Development and Validation of a Predictive Model to Identify Individuals Likely to Have Undiagnosed Chronic Obstructive Pulmonary Disease Using an Administrative Claims Database

Chad Moretz, ScD, MS; Yunping Zhou, MS; Amol D. Dhamane, BPharm, MS; Kate Burslem, MSc; Kim Saverno, PhD, RPh; Gagan Jain, PhD, MBA; Giovanna Devercelli, PhD, MBA; Shuchita Kaila, PhD; Jeffrey J. Ellis, PharmD, MS, BCPS; Gemzel Hernandez, MD, FCCP; and Andrew Renda, MD, MPH

ABSTRACT

BACKGROUND: Despite the importance of early detection, delayed diagnosis of chronic obstructive pulmonary disease (COPD) is relatively common. Approximately 12 million people in the United States have undiagnosed COPD. Diagnosis of COPD is essential for the timely implementation of interventions, such as smoking cessation programs, drug therapies, and pulmonary rehabilitation, which are aimed at improving outcomes and slowing disease progression.

OBJECTIVE: To develop and validate a predictive model to identify patients likely to have undiagnosed COPD using administrative claims data.

METHODS: A predictive model was developed and validated utilizing a retrospective cohort of patients with and without a COPD diagnosis (cases and controls), aged 40-89, with a minimum of 24 months of continuous health plan enrollment (Medicare Advantage Prescription Drug [MAPD] and commercial plans), and identified between January 1, 2009, and December 31, 2012, using Humana's claims database. Stratified random sampling based on plan type (commercial or MAPD) and index year was performed to ensure that cases and controls had a similar distribution of these variables. Cases and controls were compared to identify demographic, clinical, and health care resource utilization (HCRU) characteristics associated with a COPD diagnosis. Stepwise logistic regression (SLR), neural networking, and decision trees were used to develop a series of models. The models were trained, validated, and tested on randomly partitioned subsets of the sample (Training, Validation, and Test data subsets). Measures used to evaluate and compare the models included area under the curve (AUC); index of the receiver operating characteristics (ROC) curve; sensitivity, specificity, positive predictive value (PPV); and negative predictive value (NPV). The optimal model was selected based on AUC index on the Test data subset.

RESULTS: A total of 50,880 cases and 50,880 controls were included, with MAPD patients comprising 92% of the study population. Compared with controls, cases had a statistically significantly higher comorbidity burden and HCRU (including hospitalizations, emergency room visits, and medical procedures). The optimal predictive model was generated using SLR, which included 34 variables that were statistically significantly associated with a COPD diagnosis. After adjusting for covariates, anticholinergic bronchodilators (OR=3.336) and tobacco cessation counseling (OR=2.871) were found to have a large influence on the model. The final predictive model had an AUC of 0.754, sensitivity of 60%, specificity of 78%, PPV of 73%, and an NPV of 66%.

CONCLUSIONS: This claims-based predictive model provides an acceptable level of accuracy in identifying patients likely to have undiagnosed COPD in a large national health plan. Identification of patients with undiagnosed COPD may enable timely management and lead to improved health outcomes and reduced COPD-related health care expenditures.

J Manag Care Spec Pharm. 2015;21(12):1149-59

Copyright© 2015, Academy of Managed Care Pharmacy. All rights reserved.

What is already known about this subject

- Despite the importance of early detection, it is estimated that of the 26.8 million people with chronic obstructive pulmonary disease (COPD) in the United States, 12 million (45%) remain undiagnosed.
- Administrative claims databases have been used to develop predictive models to identify patients with undiagnosed COPD.
- To date, the most robust algorithm, using medical and pharmacy administrative claims, had a positive predictive value of 24.9% and was limited to a single health maintenance organization located in New Mexico.

What this study adds

- This study identified a claims-based, highly predictive model for detecting undiagnosed COPD patients.
- This study's predictive model had increased generalizability and improved performance measures, including positive predictive value, compared with previously developed models.

Chronic obstructive pulmonary disease (COPD) is a progressive disease of the airways characterized by persistent airflow limitation, dyspnea, cough, and sputum production. The worldwide prevalence of COPD is expected to rise because the aging population, decreased likelihood of dying from other diseases, and the burgeoning epidemic of smoking.^{1,2} In the United States, COPD-related medical costs were estimated to be \$32.1 billion in 2010, including \$29.5 billion in direct health care costs.³ Because of the irreversible nature of lung damage in COPD, early detection is important when implementing behavioral changes (e.g., smoking cessation) and initiating therapies that could relieve symptoms, reduce the frequency and severity of exacerbations, and improve health status and exercise tolerance.^{4,5}

Despite the importance of early detection, there is often a delay in the diagnosis of COPD. In the United States, it is estimated that approximately 26.8 million people have COPD and that, of these, 12 million (45%) remain undiagnosed.⁶ An analysis of data from the Third National Health and Nutrition

Examination Survey (NHANES III) showed that 63% of patients with low lung function were undiagnosed.⁷ A recent retrospective cohort study found that primary care opportunities for COPD diagnosis were missed in 85% of patients in the 5 years before a formal COPD diagnosis.⁸ There are several explanations for delayed diagnosis of COPD. Patients with undiagnosed COPD may be in the early stages of the disease with limited symptomology or may experience respiratory symptoms at a rate similar to patients without COPD.^{9,10} Further, primary care physicians may lack access to pulmonary function testing equipment (i.e., spirometry) thus inhibiting their ability to properly diagnose a symptomatic patient.^{11,12} Collectively, these factors indicate substantial opportunity for improvement in the detection of undiagnosed COPD.

Administrative claims data, including demographic and health care resource utilization (HCRU) information, may be a useful source for identifying patients with undiagnosed COPD. An algorithm derived from medical and pharmacy administrative claims data by Mapel et al. (2006) identified 19 characteristics that were statistically significantly associated with undiagnosed COPD in adults 40 years of age and older.¹³ This algorithm exhibited a high degree of accuracy in correctly identifying patients without COPD (negative predictive value [NPV] of 95.5%) but was only able to correctly identify approximately 1 in 4 patients with undiagnosed COPD (positive predictive value [PPV] of 24.9%). While this algorithm may be suitable for practical application in the single health maintenance organization within which it was developed, its lack of generalizability and low PPV are limitations.

The objective of this study was to develop and validate a predictive model to identify patients likely to have undiagnosed COPD within a health plan, based on a broad set of demographic, clinical, and HCRU characteristics found in administrative claims data.

Methods

Study Design and Data Source

A predictive model was developed and validated based on a retrospective cohort study in patients with and without COPD. The study was approved by an independent institutional review board before initiation. The study was conducted using Humana's administrative claims data from January 1, 2007, to December 31, 2012 (study period). The database includes over 12 million current and former Humana members and contains patient enrollment and inpatient and outpatient medical and pharmacy claims for fully insured Medicare Advantage Prescription Drug (MAPD) and commercial plan members. Records were linked for each patient using a unique patient identifier.

