Prediction of Chronic Obstructive Pulmonary Disease (COPD) in Asthma Patients Using Electronic Medical Records

Prediction of Chronic Obstructive Pulmonary Disease (COPD) in Asthma Patients Using Electronic Medical Records

Blanca E Himes, Yi Dai, Isaac S Kohane, et al.

JAMIA 2009 16: 371-379
doi: 10.1197/jamia.M2846

Research Paper ■

Prediction of Chronic Obstructive Pulmonary Disease (COPD) in Asthma Patients Using Electronic Medical Records

Blanca E. Himes, PhD, Yi Dai, Isaac S. Kohane, MD, PhD, Scott T. Weiss, MD, MS, Marco F. Ramoni, PhD

Abstract **Objective:** Identify clinical factors that modulate the risk of progression to COPD among asthma patients using data extracted from electronic medical records.

Design: Demographic information and comorbidities from adult asthma patients who were observed for at least 5 years with initial observation dates between 1988 and 1998, were extracted from electronic medical records of the Partners Healthcare System using tools of the National Center for Biomedical Computing "Informatics for Integrating Biology to the Bedside" (i2b2).

Measurements: A predictive model of COPD was constructed from a set of 9,349 patients (843 cases, 8,506 controls) using Bayesian networks. The model's predictive accuracy was tested using it to predict COPD in a future independent set of asthma patients (992 patients; 46 cases, 946 controls), who had initial observation dates between 1999 and 2002.

Results: A Bayesian network model composed of age, sex, race, smoking history, and 8 comorbidity variables is able to predict COPD in the independent set of patients with an accuracy of 83.3%, computed as the area under the Receiver Operating Characteristic curve (AUROC).

Conclusions: Our results demonstrate that data extracted from electronic medical records can be used to create predictive models. With improvements in data extraction and inclusion of more variables, such models may prove to be clinically useful.

■ **J Am Med Inform Assoc.** 2009;16:371–379. DOI 10.1197/jamia.M2846.

Introduction

Electronic medical records (EMRs) have been widely heralded for their potential to improve the quality of patient care.[1] Less obvious is the use of such records for clinical research, in particular, to develop predictive tools with an eye to the future of personalized medicine. The Partners Healthcare Research Patient Data Registry (RPDR), which contains information on over three million patients that have been treated in Partners-affiliated hospitals, is one of the larger collections of electronic medical records from academic medical centers. The RPDR has been used for a multitude of studies from quality improvement to drug efficacy and genomic discovery.[2] The computational infrastructure developed by the National Center for Biomedical Computing "Informatics for Integrating Biology to the Bedside" (i2b2) has been used to develop an "asthma data mart", which was refined and analyzed using the i2b2 workbench.[3–5] The asthma data mart contains codified annotations of patient records (e.g., billing codes) that have been augmented with additional concepts/phenotypes extracted from the textual notes of medical records using Natural Language processing (NLP) techniques.[6] The authors tested whether an effective clinical predictor could be designed using data extracted from the asthma data mart's hospital admission and emergency room visit EMRs.

Chronic obstructive pulmonary disease (COPD), a slowly progressive disease characterized by increased nonreversible airflow limitation, is a leading cause of morbidity and mortality worldwide.[7] A heterogeneous disease process, COPD differs among patients in its development, pathology, and comorbidity, and is traditionally divided into two types: chronic bronchitis and emphysema.[8] One of the risk factors for COPD is asthma.[7] Asthma is a complex respiratory disease characterized by airway hyperresponsiveness and reversible airflow limitation that often develops early in life. Patients with asthma seem to be at an increased risk to develop COPD, which is a significant source of morbidity

Affiliations of the authors: Harvard-MIT Division of Health Sciences and Technology (BEH, ISK, MFR), Cambridge, MA; Children's Hospital, Informatics Program Technology, Harvard Medical School (BEH, ISK, MFR), Boston, MA; Partners Healthcare Center for Personalized Genetic Medicine (BEH, ISK, STW, MFR), Boston, MA; Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School (BEH, STW), Boston, MA; Wellesley College (YD), Wellesley, MA.

