IDENTIFICATION OF PREDICTIVE CLINICAL BIOMARKERS FOR CHRONIC OBSTRUCTIVE PULMONARY DISEASE USING REAL WORLD EVIDENCE

Gang (Michael) Liu, Patrudu Makena, Kyung Soo Hong, Eric Scott and GL Prasad
RAI Services Company, 401 N Main Street, Winston-Salem, NC 27101, USA

RAI SERVICES COMPANY

Abstract

Identification of predictive biomarkers and quantification of individual risk for developing smoking-associated diseases such as Chronic Obstructive Pulmonary Disease (COPD) aids in evaluating and predicting the health effects from tobacco products. This study aimed to identify predictive biomarker(s) for COPD in U.S. smokers by leveraging a Real World Evidence (RWE) approach. We performed a retrospective analysis of smokers' electronic health records prior to COPD diagnosis dates from the Explorys database available from IBM Watson Health. Electronic health records from 181,250 smokers with COPD and 2.2 million smokers without COPD were analyzed for 75 selected health measures and 900 derived clinical features based on the selected biomarkers at the subject level. A computational model built around RWE data predicted development of COPD with 76% precision (true positive rate) and 0.801 Area Under the Receiver Operating Characteristics Curve on the subject level outcome. A set of 32 biomarkers (e.g., coagulation tissue factor, cholesterol, erythrocytes) and 96 clinical features (different ways a given biomarker is reported or analyzed) were identified to have predictive power in modeling development of COPD. Taken collectively, top clinical biomarkers identified to have high predictability were platelets, cholesterol in HDL, coagulation tissue factor, age and leukocytes. These findings from RWE data help in building an individual risk scoring model to estimate the likelihood of smokers developing COPD.

Introduction

- No epidemiological data are available for next generation tobacco products (NGP) such as electronic nicotine delivery systems (ENDS) and tobacco heated products (THP).
- Quantification of the individual health risks of these NGP products is a challenging task without epidemiological data.
- Development of a biomarker-based disease risk scoring model might be useful for evaluating and predicting the health effects from switching to NGP use in cigarette smokers.
- Real world evidence refers to observational data in electronic health records, rather than randomized clinical trials, and it plays an increasingly important role in health care decisions.
- Explorys database from IBM contains longitudinal electronic health records from over 500,000 COPD patients and 4.4 million smoking patients.

Objectives

- To identify predictive clinical biomarkers for COPD in U.S. smokers using Explorys health records.
- To develop a biomarker-based computer model for predicting the development of COPD in smokers.

Methods

Cohort Selection

- Data Source:
  - Explorys, a database of electronic health records from IBM, contains longitudinal data captured from 39 health partners, 400 acute care facilities, 400,000 providers and physicians, and 55 million patients.
- Inclusion and Exclusion criteria:
  - Smokers: patients with medical record of tobacco smoking including cigarette, cigar and pipe smoking
  - COPD patients: patients with ICD codes 491 and J44
  - The retrospective data prior to the defined index date (COPD Diagnosis [Dx]) were analyzed
- Cohorts and Clinical Biomarker Data:
  - COPD smokers and Non-COPD smokers were selected as two cohorts for analysis
  - 75 clinical variables were selected based on their potential relationships to COPD and smoking

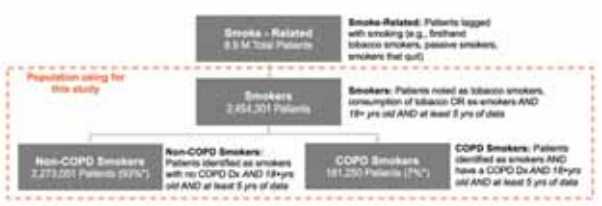
Exploratory Analyses

- Distribution Analysis
  - The distribution of gender, top comorbidities of COPD and age distribution were analyzed in the COPD smokers and non-COPD smokers cohorts.
- Outlier Analysis
  - An optimal percentile value was identified beyond which the values were classified as outliers and these values were removed to meet normal distribution assumptions.
- Biomarker Variables
  - Age, gender, and absolute values of 75 biomarker measurements
  - Additional 13 variables were generated for each biomarker including minimum, maximum, average biomarker values, etc.

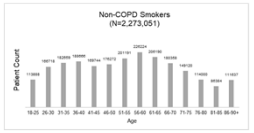
Correlation Analysis and Predictive Modeling

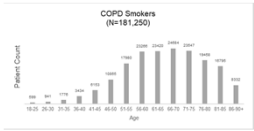
- Correlation Analysis
  - Pearson Correlation Coefficient was computed for each variable to describe its association with COPD Dx
- Predictive Model
  - Down-sampling the size of non-COPD smokers for population balance between two cohorts
  - Elimination of redundant features based on multicollinearity
  - Feature selection to select performance-limiting biomarker variables
  - Train and validation of predictive model for prediction of COPD risk using linear regression model

Results: Cohort Selection and Age/Gender Distribution



	COPD Smokers (N=181,250)	Non-COPD Smokers (N=2,273,051)
Female Prevalence	47%	50%
Male Prevalence	53%	50%





- In Exploratory Analyses, Non-COPD Smokers (~2.2 million) and COPD Smokers (~0.18 million) were used as two cohorts.
- No observable differences were found between the prevalence of females and males in the two cohorts.
- COPD smokers show a negative skew in the age distribution curve with a peak for the 66-70 years age group.