Patient Selection

Patients with (cases) and without (controls) a COPD diagnosis were identified. Cases were identified by the presence of at least

2 medical claims with a primary or secondary COPD diagnosis (*International Classification of Diseases, Ninth Revision, Clinical Modification* [ICD-9-CM]) code from January 1, 2009, to December 31, 2012 (identification period): chronic bronchitis (491.xx); emphysema (492.xx); or COPD, unspecified (496.xx). Two medical claims with a diagnosis of COPD had to occur on separate dates with the second COPD medical claim occurring within 90 days of the first. The index date for cases was defined as the first chronologically occurring date of a medical claim with a COPD diagnosis during the identification period. Patients who had a medical claim with a primary or secondary diagnosis of COPD during a 24-month period before the index date (pre-index period) were excluded from the case cohort.

Controls were identified using a multistep process. First, all patients enrolled during the study period were identified. Second, patients with at least 1 medical claim with a diagnosis of COPD during the study period were excluded. Finally, patients with less than 24 months (731 days) of continuous enrollment were excluded. The index date of the controls was defined as the 731st day of the most recent continuous enrollment period.

Patients were eligible for inclusion in the study if they were aged 40 to 89 years on the index date and had a minimum of 24 months of continuous enrollment. Patients with claims for any of the following ICD-9 CM diagnosis codes at any time during the study period were excluded: cystic fibrosis (277.0); pulmonary tuberculosis (011); or malignant neoplasms (140-172, 174-209.3, or 209.7). Stratified random sampling, based on plan type (commercial or MAPD) and year of index date, was performed to select 1 control for each case, so that the 2 cohorts (cases and controls) had similar distributions of these variables. No other variable (demographic or clinical) apart from plan type and year of index date was used to select a control for a case.

Model Development

Demographic, clinical, and HCRU characteristics were selected for predictive model development through a review of the published literature and based on input from the research team's clinical experts.¹³⁻²⁰ Sixty-one demographic, clinical, and HCRU characteristics were assessed in the 2 cohorts (cases and controls) during the pre-index period and compared using analysis of variance for continuous variables and chi-square tests for categorical variables. These characteristics were included as covariates in the model development. Demographic characteristics included age (at index date), gender, race/ethnicity, geographic region, and plan type. Clinical and HCRU characteristics included comorbidities, all-cause hospitalizations, all-cause emergency room (ER) visits, airflow and cardiopulmonary exercise tests, chest X-rays, and medications. Since smoking is known to be highly associated with COPD,¹⁴ and smoking status was not available in the claims database,

Development and Validation of a Predictive Model to Identify Individuals Likely to Have Undiagnosed Chronic Obstructive Pulmonary Disease Using an Administrative Claims Database

TABLE 1 Diagnosis Codes and Medical Services/Procedure Codes Used to Define Clinical Characteristics and Utilization

Diagnosis	ICD-9-CM Codes
Respiratory conditions	
Asphyxia	799.0x
Asthma	493.xx
Bronchitis (not chronic)	466.xx, 490.xx
Pneumonia or influenza	480.xx, 481.xx, 482.xx, 483.xx, 484.xx, 485.xx, 486.xx, 487.xx
Respiratory infection	460.xx, 461.xx, 462.xx, 463.xx, 464.xx, 465.xx
Respiratory symptoms	786.0x, 786.1x, 786.2x, 786.3x, 786.4x, 786.52
Cardiovascular conditions	
Aortic aneurysm	441.xx
Arterial circulatory disease	442.xx, 443.xx, 444.xx, 445.xx, 446.xx, 447.xx
Atherosclerosis	440.xx
Cor pulmonale	415.xx, 416.xx
Heart failure	428.xx
Hypertension	401.xx, 402.xx, 403.xx, 404.xx, 405.xx
Ischemic heart disease	410.xx, 411.xx, 412.xx, 413.xx, 414.xx
Valve disease	424.0x, 424.1x, 424.2x, 424.3x
Congenital abnormalities	
Ehlers-Danlos syndrome	756.83
Marfan syndrome	759.82
Endocrine or metabolic disorders	
Alpha 1-antitrypsin deficiency	273.4
Diabetes	250.00, 250.02, 250.10, 250.12, 250.20, 250.22, 250.30, 250.32, 250.40, 250.42, 250.50, 250.52, 250.60, 250.62, 250.70, 250.72, 250.80, 250.82, 250.90, 250.92
Miscellaneous disorders	
Cutis laxa	701.8
Depression	296.3x, 296.2x, 311.xx
Edema	782.3
Hematuria	599.7x
Human immunodeficiency virus	042.xx
Peptic ulcer	531.xx, 532.xx, 533.xx
Musculoskeletal disorders	
Osteoarthritis	715.xx
Osteoporosis	733.0x
Miscellaneous Medical Service/Procedure	
ICD-9-CM or CPT or HCPCS Codes	
Airflow test	94010, 94014, 94015, 94016, 94060, 94070, 94150, 94200, 94240, 94370, 94375, 94620, 94621, 94681, 94720
Cardiopulmonary exercise test	93015
Chest x-ray	71010, 71015, 71020, 71021, 71022, 71023, 71030, 71034, 71035
ER visit (all cause)	
Hospitalization (all cause)	
Tobacco cessation counseling	305.1, V15.82, V65.42, 649.0, 989.84, E869.4, 99406, 99407, 1034F, 4000F, G0436, G0437, S9075, S9453, C9801, C9802, G8455, D1320, G0375, G0376, G8402, G8403, G8453, G8454

CPT= Current Procedural Terminology; HCPCS= Healthcare Common Procedure Coding System; ICD-9-CM= International Classification of Diseases, Ninth Revision, Clinical Modification.

tobacco cessation counselling and medications were used as proxies. The medical and pharmacy claims-based definitions are provided in Tables 1 and 2, respectively. No data from the 60-day period before the index date were used, since this period was likely to be reflective of HCRU patterns and clinical parameters related to the initial diagnosis of COPD in the case cohort.¹³ The RxRisk-V prescription claims-based comorbidity

index and the Deyo Charlson Comorbidity Index (DCCI) were used in the development of the predictive model to adjust for overall comorbidity burden and the likelihood of 12-month mortality, respectively.²¹⁻²⁶

Three commonly used modeling approaches, stepwise logistic regression (SLR),^{27,28} decision tree (DT),^{27,29} and neural networking (NN),^{27,30} were pursued in order to develop and select

Development and Validation of a Predictive Model to Identify Individuals Likely to Have Undiagnosed Chronic Obstructive Pulmonary Disease Using an Administrative Claims Database