Supported by the following NIH grants: 5U54LM008748-02 (National Centers for Biomedical Computing), 2U01HL065899 (National Heart, Lung, and Blood Institute) and 2T15LM007092-16 (NLM). The authors thank Shawn Murphy, MD, PhD and Vivian Gainer, MS of the Massachusetts General Hospital Laboratory of Computer Science at Harvard Medical School, for facilitating access to the i2b2 asthma data mart.

Correspondence: Blanca E. Himes, PhD, Channing Laboratory, 181 Longwood Ave, Boston, MA 02115; e-mail: <blanca_himes@hms.harvard.edu>.

Received for review: 05/04/08; accepted for publication: 01/30/09

and mortality among asthmatics.[8–10] Smoking is the known major risk factor for COPD yet only 25% of smokers get COPD.[11] According to the "Dutch Hypothesis" for the development of COPD it is the inherited susceptibility to increased airway responsiveness and allergy in asthmatics that leads to the development of COPD in some subjects.[12,13] However, particularly in the clinical setting, a thorough understanding of which asthma patients go on to develop COPD has not been achieved. The ability to predict which asthma patients develop COPD would be useful to understand the pathology underlying the development of this disease and to alter the clinical course of these patients. Traditional studies that look into the development of COPD are longitudinal epidemiological studies that focus on initially normal subjects and use a wide variety of clinical measures, including lung function tests, and medical and smoking history to follow and evaluate disease onset rather than progression.[14] These studies typically include a small numbers of milder asthmatics and have not given a clear picture of the relationship of asthma to COPD.

In this work, we created a predictive model of COPD in asthma patients using demographic and comorbidity data extracted from the i2b2 asthma data mart. The predictive model was created using Bayesian networks, multivariate models that are able to account for simultaneous associations and interactions among variables to make predictions of COPD. Bayesian networks have been successfully used in a wide variety of medical applications, from public health surveillance systems,[15] to the classification of brain tumors based on radiological data[16] and ovarian tumors based on the integration of clinical and literature data.[17] Mortality is a favored outcome to predict with Bayesian networks as illustrated by studies predicting risk of death among cardiac surgery patients[18] and sickle cell anemia patients.[19] Additionally, Bayesian networks have been used to predict clinical phenotypes including the diagnosis of acute appendicitis,[20] the assessment of ballistic penetrating trauma,[21] and the identification of patients at risk for asthma exacerbations.[22] The authors describe here the extension of the application of Bayesian networks to the task of prediction of a clinical phenotype using data extracted from EMRs.

Methods

Data Collection

A set of 10,341 asthma patients from the i2b2 asthma data mart was obtained (Fig 1a). Records from these patients

