Allergy Clin Immunol* 2011; 127(2): 363.e1–369.e3.

12. Wish J, Schultman K, Law A, et al. Healthcare expenditure and resource utilization in patients with anemia and chronic kidney disease: a retrospective claims database analysis. *Kidney Blood Press Res* 2009; 32(2): 110–118.

13. Huber CA, Schneeweiss S, Signorell A, et al. Improved prediction of medical expenditures and health care utilization using an updated chronic disease score and claims data. *J Clin Epidemiol* 2013; 66(10): 1118–1127.

14. Halpern MT and Yabroff KR. Prevalence of outpatient cancer treatment in the United States: estimates from the medical panel expenditures survey (MEPS). *Cancer Invest* 2008; 26(6): 647–651.

15. Lau AC, Yam LY and Poon E. Hospital re-admission in patients with acute exacerbation of chronic obstructive pulmonary disease. *Resp Med* 2001; 95(11): 876–884.

16. Garcia-Aymerich J, Farrero E, Felez MA, et al. Risk factors of readmission to hospital for a COPD exacerbation: a prospective study. *Thorax* 2003; 58(2): 100–105.

17. Guan WJ, Ran PX and Zhong NS. Prevention and management of COPD in China: successes and major challenges. *Lancet Respir Med* 2016; 4(6): 428–430.

18. Wilson L, Devine EB and So K. Direct medical costs of chronic obstructive pulmonary disease: chronic bronchitis and emphysema. *Resp Med* 2000; 94(3): 204–213.

19. DeSalvo KB, Jones TM, Peabody J, et al. Health care expenditure prediction with a single item, self-rated health measure. *Med Care* 2009; 47(4): 440–447.

20. Maeng DD, Stewart WF, Yan X, et al. Use of electronic health records for early detection of high-cost, low back pain patients. *Pain Res Manag* 2015; 20(5): 234–240.

21. Lee SM, Kang JO and Suh Y. Comparison of hospital charge prediction models for colorectal cancer patients: neural network vs decision tree models. *J Korean Med Sci* 2004; 19(5): 677–681.

22. Peterson L, Byrne MM and Petersen L. A decision-theoretic approach to identifying future high-cost patients. *Med Care* 2006; 44(9): 842–849.
28. Lai YH, Wang TY and Yang HH. Network-based analysis of comorbidities: case study of diabetes mellitus. In: Wang L, Uesugi S, Ting IH, et al. (eds) Multidisciplinary social networks research. Berlin: Springer, 2015, pp. 210–222.
29. Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987; 40(5): 373–383.
30. Sundararajan V, Henderson T, Perry C, et al. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. J Clin Epidemiol 2004; 57(12): 1288–1294.
31. Farley JF, Harley CR and Devine JW. A comparison of comorbidity measurements to predict healthcare expenditures. Am J Manag Care 2006; 12(2): 110–119
32. Marjit S, Mukherji A and Sarkar S. Pareto efficiency, inequality and distribution neutral fiscal policy—an overview. In: Abergel F, Chakrabarti BK, Chakrabarti A, et al. (eds) New perspectives and challenges in econophysics and sociophysics. Cham: Springer, 2019, pp. 191–202.
33. Pocatilu P, Alecu F and Vetrici M. Measuring the efficiency of cloud computing for e-learning systems. WSEAS Trans Comput 2010; 9(1): 42–51.
34. Gondra I. Applying machine learning to software fault-proneness prediction. J Syst Software 2008; 81(2): 186–195.
35. Etemad LR and McCollam P. Predictors of high-cost managed care patients with acute coronary syndrome. Curr Med Res Opin 2005; 21(12): 1977–1984.
36. IzadShenas SA, Raahemi B, HosseinTekieh M, et al. Identifying high-cost patients using data mining techniques and a small set of non-trivial attributes. Comput Biol Med 2014; 53: 9–18.
37. Ichikawa D. How can machine-learning methods assist in virtual screening for hyperuricemia? A healthcare machine-learning approach. J Biomed Inform 2016; 64: 20–24.
38. Adam-Bourdarios C, Cowan G, Germain C, et al. The Higgs boson machine learning challenge. In: International conference on high-energy physics and machine learning, 2014, vol. 664, pp. 19–55, http://proceedings.mlr.press/v42/cowa14.pdf
39. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018; 1: 18.
40. Spathis D and Vlamos P. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. Health Inform J 2017; 25: 811–827.
41. Maciejewski ML, Liu CF and Fihn SD. Performance of comorbidity, risk adjustment, and functional status measures in expenditure prediction for patients with diabetes. Diabetes Care 2009; 32(1): 75–80.
42. Fleishman JA and Cohen J. Using information on clinical conditions to predict high-cost patients. Health Serv Res 2010; 45(2): 532–552.
43. Friedman B, Jiang HJ, Elixhauser A, et al. Hospital inpatient costs for adults with multiple chronic conditions. Med Care Res Rev 2006; 63: 327–346.
44. Pirson M, Dramaix M, Leclercq P, et al. Analysis of cost outliers within APR-DRGs in a Belgian general hospital: two complementary approaches. Health Policy 2006; 76(1): 13–25.
45. Kuo RJ, Cheng WC, Lien WC, et al. A medical cost estimation with fuzzy neural network of acute hepatitis patients in emergency room. Comput Methods Programs Biomed 2015; 122(1): 40–46.
46. Chang HY, Lee WC and Weiner J. Comparison of alternative risk adjustment measures for predictive modeling: high risk patient case finding using Taiwan's National Health Insurance claims. BMC Health Serv Res 2010; 10: 343.
### Appendix 1