TABLE 2 Codes Used to Define Medications

Medication	Definition
Respiratory medications	
Phosphodiesterase 4 (PDE4) inhibitors	GPI-4: 4445
Xanthines	GPI-4: 4430
Anticholinergic bronchodilators	GPI-4: 4410
Anticholinergic beta-agonist combination agents	GPI-10: 4420990201
Inhaled corticosteroids	GPI-4: 4440
Long-acting beta-agonists	GPI-10: 4420101210, 4420102710, 4420104220, 4420105810
Long-acting beta-agonist/inhaled corticosteroid combination	GPI-10: 4420990241, 4420990270, 4420990290
Oral corticosteroids	GPI-10: 2210001000, 2210001510, 2210002000, 2210002500, 2210003000, 2210003510, 2210004010, 2210004020, 2210004000, 2210004500, 2210005010, 2210005020, 2210005000
Short-acting beta2-agonists	GPI-8: 44201010, 44201045, 44201050, 44201055, 44201060
Asthma and bronchodilator agent combination	GPI-4: 4499
Mucolytics	GPI-10: 4320000310, 4330001000, 8030300200, 8125001600, 9300000700
Oxygen	HCPCS: E0424, E0425, E0430, E0431, E0433, E0434, E0435, E0439, E0440, E0441, E0442, E0443, E0444, E1390, E1391, E1392, E1405, E1406, K0738, K0741, S8120, S8121
Nonrespiratory medications	
Antibiotics	GPI-4, GPI-10, and/or HCPCS: amoxicillin = 0120001010 amoxicillin/clavulanate = 0199000220 ampicillin = 0120002020, 0120002030, 9642664925, J0290 piperacillin/tazobactam = 0199000270, 0199000272, J2543 azithromycin = 0340001000, J0456 clarithromycin = 0350001000 doxycycline monohydrate = 0400002000 doxycycline hyclate = 0400002010, 0400002015 doxycycline calcium = 0400002020 ciprofloxacin = 0500002000, 0500002005, 0500002010, 0500002011, J0744 gatifloxacin = 0500008200, 0500008210, J1590 levofloxacin = 0500003400, 0500003411, J1956 moxifloxacin = 0500003710, 0500003712, J2280 gemifloxacin = 0500008310 telithromycin = 1621007000 sulfamethoxazole/trimethoprim = 1699000230 cefactor = 0220004000, 0220004010 cefprozil = 0220006200 cefuroxime axetil = 0220006505 cefuroxime sodium = 0220006510, 0220006511, 0220006512, 0220006513, J0697 cefdinir = 0230004000 cefixime = 0230006000 cefpodoxime = 0230006510 ceftibuten = 0230008300 cefditoren pivoxil = 0230004520 cefotaxime = 0230007510, 0230007511, J0698 ceftazidime = 0230008000, 0230008011, 0230008012, 0230008014, J0713 ceftriaxone = 0230009010, 0230009011, 0230009012, 0230009013, J0696 cefepime = 0240004010, 0240004012, J0692
Smoking cessation medications	GPI-4: 6210
Cardiovascular medications	GPI-4: 8320, 8310, 3940, 8515, 3610, 3615, 3310, 3320, 3400, 3710, 3720, 3740, 3750, 3760, 3799, 3699, 3120 AHFS: 241200
Influenza vaccination or medication to treat influenza	GPI-4 or GPI-8: 1250, 17100020, 7320001010, 9642660324 HCPCS: Q2034, Q2035, Q2036, Q2037, Q2038, Q2039 CPT: 90653, 90654, 90655, 90656, 90657, 90658, 90660, 90662, 90672, 90685, 90686, 90687, 90688
Pneumococcal vaccination	GPI-8: 17200065 CPT: 90669, 90670, 90732
Vitamin B complex	AHFS: 880800
Antidepressants	GPI-4: 5810, 5816, 5818, 5830, 5820
Leukotriene inhibitors	GPI-4: 4450
Antipsychotics	AHFS: 281608

AHFS = American Hospital Formulary Service; CPT = Current Procedural Terminology; GPI = generic product identifier; HCPCS = Healthcare Common Procedure Coding System.

Development and Validation of a Predictive Model to Identify Individuals Likely to Have Undiagnosed Chronic Obstructive Pulmonary Disease Using an Administrative Claims Database

FIGURE 1 Selection of Patients for the Case and Control Cohorts

TABLE 3 Patient Demographics

Characteristic, n (%)	Cases (n=50,880)	Controls (n=50,880)	P Value ^a
Age (years), mean (SD)	71.4 (10.1)	68.3 (9.9)	<0.001
Age (years) category			<0.001
40-49	1,722 (3.4)	3,342 (6.6)	
50-59	5,089 (10.0)	4,879 (9.6)	
60-69	13,032 (25.6)	20,406 (40.1)	
70-79	19,718 (38.8)	16,376 (32.2)	
80-89	11,319 (22.2)	5,877 (11.6)	
Gender			<0.001
Male	23,585 (46.4)	21,996 (43.2)	
Female	27,295 (53.6)	28,884 (56.8)	
Race/ethnicity			<0.001
White/Caucasian	39,836 (78.3)	38,672 (76.0)	
Black	4,823 (9.5)	5,288 (10.4)	
Hispanic	966 (1.9)	857 (1.7)	
Other	1,160 (2.3)	1,931 (3.8)	
Unknown	4,095 (8.0)	4,132 (8.1)	
Geographic region			<0.001
Northeast	893 (1.8)	1,219 (2.4)	
Midwest	12,584 (24.7)	13,915 (27.3)	
South	33,262 (65.4)	29,116 (57.2)	
West	4,141 (8.1)	6,630 (13.0)	
Plan type			1.000
Commercial	4,056 (8.0)	4,056 (8.0)	
MAPD	46,824 (92.0)	46,824 (92.0)	

^aP values for age were determined by analysis of variance; P values for categorical variables were determined by chi-square analysis.

MAPD = Medicare Advantage Prescription Drug; SD = standard deviation.

the optimal predictive model for identifying patients likely to have undiagnosed COPD. The sample was randomly partitioned into Training (40%), Validation (30%), and Test (30%) subsets, which were used for model development, validation, and comparison, respectively. Each subset had an equal number of cases and controls to create a balanced sample (50% cases and 50% controls). Predictive models developed on a balanced sample can better detect a rare class (e.g., true positive) of a target variable³¹; however, this leads to an overrepresentation of case prevalence. To adjust assessments and prediction estimates to match the real population,^{32,33} the “population prior” (i.e., the true Humana population COPD prevalence) was specified to be 2.7%. For model selection, the misclassification rate and profit/loss function in the Validation subset were used. The reciprocal of true COPD prevalence was set to the profit for true positive, and reciprocal of non-COPD prevalence was set to the profit for true negative in a decision matrix to achieve an acceptable conforming cutoff value.^{32,33} Stepwise, backward, and forward model selection methods with 0.05 statistically significant levels of entry and exit were used to develop the SLR model.^{28,34} Gini splitting rules and gradient boosting models were used to build the DT model.²⁹ Multilayer perceptron with 1 hidden layer and

^aStudy period: January 1, 2007, to December 31, 2012.