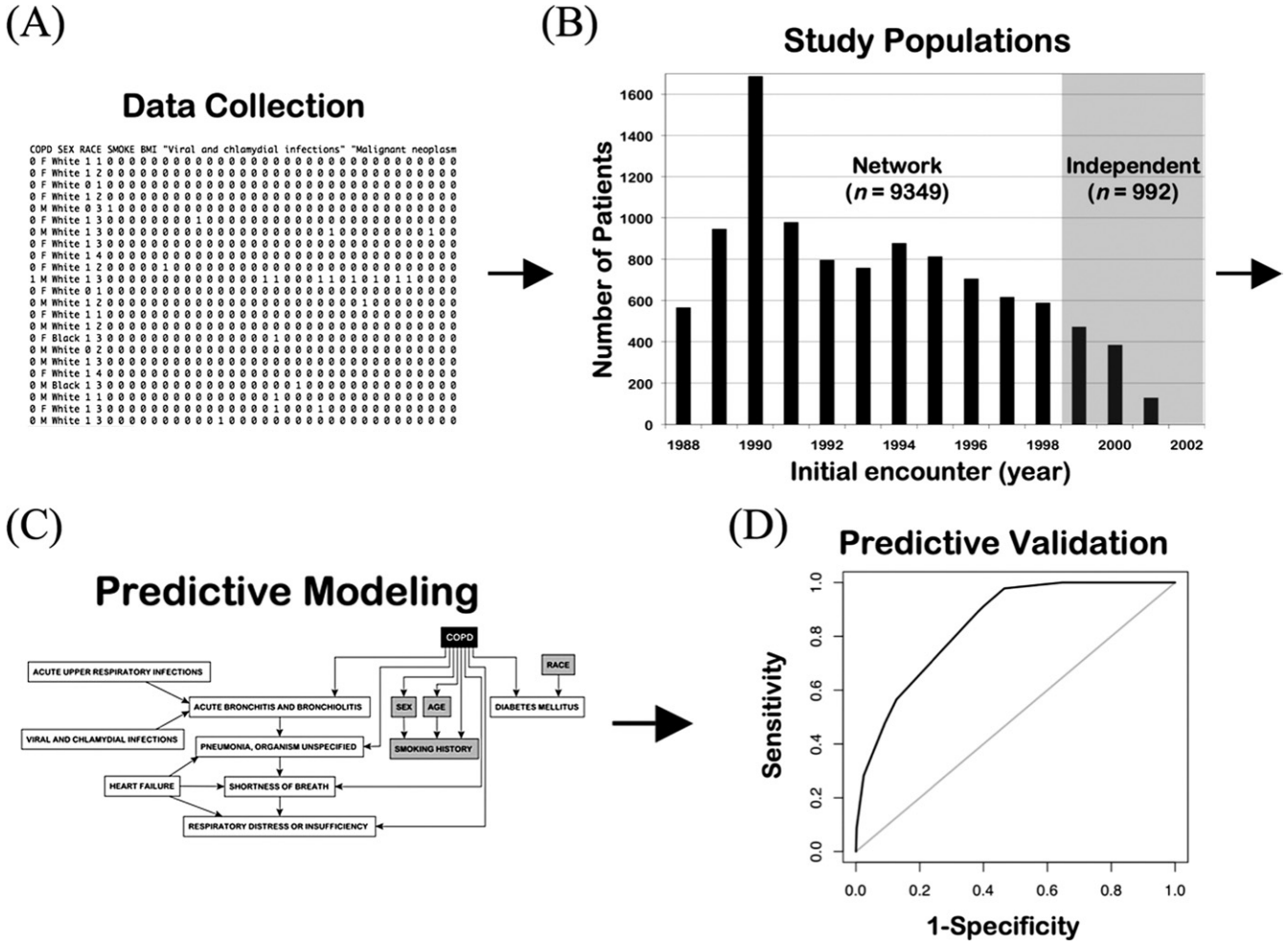

Figure 1. Procedure outline. (A) Clinical data was extracted from electronic medical records of asthma patients using i2b2 tools. (B) Patients were divided into two groups, network and independent, according to initial observation date. (C) A predictive model was created using Bayesian networks with data from the network group of patients. (D) The performance of the predictive model was evaluated using receiver operating characteristic (ROC) curves with data from the independent group of patients.

have been parsed for coded data and variables such as smoking history extracted by NLP of unstructured text.[6] The collection and study of these data is approved by the Institutional Review Board of Partners Healthcare System.