#### Table 9. ICD-10 coding of comorbidities.

| Comorbidity information                  | ICD-10 coding |
|------------------------------------------|---------------|
| Chronic lower respiratory diseases       | J4            |
| AECOPD                                   | J44.1         |
| Heart disease                            | I01.9, I05, I06, I07, I08, I09, I27.9, I27.1, I26.0, I20, I21, I22, I23, I24, I25, I3, I4, I51, I52, I11.0, I11.9, I10.8, I10.9, I11.3, I13.1, I13.2, I13.9, E05.9, E05.9, M05.3+, Q24.9, O10.1, O10.3, Z50.0, Z82.4, I52.0*, I52.1*, I52.8* |
| Digestive system disease                 | K2, K3, K4, K5, K6, K7, K8, K9 |
| Hypertension                             | I10, I11, I12, I13, I14, I15 |
| Pulmonary emphysema                      | J43.0, J43.1, J43.8, J43.9, J44.8, J68.4, J98.2, J98.3, P25.0 |
| Chronic pulmonary heart disease          | I27.9         |
| Chronic bronchitis                       | J42, J44.8    |
| Other respiratory diseases               | J95, J96, J97, J98, J99 |
| Disease of the genitourinary system      | N             |
| Pneumonia                                | J12, J13, J14, J15, J16, J17, J18 |
| Cerebrovascular disease                  | I60, I61, I62, I63, I64, I65, I66, I67, I68, I69 |
| Metabolic disease                        | E7, E8, E9   |
| Coronary artery disease                  | I20.0, I21.9, I25.1, I25.4, I25.8, I24.0, I24.8, I97.9, Q24.5 |
| Diabetes                                 | E10, E11, E12, E13, E14, E71.0, E74.8 |
| Acute lower respiratory infection        | J2            |
| Cardiac failure                          | I50.0, I50.1, I50.9 |
| Spondylodynia                            | M45, M46, M47, M48, M49 |
| Other back diseases                      | M50, M51, M52, M63, M54 |
| Nervous system disease                   | G             |
| No specific symptoms                     | R             |
| Bone disease                             | M8            |
| Arthroplogosis                           | I00, L40.5+, M00.0, M00.1, M00.2, M00.8, M00.9, M05.3+, M05.8, M05.9, M06.0, M06.1, M06.2, M06.3, M06.4, M06.8, M06.9, M08.0, M08.1, M08.2, M08.3, M08.4, M08.8, M08.9, M12.0, M12.1, M12.2, M12.3, M12.4, M12.5, M12.8, M13.0, M13.1, M13.8, M13.9, M13.2, M15.9, M19.9, M45, M46, M47, M75.0 |
| Trauma                                   | S, T          |
| Lens diseases                            | H25, H26, H27, H28 |
| Asthma                                   | J45           |
| Blood disease                            | D5, D6, D7   |
| Acute upper respiratory infection        | J00, J01, J02, J03, J04, J05, J06 |
| Cancer                                   | C, Z51.1     |
| Lymphatic vessel disease                 | I70, I71, I72, I74, I77, I79.0, I79.1, I79.2, I79.8, I80, I81, I82, I83, I84, I85, I86, I87, I88, I89 |
| Pleural diseases                         | J90, J91, J92, J93, J94 |
| High cholesterol                         | E78.0        |
| Inner ear disease                        | H8           |
| Herpes zoster                            | B02          |
| Gastric ulcer                            | K25          |
| Stenocardia                              | I20.0, I20.1, I20.8, I20.9 |