^bICD-9-CM 491.xx, 492.xx, 496.xx.

^cIdentification period: January 1, 2009, to December 31, 2012.

^dPre-index period: 24-month period before index date.

^eIndex date for Case Cohort: date of the first COPD claim.

^fICD-9-CM 277.0.

^gICD-9-CM 011.

^hICD-9-CM 140-172, 174-209.3, 209.7.

ⁱIndex date for Control Cohort: 731st day of most recent continuous enrollment period.

COPD = chronic obstructive pulmonary disease; ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modification; MAPD = Medicare Advantage Prescription Drug.

Development and Validation of a Predictive Model to Identify Individuals Likely to Have Undiagnosed Chronic Obstructive Pulmonary Disease Using an Administrative Claims Database

TABLE 4 Patient Clinical Characteristics and Health Care Utilization During Pre-index Period

Characteristic, n (%)	Cases (n = 50,880)	Controls (n = 50,880)	P Value ^a
Deyo Charlson Comorbidity Index, mean (SD)	1.73 (2.0)	0.87 (1.4)	<0.001
RxRisk-V comorbidity score, mean (SD)	6.22 (3.8)	4.47 (3.3)	<0.001
Respiratory conditions			
Asphyxia	2,269 (4.5)	604 (1.2)	<0.001
Asthma	7,084 (13.9)	2,592 (5.1)	<0.001
Bronchitis (not chronic)	12,015 (23.6)	6,555 (12.9)	<0.001
Pneumonia or influenza	4,892 (9.6)	1,777 (3.5)	<0.001
Respiratory infection	13,917 (27.4)	11,879 (23.3)	<0.001
Respiratory symptoms	22,770 (44.8)	12,615 (24.8)	<0.001
Cardiovascular conditions			
Aortic aneurysm	2,046 (4.0)	765 (1.5)	<0.001
Arterial circulatory disease	9,759 (19.2)	4,268 (8.4)	<0.001
Atherosclerosis	7,076 (13.9)	3,154 (6.2)	<0.001
Cor pulmonale	2,781 (5.5)	1,038 (2.0)	<0.001
Heart failure	9,185 (18.1)	2,914 (5.7)	<0.001
Hypertension	39,747 (78.1)	34,677 (68.2)	<0.001
Ischemic heart disease	17,791 (35.0)	9,759 (19.2)	<0.001
Valve disease	8,653 (17.0)	4,733 (9.3)	<0.001
Congenital abnormalities			
Ehlers-Danlos syndrome	1 (0.0)	2 (0.0)	0.564
Marfan syndrome	5 (0.0)	3 (0.0)	0.480
Endocrine or metabolic disorders			
Alpha 1-antitrypsin deficiency	14 (0.0)	4 (0.0)	0.018
Diabetes	19,084 (37.5)	14,802 (29.1)	<0.001
Miscellaneous disorders			
Cutis laxa	102 (0.2)	79 (0.2)	0.087
Depression	9,612 (18.9)	6,421 (12.6)	<0.001
Edema	7,995 (15.7)	4,533 (8.9)	<0.001
Hematuria	3,489 (6.9)	2,769 (5.4)	<0.001
HIV	101 (0.2)	75 (0.1)	0.050
Peptic ulcer	1,391 (2.7)	864 (1.7)	<0.001
Miscellaneous medical service/procedure			
Airflow test	4,903 (9.6)	2,039 (4.0)	<0.001
Cardiopulmonary exercise test	5,844 (11.5)	4,301 (8.5)	<0.001
Chest x-ray	26,566 (52.2)	16,622 (32.7)	<0.001
ER visit (all cause)	22,533 (44.3)	15,570 (30.6)	<0.001
Hospitalization (all cause)	15,924 (31.3)	9,682 (19.0)	<0.001

Characteristic, n (%)	Cases (n = 50,880)	Controls (n = 50,880)	P Value ^a
Miscellaneous medical service/procedure			
Tobacco cessation counseling	10,459 (20.6)	4,201 (8.3)	<0.001
Musculoskeletal disorders			
Osteoarthritis	18,473 (36.3)	14,452 (28.4)	<0.001
Osteoporosis	7,591 (14.9)	6,180 (12.1)	<0.001
Nonrespiratory medications			
Antibiotics	31,277 (61.5)	25,201 (49.5)	<0.001
Smoking cessation medications	1,261 (2.5)	408 (0.8)	<0.001
Cardiovascular medications	39,704 (78.0)	34,653 (68.1)	<0.001
Influenza vaccination or medication to treat influenza	8,367 (16.4)	8,906 (17.5)	<0.001
Pneumococcal vaccination	4,722 (9.3)	5,518 (10.8)	<0.001
Vitamin B complex	2,154 (4.2)	1,361 (2.7)	<0.001
Antidepressants	14,699 (28.9)	10,455 (20.5)	<0.001
Leukotriene inhibitors	1,605 (3.2)	699 (1.4)	<0.001
Antipsychotics	2,668 (5.2)	1,677 (3.3)	<0.001
Respiratory medications			
Any respiratory medication	20,018 (39.3)	10,451 (20.5)	<0.001
Phosphodiesterase 4 (PDE4) inhibitors	2 (0.0)	0 (0.0)	0.157
Xanthines	213 (0.4)	47 (0.1)	<0.001
Anticholinergic bronchodilators	1,207 (2.4)	145 (0.3)	<0.001
Anticholinergic beta-agonist combination agents	1,628 (3.2)	233 (0.5)	<0.001
Inhaled corticosteroids	1,196 (2.4)	418 (0.8)	<0.001
Long-acting beta-agonists	117 (0.2)	31 (0.1)	<0.001
Long-acting beta-agonist/ inhaled corticosteroid combination	3,226 (6.3)	790 (1.6)	<0.001
Oral corticosteroids	12,169 (23.9)	7,967 (15.7)	<0.001
Short-acting beta2-agonists	9,177 (18.0)	2,894 (5.7)	<0.001
Asthma and bronchodilator agent combination	46 (0.1)	22 (0.0)	0.004
Mucolytics	332 (0.7)	133 (0.3)	<0.001
Oxygen	1,865 (3.7)	402 (0.8)	<0.001
Number of total antibiotic prescriptions, mean (SD)	2.00 (2.9)	1.32 (2.3)	<0.001
Number of total oral corticosteroid prescriptions, mean (SD)	0.65 (2.2)	0.34 (1.4)	<0.001

^aP values for Deyo Charlson Comorbidity Index and RxRisk-V comorbidity score were determined by analysis of variance; P values for categorical variables were determined by chi-square analysis.

ER = emergency room; HIV = human immunodeficiency virus; SD = standard deviation.

back propagation with conjugate gradient training techniques were used to develop the NN model.³⁰ All models were cross-validated using the Training and Validation subsets.