Results: Exploratory Analyses

Comorbidity	COPD Smokers (N=181,250)	Non-COPD Smokers (N=2,273,051)
Hypertension	97% (n=175,681)	43% (n=986,701)
Asthma	89% (n=161,534)	12% (n=262,502)
Hyperlipemia	67% (n=121,659)	36% (n=828,687)
Coronary Atherosclerosis	44% (n=79,822)	13% (n=303,877)
Shortness of Breath	40% (n=73,392)	14% (n=326,759)
Upper Respiratory Infections	40% (n=73,221)	32%(n=732,778)
Esophageal Reflux	38% (n=68,743)	23% (n=517,565)
Malaise and Fatigue	33% (n=60,042)	24% (n=547,518)
Hypercholesteremia	28% (n=51,113)	16% (n=369,935)
Congestive Heart Failure	21% (n=37,771)	6% (n=177,057)
Pneumonia	26% (n=47,381)	8% (n=181,844)

Diagnosis (Dx)	Years (yr) Prior to COPD Diagnosis						Percent of Diagnosis that happen within 1 yr of Dx
	5+	4-5	3-4	2-3	1-2	0-1	
Asthma	14,107	5,683	6,665	8,315	10,526	116,238	72.0%
Pneumonia	7,039	2,845	3,470	4,514	6,043	23,470	49.5%
Congestive Heart Failure	6,712	2,725	3,126	4,164	5,125	15,919	42.1%
Shortness of Breath	13,938	5,520	6,665	8,227	9,965	29,077	40%
Atrial Fibrillation	6,879	2,479	2,885	3,418	3,992	10,080	33.9%
Acute Bronchitis	11,415	3,598	3,958	4,645	5,510	13,023	30.9%
Obesity	8,970	3,197	3,776	4,573	5,430	10,486	28.8%
Upper Respiratory Infections	16,802	6,918	7,967	9,550	11,548	20,436	27.9%
Anemia	11,382	3,993	4,636	5,602	6,433	12,279	27.7%
Coronary Atherosclerosis	19,892	7,943	8,739	10,227	11,173	21,848	27.4%
Esophageal Reflux	18,880	6,233	7,140	8,460	9,311	18,719	27.2%
Malaise and Fatigue	16,248	5,434	6,229	7,509	8,658	15,964	26.6%
Hypertension	53,451	17,766	17,831	19,910	21,074	45,649	26.0%
Hyperlipidemia	40,248	12,135	12,392	14,058	14,887	27,939	23.0%
Hypothyroidism	12,039	3,355	3,561	4,195	4,349	7,940	22.4%
Hypercholesterolemia	18,559	4,857	5,114	5,750	6,178	10,655	20.8%

- Chi-square analysis identified hypertension, asthma, hyperlipidemia, coronary atherosclerosis, shortness of breath, upper respiratory infections, esophageal reflux, malaise and fatigue, and hypercholesterolemia as significantly different between non-COPD smokers and COPD smokers.
- Time series analysis identified asthma, pneumonia, congestive heart failure and shortness of breath are the top 4 comorbidities occurring within 1 year prior to COPD Diagnosis.

Results: Correlation Analysis

Statistical Variables	Summary
Duration of First Observation to Index	Duration in months from index date to first observation for each biomarker at patient level
Duration of Last Observation to Index	Duration in months from index date to last observation for each biomarker at patient level
Frequency	Number of times each biomarker observation was taken at patient level
Min. of Values	Minimum value for each biomarker observation at patient level
Max. of Values	Maximum value for each biomarker observation at patient level
Average of Values	Average of all observation values for each biomarker at patient level
Standard Deviation of Values	Standard deviation of all observation values for each Biomarker at patient level
Median	Median of all observation values for each biomarker at patient level
Delta (first observation to last observation values)	The difference or change from first to last observation value for each biomarker at patient level
Observation Occurrence	Binary flag to mark occurrence of observation at patient level
Average Cohort Value	Average of biomarker values at a cohort level
Median Cohort Value	Median of biomarker values at a cohort level
Duration of Comorbidities to Index	Duration in months from index date to first diagnosis date of 4 comorbidities – asthma, shortness of breath, pneumonia and congestive heart failure – at patient level