(Continued)
Table 9. (Continued)

| Comorbidity information                  | ICD-10 coding                                      |
|------------------------------------------|----------------------------------------------------|
| Uarthritis                               | M10                                               |
| Malnutrition                             | E4                                                |
| Depression                               | F03, F25.1, F31.0, F31.1, F31.2, F31.3, F31.4, F31.5, F31.6, F31.7, F31.8, F32.0, F32.1, F32.3, F32.8, F32.9, F32.2, F33.0, F33.1, F33.2, F33.3, F33.4, F33.8, F33.9, F34.1, F34.1, F38.1, F41.2 |
| Pulmonary heart disease                  | BNF061                                            |
| Stroke                                   | I64, I69.4, I67.9, Z82.3                           |
| Spontaneous pneumothorax                 | J93.0, J93.1                                      |
| Chronic respiratory failure              | J96.1                                             |
| Emotional disorder                       | F3                                                |
| Anxiety disorder                         | F40, F41                                          |
| Cystic kidney disease                    | BNL040                                            |
| Secondary polycythemia                   | D75.1                                             |
| Flu                                      | J09, J10, J11                                     |

Table 10. Comorbidities’ proportion.

| Comorbidity information                  | Patients’ proportion in 2011 (%) |
|------------------------------------------|---------------------------------|
| **Chronic lower respiratory diseases**   | 71.99                           |
| AECOPD                                   | 49.68                           |
| **Heart disease**                        | 28.92                           |
| **Digestive system disease**             | 24.97                           |
| **Hypertension**                         | 18.30                           |
| **Pulmonary emphysema**                  | 17.41                           |
| **Chronic pulmonary heart disease**      | 17.10                           |
| **Chronic bronchitis**                   | 15.87                           |
| **Other respiratory diseases**           | 12.13                           |
| **Disease of the genitourinary system**  | 11.44                           |
| Pneumonia                                | 9.65                            |
| Cerebrovascular disease                  | 8.96                            |
| Metabolic disease                        | 8.69                            |
| Coronary artery disease                  | 8.14                            |
| Diabetes                                 | 5.93                            |
| Acute lower respiratory infection        | 5.73                            |
| Cardiac failure                          | 5.67                            |
| Spondylodynia                            | 5.58                            |
| Other back diseases                      | 5.46                            |
| Nervous system disease                   | 4.34                            |
| No specific symptoms                     | 4.02                            |
| Bone disease                             | 3.66                            |
| Arthrophlogosis                          | 3.29                            |
| Trauma                                   | 2.93                            |
| Lens diseases                            | 2.75                            |
| Asthma                                   | 2.72                            |

(Continued)
**Table 10.** (Continued)

| Comorbidity information                      | Patients’ proportion in 2011 (%) |
|----------------------------------------------|----------------------------------|
| Blood disease                                | 2.58                             |
| Acute upper respiratory infection            | 2.39                             |
| Cancer                                       | 2.09                             |
| Lymphatic vessel disease                     | 1.97                             |
| Pleural diseases                             | 1.34                             |
| High cholesterol                             | 1.20                             |
| Inner ear disease                            | 1.16                             |
| Herpes zoster                                | 0.90                             |
| Gastric ulcer                                | 0.78                             |
| Stenocardia                                  | 0.57                             |
| Uarthritis                                   | 0.54                             |
| Malnutrition                                 | 0.50                             |
| Depression                                   | 0.27                             |
| Pulmonary heart disease                      | 0.26                             |
| Stroke                                       | 0.19                             |
| Spontaneous pneumothorax                     | 0.19                             |
| Chronic respiratory failure                  | 0.15                             |
| Emotional disorder                           | 0.14                             |
| Anxiety disorder                             | 0.12                             |
| Cystic kidney disease                        | 0.06                             |
| Secondary polycythemia                       | 0.04                             |
| Flu                                          | 0.01                             |

The comorbidities whose proportions are greater than 11 percent are bolded, and we take these as new attributes.