Model performance measures were calculated using the Test subset for each model. These included the area under the curve (AUC) index of the receiver operating characteristics (ROC) curve, sensitivity, specificity, false positive rate, false negative rate, PPV, NPV, and overall classification rate. The best performing model for each modeling approach was

selected. These 3 models were then compared and the optimal predictive model selected. Model selection was based on the AUC index. The AUC index reflects the level of discrimination between individuals with and without the outcome of interest (i.e., COPD diagnosis)—the higher the AUC index, the better the discrimination. Two additional models were developed by plan type (MAPD and commercial) as sensitivity analyses. Identical methods were used; however, the population priors were based on the prevalence of COPD in the MAPD and

Development and Validation of a Predictive Model to Identify Individuals Likely to Have Undiagnosed Chronic Obstructive Pulmonary Disease Using an Administrative Claims Database

TABLE 5 Performance Measures for the Predictive Models (Test Subset)

Model	AUC	Sensitivity ^b (%)	Specificity ^c (%)	FP Rate ^d (%)	FN Rate ^e (%)	PPV ^f (%)	NPV ^g (%)	Overall Classification Rate ^h (%)
NN	0.757	63.7	75.0	25.0	36.3	71.8	67.4	69.3
SLR ^a	0.754	60.0	78.3	21.7	40.1	73.4	66.2	69.1
DT	0.729	62.1	73.4	26.6	37.9	70.0	65.9	67.7

^aThe SLR model was selected as the optimal model.

^bSensitivity (true positive rate): $TP/(TP+FN)$; the proportion of all patients with a COPD diagnosis correctly predicted by the model to have a COPD diagnosis.

^cSpecificity (true negative rate): $TN/(TN+FP)$; the proportion of all patients without a COPD diagnosis correctly predicted by the model to not have a COPD diagnosis.

^dFP Rate: Type I Error (false alarm); $1 - \text{specificity}$.

^eFN Rate: Type II Error (miss); $1 - \text{sensitivity}$.

^fPPV: $TP/(TP+FP)$; the proportion of patients predicted by the model to have a COPD diagnosis who truly had a COPD diagnosis.

^gNPV: $TN/(TN+FN)$; the proportion of patients predicted by the model not to have a COPD diagnosis who truly did not have a COPD diagnosis.

^hOverall Classification Rate: $(TP+TN)/\text{Total Observations}$; proportion of model predictions that were accurate.

AUC = area under the curve; COPD = chronic obstructive pulmonary disease; DT = decision tree; FN = false negative; FP = false positive; NN = neural network; NPV = negative predictive value; PPV = positive predictive value; SLR = stepwise logistic regression; TN = true negative; TP = true positive.

commercial populations (4.2% and 0.55%, respectively). Data analyses were conducted using SAS Enterprise Guide version 4.3 and Enterprise Miner 12.1 (SAS Institute, Cary, NC).

Results

A total of 50,880 patients met the selection criteria for the cases, and 1,802,705 patients met the selection criteria to serve as a potential control. Stratified random sampling, based on index year and plan type, was used to select an equal number of controls (50,880) to the number of cases (Figure 1). Statistically significant differences between cases and controls were observed across all demographic characteristics including age, race/ethnicity, geographic region, and gender (Table 3; $P < 0.001$). Mean age was statistically significantly higher for the cases compared with the controls (71.4 years vs. 68.3 years, respectively). In both cases and controls, the majority of patients was female, of white/Caucasian ethnicity, and geographically located in the South.

Clinical characteristics and HCRU are described in Table 4. DCCI and RxRisk-V scores were statistically significantly higher for cases versus controls ($P < 0.001$), indicating a higher comorbidity burden. A higher proportion of cases had comorbid respiratory and cardiovascular conditions, as well as diabetes, compared with controls ($P < 0.001$). More use of medical services was observed in cases compared with controls (Table 4). A higher proportion of cases had at least 1 all-cause hospitalization or ER visit, and tobacco cessation counseling and oxygen therapy were more common in cases compared with controls ($P < 0.001$). The proportion of patients with at least 1 claim for any respiratory-related medication or smoking cessation medication was higher in cases compared with controls ($P < 0.001$).

Performance measures for the SLR, DT, and NN models are provided in Table 5. The SLR and NN models performed

similarly in terms of AUC (0.754 and 0.757, respectively), followed by the DT model (0.729). The SLR and NN models also performed similarly on other performance measures, although the SLR model had slightly higher specificity (78.3% vs. 75.0%) and PPV (73.4% vs. 71.8%). The SLR model was selected as the optimal predictive model, since it performed similarly to the NN model (based on AUC) and was considered to be a more widely applicable and understood method for predictive modeling. Another advantage of the SLR model is that it allows the impact of individual risk factors to be studied. The SLR model had a sensitivity of 60%, specificity of 78%, PPV of 73%, and NPV of 66% when applied to the Test data subset.

The optimal predictive model is described in Table 6. The model contained 34 variables that were statistically significantly associated with COPD. After adjusting for covariates, anticholinergic bronchodilators (odds ratio [OR] = 3.336), tobacco cessation counseling (OR = 2.871), anticholinergic beta-agonist combination agents (OR = 2.675), and smoking cessation medication (OR = 2.317) were found to make a large contribution to the model. Other variables included demographic characteristics (region, age, and gender); comorbidities (DCCI, heart failure, asthma, aortic aneurysm, pneumonia or influenza, asphyxia, atherosclerosis, bronchitis [not chronic], respiratory symptoms, arterial circulatory disease, ischemic heart disease, depression, diabetes, and hypertension); and HCRU (all-cause hospitalization, cardiopulmonary exercise test, tobacco cessation counseling, oxygen, and medications [long-acting beta-agonists/inhaled corticosteroids combination, short-acting beta2-agonist, respiratory medications, antidepressants, number of total oral corticosteroid prescriptions, cardiovascular medications, and oral corticosteroid, vaccination/medication to treat influenza, pneumococcal vaccination, and RxRisk-V]).