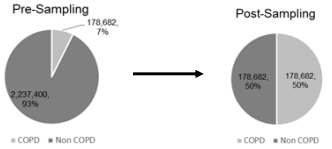
Biomarker #	Variable	Correlation to COPD Dx
77	Asthma Dx	0.433
80	Shortness of Breath Dx	0.186
70	Age at Index	0.166
78	Pneumonia Dx	0.153
79	Congestive Heart Failure Dx	0.151
25	Coagulation Tissue Factor Induced INR*	0.122
31	Troponin I Cardiac	0.121
25	Average Value of Coagulation Tissue Factor Induced INR	0.119
25	Median Value of Coagulation Tissue Factor Induced INR	0.117
36	Oxygen	0.115
15	Max Value of Eosinophils / 100 Leukocytes	0.114
25	Min Value of Coagulation Tissue Factor Induced INR	0.114
15	Standard Deviation of Eosinophils / 100 Leukocytes	0.112
25	Max Value of Coagulation Tissue Factor Induced INR	0.110
22	Max Value of Glucose	0.106
41	Base Excess	0.105

Biomarker #	Variable	Correlation to COPD Dx
77	Duration between first Asthma Dx prior to index date and index date	-0.477
78	Duration between first Pneumonia Dx prior to index date and index date	-0.246
79	Duration between first Congestive Heart Failure Dx prior to index date and index date	-0.236
80	Duration between first Shortness of Breath Dx prior to index date and index date	-0.185
7	Duration between last Heart Rate Observation prior to index date and index date	-0.145
1	Duration between last BMI Observation prior to index date and index date	-0.140
2	Duration between last Hematocrit Observation prior to index and index date	-0.134
4	Duration between last Platelets Observation prior to index date and index date	-0.127
3	Duration between last Leukocytes Observation prior to index date and index date	-0.127
6	Duration between last Hemoglobin Observation prior to index date and index date	-0.120
8	Duration between last Erythrocytes Observation prior to index date and index date	-0.110
24	Duration between last Erythrocyte Mean Corpuscular Hemoglobin Concentration prior to index and index date	-0.107

*Coagulation Tissue Factor Induced INR (International Normalized Ratio): a clinical measure of warfarin dosage, liver damage and/or vitamin K status. It is defined as ratio of a patient prothrombin time to a normal sample

- Asthma, shortness of breath, age, pneumonia, coagulation tissue factor induced INR and 24 other biomarkers were correlated to the COPD Dx.

Results: Balancing Cohort Size*



In Correlation Analysis and Prediction Modeling, a subset of non-COPD smokers (~179k patients, same number as COPD smokers) was generated as the cohort non-COPD smokers after randomized sampling.

* Note: the number of COPD smokers drop from ~181k to ~179k is due to the fact that ~179k patients had at least 1 of the selected 75 biomarkers considered for modeling

Results: Predictive Model

Feature	Coefficient	P-Value	Odds Ratio	Summary
Platelets Observation	0.884	<0.001	**2.42**	**The odds of having COPD for smokers that have had at least 1 Platelets test are 142% higher as compared with those that have not had a Platelets test**
Cholesterol in HDL Observation	0.460	<0.001	**1.58**	**The odds of having COPD for smokers that have had at least 1 Cholesterol in HDL test are 58% higher as compared with those that have not had a Cholesterol in HDL test**
Coagulation Tissue Factor Induced INR Observation	0.398	<0.001	**1.49**	**The odds of having COPD for smokers that have had at least 1 Coagulation Tissue Factor Induced INR test are 49% higher as compared with those that have not had a Coagulation Tissue Factor Induced INR test**
Coagulation Tissue Factor Induced INR Min Value	0.270	<0.001	**1.31**	**The odds of having COPD for smokers increase by 31% with unit increase in Coagulation Tissue Factor Induced INR min value**
Coagulation Tissue Factor Induced INR Standard Deviation	0.220	0.019	**1.25**	**The odds of having COPD for smokers increase by 25% with unit increase in Coagulation Tissue Factor Induced INR standard deviation value**
Troponin Test Observation	0.200	<0.001	**1.22**	**The odds of having COPD for smokers that have had at least 1 Troponin test are 22% higher as compared with those that have not had a Troponin test**
Eosinophils per 100 Leukocytes Max Value	0.084	<0.001	**1.09**	**The odds of having COPD for smokers increase by 9% with unit increase in Eosinophils per 100 Leukocytes max value**
Coagulation Tissue Factor Induced INR Max Value	0.057	0.002	**1.06**	**The odds of having COPD for smokers increase by 6% with unit increase in Coagulation Tissue Factor Induced INR max value**
Age at index	0.045	<0.001	**1.05**	**The odds of having COPD for smokers increase by 5% with unit increase in Age**

- A logistic regression model with 96 features was developed with 76% precision and 0.801 AUC, after removing all comorbidity features.
- The top 9 features related to increased likelihood of having COPD are shown in the table above, with clinical features related to platelets, cholesterol in HDL, and coagulation tissue factor having highest odds ratio.

Conclusions

- A biomarker-based linear regression model for the risk of developing COPD was developed, with top clinical biomarker features related to platelet, cholesterol in HDL, coagulation tissue factor, leukocytes and age.
- Our retrospective analysis of smokers' electronic health records identified high prevalence of cardiovascular-related comorbidities and predictive biomarkers, suggesting COPD smokers were likely to experience a cardiovascular-related comorbidity event prior to COPD Dx.
- These findings aid in an individual risk scoring model to estimate the likelihood of smokers developing COPD.