Development and Validation of a Predictive Model to Identify Individuals Likely to Have Undiagnosed Chronic Obstructive Pulmonary Disease Using an Administrative Claims Database

TABLE 6 Parameter Estimates for the Stepwise Logistic Regression Model (Optimal Predictive Model)

Parameter	Odds Ratio	95% CI		Estimate	Standard Error	Wald Chi-square	P Value
		LL	UL				
Anticholinergic bronchodilators	3.336	2.354	4.727	1.2047	0.1476	66.62	<0.001
Tobacco cessation counseling	2.871	2.670	3.086	1.0545	0.0350	909.41	<0.001
Anticholinergic beta-agonist combination agents	2.675	2.122	3.372	0.9839	0.1201	67.07	<0.001
Smoking cessation medications	2.317	1.964	2.734	0.8404	0.1003	70.27	<0.001
Geographic region							
South vs. West	2.020	1.916	2.129	0.7031	0.0383	337.35	<0.001
Midwest vs. West	1.635	1.571	1.701	0.4915	0.0415	140.36	<0.001
Northeast vs. West	1.102	1.085	1.121	0.0976	0.0851	1.31	0.252
Long-acting beta-agonist/inhaled corticosteroid combination	1.804	1.653	1.969	0.5901	0.0755	61.10	<0.001
Oxygen	1.724	1.531	1.941	0.5444	0.1113	23.92	<0.001
Short-acting beta2-agonist	1.618	1.538	1.702	0.4812	0.0539	79.76	<0.001
Heart failure	1.593	1.531	1.657	0.4656	0.0435	114.43	<0.001
Respiratory medications	1.525	1.452	1.602	0.4221	0.0593	50.65	<0.001
Asthma	1.470	1.417	1.524	0.3850	0.0484	63.28	<0.001
Aortic aneurysm	1.394	1.327	1.463	0.3318	0.0749	19.63	<0.001
Pneumonia or influenza	1.385	1.339	1.433	0.3257	0.0534	37.17	<0.001
Asphyxia	1.380	1.302	1.463	0.3221	0.0921	12.22	<0.001
Atherosclerosis	1.348	1.315	1.382	0.2989	0.0425	49.36	<0.001
Bronchitis (not chronic)	1.313	1.291	1.335	0.2721	0.0318	73.27	<0.001
Respiratory symptoms	1.308	1.290	1.326	0.2683	0.0267	101.03	<0.001
Arterial circulatory disease	1.267	1.245	1.290	0.2370	0.0379	39.16	<0.001
Ischemic heart disease	1.218	1.204	1.233	0.1974	0.0306	41.52	<0.001
Antidepressants	1.179	1.166	1.192	0.1649	0.0339	23.65	<0.001
Depression	1.163	1.150	1.176	0.1509	0.0369	16.67	<0.001
Deyo Charlson Comorbidity Index	1.126	1.123	1.128	0.1183	0.0101	136.08	<0.001
RxRisk-V	1.057	1.057	1.058	0.0559	0.0056	98.96	<0.001
Age	1.044	1.044	1.044	0.0432	0.0013	1202.28	<0.001
Number of total oral corticosteroid prescriptions	1.032	1.032	1.033	0.0319	0.0078	16.81	<0.001
Cardiovascular medications	0.857	0.848	0.867	-0.1543	0.0364	17.96	<0.001
Diabetes	0.848	0.840	0.856	-0.1647	0.0298	30.61	<0.001
Hypertension	0.834	0.824	0.844	-0.1816	0.0323	31.53	<0.001
Influenza vaccination or medication to treat influenza	0.823	0.813	0.832	-0.1954	0.0298	42.94	<0.001
Hospitalization (all cause)	0.797	0.786	0.808	-0.2268	0.0305	55.11	<0.001
Cardiopulmonary exercise test	0.792	0.777	0.807	-0.2332	0.0408	32.63	<0.001
Pneumococcal vaccination	0.759	0.744	0.775	-0.2755	0.0375	54.01	<0.001
Female	0.730	0.719	0.740	-0.3152	0.0233	182.87	<0.001
Oral corticosteroid	0.689	0.660	0.719	-0.3729	0.0591	39.76	<0.001
Intercept	0.016	0.007	0.034	-4.1657	0.0961	1880.79	<0.001

CI = confidence interval; LL = lower limit; UL = upper limit.

Sensitivity Analyses

The population used to develop the MAPD plan model was composed of 46,824 cases and an equal number of controls. The SLR and NN models performed similarly in terms of AUC (0.764 and 0.766, respectively). The SLR model was selected as the final model for similar reasons to those applied for the primary model selection. The SLR model contained 43 variables and had a sensitivity of 66.9%, specificity of 72.9%, PPV of 71.2%, and NPV of 68.8%.

The population used to develop the commercial plan model was composed of 4,056 cases and an equal number of controls.

The SLR model performed best in terms of AUC (0.810) and was selected as the final model. This model had 23 variables and had a sensitivity of 53.1%, specificity of 89.8%, PPV of 83.9%, and NPV of 65.7%. Findings from the sensitivity analyses confirmed the selection of SLR as the optimal predictive modeling approach.

Discussion

This study showed that a claims-based predictive model provides a method to identify those patients likely to have undiagnosed COPD within a national health plan in the United States.

Development and Validation of a Predictive Model to Identify Individuals Likely to Have Undiagnosed Chronic Obstructive Pulmonary Disease Using an Administrative Claims Database

The PPV of 73.4% implies that for every 4 patients identified by the model as likely to have COPD, approximately 3 of them will have undiagnosed COPD based on the operational definition of a COPD diagnosis used in this study. While higher rates of sensitivity and specificity may be preferred for diagnostic tools,³⁵ the levels of sensitivity and specificity provided by the SLR predictive model in this study may be of value to payers as a screening tool, allowing them to target appropriate health care interventions to health plan members who may have undiagnosed COPD with a reasonable level of accuracy.

The strongest predictors in the optimal model were the use of anticholinergic bronchodilators and tobacco cessation counseling. Since smoking is the primary cause of COPD,^{14,36,37} the impact of tobacco cessation counseling, as well as the use of smoking cessation medications was expected. Of the comorbidities that were evaluated, heart failure had the highest OR (1.593) in the predictive model. Heart failure and COPD may have some similar symptoms that can mask the diagnosis of either disease, and there is evidence to support screening for the presence of COPD in patients that have been diagnosed with heart failure.³⁸

The predictive model developed in this study compares favorably to other published predictive models aimed at identifying those with undiagnosed COPD.^{13,39-41} The algorithm, using medical and pharmacy administrative claims and developed by Mapel et al., had a sensitivity of 60.5%, specificity of 82.1%, PPV of 24.9%, and NPV of 95.5%.¹³ It is noteworthy that the NPV was higher in the Mapel et al. model, while the PPV was higher in the model developed in this study. PPV is the likelihood that a patient actually has undiagnosed COPD when the model returns a positive result and, as such, determines the model's degree of efficiency for screening those with a positive result.

A possible explanation for differences in performance metrics between the current model and the Mapel et al. algorithm is the use of different independent variables. For example, the Mapel et al. study used a composite variable to indicate respiratory medication use.¹³ This variable (Respiratory Rx) included antimuscarinics, antispasmodics, sympathomimetic (adren-ergic) agents, antitussives, expectorants, mucolytic agents, and adrenal agents, as categorized by the American Hospital Formulary Service. Among the independent variables in the Mapel et al. model, the Respiratory Rx variable made a significant contribution, based on F value. Of note, this composite variable contained antitussives and expectorants, which are used for common respiratory illnesses, as well as COPD symptoms such as cough and sputum production. Such a non-specific respiratory medication use variable may explain the high number of false positives (6,770) relative to true positives (2,240) and thus the low PPV (24.9%). By comparison, the current study's predictive model contained variables at the drug class level (e.g., long-acting beta-agonists/inhaled corticoste-

roid combinations) as well as a composite respiratory medications variable, which was defined more restrictively (Table 1) than in the Mapel et al. study.¹³ The use of drug class-level variables may allow for a better distinction of medications used primarily for chronic airway disease versus those used to treat symptoms of acute viral or bacterial respiratory infections.

Several studies have demonstrated improvements in lung function, dyspnea, quality of life, and reduced risk of exacerbations through earlier pharmacologic intervention and serve as the foundation of the recommendations found in the Global Strategy for the Diagnosis, Management and Prevention of COPD (GOLD).¹⁴ The predictive model in the current study may allow a payer or managed care pharmacy to identify patients likely to have undiagnosed COPD and encourage the clinical screening of COPD through spirometry. This model creates the opportunity for proactive diagnosis as well as for more targeted and timely support. Managed care pharmacy is typically limited to providing disease management and education programs, such as smoking cessation counseling, medication therapy management, and transitions of care programs for established COPD patients. In addition, managed care pharmacy may also play a vital role in appropriate treatment and management of COPD for "new" patients identified through this predictive model.

Limitations

There are several limitations associated with this study that should be considered when interpreting the results. As is common with all administrative claims databases, the database used for this study lacked some clinical parameters that have been shown to be strongly associated with COPD, such as smoking status. Lack of a smoking status variable could be expected to reduce the predictive ability of the model. Tobacco cessation counseling and use of smoking cessation medications were included as proxies in order to address this potential limitation.^{13,39-41} Administrative claims may contain coding errors of omission and commission and incomplete claims information. COPD diagnosis was determined using claims with ICD-9-CM diagnosis codes indicative of COPD. This operational classification may have resulted in misclassification in some cases, since airflow testing (e.g., FEV1) results were not available to confirm COPD diagnosis. The patient population included in this study may not be representative of the general U.S. population, thus, limiting the generalizability of the predictive model. Finally, the predictive model will only be able to identify patients likely to have COPD. Further testing, including spirometric and symptom evaluation, will be required to confirm the clinical diagnosis of COPD.

Conclusions

This claims-based predictive model provides an acceptable level of accuracy in identifying patients likely to have undiag-

Development and Validation of a Predictive Model to Identify Individuals Likely to Have Undiagnosed Chronic Obstructive Pulmonary Disease Using an Administrative Claims Database

nosed COPD in a large health plan that includes commercial and MAPD members. Identification of patients with a high risk of having undiagnosed COPD may help in the timely referral for a diagnostic evaluation and management of the disease and possibly lead to improved health outcomes.

Authors

CHAD MORETZ, ScD, MS, is Research Lead; YUNPING ZHOU, MS, is Research Scientist; and JEFFREY J. ELLIS, PharmD, MS, BCPS, is Research Consultant, Comprehensive Health Insights, Louisville, Kentucky. AMOL D. DHAMANE, BPharm, MS, is Manager; KATE BURSLEM, MSc, is Associate Director; GAGAN JAIN, PhD, MBA, is Associate Director; GIOVANNA DEVERCELLI, PhD, MBA, is Executive Director; and SHUCHITA KAILA, PhD, is Associate Director, Health Economics & Outcomes Research, Boehringer Ingelheim Pharmaceuticals, and GEMZEL HERNANDEZ, MD, FCCP, is Director, Therapeutic Operations, Respiratory, Boehringer Ingelheim Pharmaceuticals, Ridgefield, Connecticut. KIM SAVERNO, PhD, RPh, is Director of Pharmacoeconomics, Vector Oncology, Memphis, Tennessee, and ANDREW RENDA, MD, MPH, is Director, Diabetes and Heart Disease Consumers, Retail Strategy and Execution, Humana, Louisville, Kentucky.

AUTHOR CORRESPONDENCE: Chad Moretz, ScD, MS, 545 Wilson Valley Dr., Marion, NC 28752. Tel.: 828.527.9033; E-mail: dmoretz@humana.com.

DISCLOSURES

This work was supported by Boehringer Ingelheim Pharmaceuticals (BIPI) and Humana. Burslem, Devercelli, Dhamane, Hernandez, Jain, and Kaila are employees of BIPI. Renda is an employee of Humana. Ellis, Moretz, and Zhou are employees of Comprehensive Health Insights, a fully owned subsidiary of Humana. Saverno was an employee of Comprehensive Health Insights when this research was conducted. The authors meet criteria for authorship as recommended by the International Committee of Medical Journal Editors, were fully responsible for all content and editorial decisions, and were involved at all stages of manuscript development.

Moretz, Burslem, and Renda created the study design, along with Zhou, Dhamane, Saverno, and Jain. Zhou was primarily responsible for data collection, assisted by Moretz and Saverno. Zhou took the lead in data interpretation, assisted by the other authors. The manuscript was written and revised by Moretz, Devercelli, Kaila, and Ellis, assisted by the other authors.

REFERENCES

1. Lopez AD, Shibuya K, Rao C, et al. Chronic obstructive pulmonary disease: current burden and future projections. *Eur Respir J*. 2006;27(2):397-412.
2. Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. *Lancet*. 2007;370(9589):765-73.
3. Ford ES, Murphy LB, Khavjou O, Giles WH, Holt JB, Croft JB. Total and state-specific medical and absenteeism costs of COPD among adults aged ≥ 18 years in the United States for 2010 and projections through 2020. *Chest*. 2015;147(1):31-45. Available at: <http://journal.publications.chestnet.org/article.aspx?articleid=1891096>. Accessed October 29, 2015.
4. Wilkinson TM, Donaldson GC, Hurst JR, Seemungal TA, Wedzicha JA. Early therapy improves outcomes of exacerbations of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2004;169(12):1298-303.
5. Doherty DE, Chapman KR, Martinez FJ, Belfer MH. The value of early diagnosis for effective management of chronic obstructive pulmonary disease. *J Fam Pract*. 2007;56(10 Suppl Value):S1-S24.
6. Mannino DM, Homa DM, Akinbami LJ, et al. Chronic obstructive pulmonary disease surveillance—United States, 1971-2000. *MMWR Surveill Summ*. 2002;51(6):1-16. Available at: <http://www.cdc.gov/mmwr/preview/mmwrhtml/ss5106a1.htm>. Accessed October 20, 2015.
7. Mannino DM. COPD: epidemiology, prevalence, morbidity and mortality, and disease heterogeneity. *Chest*. 2002;121(5 Suppl):121S-126S.
8. Jones RC, Price D, Ryan D, et al.; Respiratory Effectiveness Group. Opportunities to diagnose chronic obstructive pulmonary disease in routine care in the UK: a retrospective study of a clinical cohort. *Lancet Respir Med*. 2014;2(4):267-76.
9. Lindberg A, Bjerg A, Ronmark E, Larsson LG, Lundback B. Prevalence and underdiagnosis of COPD by disease severity and the attributable fraction of smoking: report from the Obstructive Lung Disease in Northern Sweden Studies. *Respir Med*. 2006;100(2):264-72.
10. Hill K, Goldstein RS, Guyatt GH, et al. Prevalence and underdiagnosis of chronic obstructive pulmonary disease among patients at risk in primary care. *CMAJ*. 2010;182(7):673-78.
11. Walters JA, Hansen EC, Johns DP, Blizzard EL, Walters EH, Wood-Baker R. A mixed methods study to compare models of spirometry delivery in primary care for patients at risk of COPD. *Thorax*. 2008;63(5):408-14.
12. Dales RE, Vandemheen KL, Clinch J, Aaron SD. Spirometry in the primary care setting: influence on clinical diagnosis and management of airflow obstruction. *Chest*. 2005;128(4):2443-47.
13. Mapel DW, Frost FJ, Hurley JS, et al. An algorithm for the identification of undiagnosed COPD cases using administrative claims data. *J Manag Care Pharm*. 2006;12(6):457-65. Available at: http://amcp.org/data/jmcp/research_458-465.pdf.
14. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global strategy for diagnosis, management, and prevention of COPD. January 2015. Available at: <http://www.goldcopd.org/guidelines-global-strategy-for-diagnosis-management.html>. Accessed October 29, 2015.
15. Mapel DW, Dutro MP, Marton JP, Woodruff K, Make B. Identifying and characterizing COPD patients in U.S. managed care. A retrospective, cross-sectional analysis of administrative claims data. *BMC Health Serv Res*. 2011;11:43.
16. Mannino DM, Thorn D, Swensen A, Holguin F. Prevalence and outcomes of diabetes, hypertension and cardiovascular disease in COPD. *Eur Respir J*. 2008;32(4):962-69.
17. Bartlett JG, Sethi S. Management of infection in exacerbations of chronic obstructive pulmonary disease. UpToDate.com. 2013.
18. Mosenifar Z, Kamangar N, Nikhanj NS, Harrington A. Chronic obstructive pulmonary disease. Medscape.com. May 28, 2013. Available at: <http://emedicine.medscape.com/article/297664-overview#aw2aab6b2b4aa>. Accessed October 21, 2015.
19. Sharafkhaneh A, Petersen NJ, Yu HJ, Dalal AA, Johnson ML, Hanania NA. Burden of COPD in a government health care system: a retrospective observational study using data from the US Veterans Affairs population. *Int J Chron Obstruct Pulmon Dis*. 2010;5:125-32.
20. Watson L, Vonk JM, Löfdahl CG, et al.; European Respiratory Society Study on Chronic Obstructive Pulmonary Disease. Predictors of lung function and its decline in mild to moderate COPD in association with gender: results from the Euroscop study. *Respir Med*. 2006;100(4):746-53.

Development and Validation of a Predictive Model to Identify Individuals Likely to Have Undiagnosed Chronic Obstructive Pulmonary Disease Using an Administrative Claims Database

21. Sloan KL, Sales AE, Liu CF, et al. Construction and characteristics of the RxRisk-V: a VA-adapted pharmacy-based case-mix instrument. *Med Care*. 2003;41(6):761-74.
22. Sales AE, Liu CF, Sloan KL, et al. Predicting costs of care using a pharmacy-based measure risk adjustment in a veteran population. *Med Care*. 2003;41(6):753-60.
23. Fishman PA, Goodman MJ, Hornbrook MC, Meenan RT, Bachman DJ, O'Keeffe Rosetti MC. Risk adjustment using automated ambulatory pharmacy data: the RxRisk model. *Med Care*. 2003;41(1):84-99.
24. Farley JF, Harley CR, Devine JW. A comparison of comorbidity measurements to predict healthcare expenditures. *Am J Manag Care*. 2006;12(2):110-19.
25. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol*. 1992;45(6):613-19.
26. Klabunde CN, Potosky AL, Legler JM, Warren JL. Development of a comorbidity index using physician claims data. *J Clin Epidemiol*. 2000;53(12):1258-67.
27. Hand DJ. Classifier technology and the illusion of progress. *Stat Sci*. 2006;21(1):1-15.
28. Sarma KS. *Predictive Modeling with SAS Enterprise Miner: Practical Solutions for Business Applications*. 2nd ed. Cary, NC: SAS Institute; 2013.
29. De Ville B. *Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner*. Cary, NC: SAS Institute; 2006.
30. Matignon R. *Neural Network Modeling Using SAS Enterprise Miner*. Bloomington, IN: AuthorHouse; 2005.
31. SAS Institute Inc. *Getting Started with SAS Enterprise Miner 7.1*. Cary, NC: SAS Institute Inc.; 2011. Available at: <https://support.sas.com/documentation/cdl/en/emgsj/64144/PDF/default/emgsj.pdf>. Accessed October 29, 2015.
32. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd ed. New York: Springer; 2009.
33. Matignon R. *Data Mining Using SAS Enterprise Miner*. Hoboken, NJ: John Wiley & Sons; 2007.
34. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer; 2010.
35. Bland, M. *An Introduction to Medical Statistics*. 3rd ed. New York: Oxford University Press; 2000.
36. U.S. Department of Health and Human Services. *The Health Consequences of Smoking: 50 Years of Progress. A Report of the Surgeon General*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health; 2014. Available at: <http://www.surgeongeneral.gov/library/reports/50-years-of-progress/full-report.pdf>. Accessed October 29, 2015.
37. Kohansal R, Martinez-Camblor P, Agusti A, Buist AS, Mannino DM, Soriano JB. The natural history of chronic airflow obstruction revisited: an analysis of the Framingham offspring cohort. *Am J Respir Crit Care Med*. 2009;180(1):3-10.
38. Mascarenhas J, Azevedo A, Bettencourt P. Coexisting chronic obstructive pulmonary disease and heart failure: implications for treatment, course and mortality. *Curr Opin Pulm Med*. 2010;16(2):106-11.
39. Mapel DW, Petersen H, Roberts MH, Hurley JS, Frost FJ, Marton JP. Can outpatient pharmacy data identify persons with undiagnosed COPD? *Am J Manag Care*. 2010;16(7):505-12.
40. Smidth M, Sokolowski I, Kærsvang L, Vedsted P. Developing an algorithm to identify people with Chronic Obstructive Pulmonary Disease (COPD) using administrative data. *BMC Med Inform Decis Mak*. 2012;12:38.
41. Belleudi V, Agabiti N, Kirchmayer U, et al. Definition and validation of a predictive model to identify patients with chronic obstructive pulmonary disease (COPD) from administrative databases. *Epidemiol Prev*. 2012;36(3-4):162-71.