Patients were included if observed for at least 5 years as determined by the dates of the earliest and latest records available, were at least 18 years of age at the initial observation date, and had race, sex, height, weight, and smoking history data available. Age at initial observation (age), race, sex, height, and weight values were lifted directly from EMRs, as they are structured information in the RPDR. Age was categorized into four groups: 18–44, 45–64, 65–74, and 75+ years old. The values of race, as extracted from EMRs, were "White," "Black," "Hispanic," and "Asian." The BMI was calculated with the average height and weight measurements for each subject's EMRs as weight in kilograms divided by height in meters squared. The BMI was categorized as follows: underweight (BMI $<$ 18), normal (18 $<=$ BMI $<$ 25), overweight (25 $<=$ BMI $<$ 30), obese (30 $<=$ BMI $<$ 40), and morbidly obese (BMI $>=$ 40). Smoking status for each EMR was determined using the method outlined in Zeng et al., 2006.[6] Briefly, the Health Information Text Extraction (HITEx) tool, which uses the Collection of Reusable Objects for Language Engineering (CREOLE) included in the General Architecture for Text Engineering (GATE) platform,[23] was used to determine for each EMR whether a subject's smoking status was "current smoker," "past smoker," "never smoker," "denies smoking," or "insufficient data." The HITEx NLP procedure to classify smoking status consisted in (1) splitting an EMR into sections, (2) finding all occurrences of regular expressions of interest (i.e., smoking keywords) (3) filtering sections into appropriate categories, (4) splitting sections into sentences, (5) extracting word fragments and their frequency from text, and (6) classifying smoking-related fragments using a support vector machine (SVM). For this study, smoking history was dichotomized into "Negative" if the smoking status was determined to be "never smoker" in 90% of a subject's EMRs and "Positive" otherwise.

Patients were determined to have comorbidities on the basis of International Classification of Diseases, Ninth Revision (ICD-9) codes being used as admission diagnosis codes for hospitalizations or primary diagnosis codes for emergency room visits. Such codes were obtained via direct extraction from billing codes in EMRs. For this study, if a patient had at least one instance of a code being used for admission and/or primary diagnosis during the entire time course of observation, then the ICD-9 code variable was assigned a value of "1," otherwise it was assigned a value of "0." The time order of diagnoses was disregarded to simplify the model, maximize the amount of information per patient, and because the authors could not be certain of a patient's medical history before the initial observation. Cases are those subjects who had COPD, determined by having a value of "1" in ICD-9 codes corresponding to at least one of the following: "Chronic Bronchitis," "Emphysema," or "Chronic Airways Obstruction, not otherwise specified." Controls are those subjects who had a value of "0" in these ICD-9 codes. The remaining 104 comorbidity variables had a value of "1" in at least 1% of the patients.

Study Populations

Patients were divided into two groups according to initial observation date (Fig 1b). A cohort of 9,349 patients (843 cases, 8,506 controls), who were initially observed between 1988 and 1998, were selected to create the predictive model of COPD. A future independent set of 992 patients (46 cases, 946 controls) had initial observation dates between 1999 and 2002. The independent group was used to test the predictive model.

Predictive Modeling

Because of the large number of variables available to create the model and the expectation that relationships among many of these variables are complex, the authors used Bayesian networks to find a predictive model. A Bayesian network is a directed acyclic graph where nodes represent variables and edges between nodes represent probabilistic relationships between variables (i.e., a node that has an incoming arrow is dependent on the node from which the arrow originates). The topology of a Bayesian network and the associated probabilistic relationships among variables can be learned directly from data, making Bayesian networks powerful for extracting complex and unbiased relationships among variables. Bayesian networks are better equipped to find complex relationships than traditional regression approaches because they are not limited to representing the dependencies of a single outcome variable on predictor variables. Among other machine learning methods that can create predictive models with many variables, including artificial neural networks[24] and support vector machines,[25] Bayesian networks have the advantage of creating an intuitive graphical representation of the complex relationships among variables that can help to understand the underlying quantitative relationships. In this work, a Bayesian network was constructed from the set of 9,349 patients (843 cases, 8,506 controls) and 109 clinical variables using the K2 algorithm, a common and efficient approach to identify the most probable network of dependency from a dataset.[26] To find such a network, a space of different network models is explored and each is scored by its posterior probability given the data. The model with maximum posterior probability is returned. The network found with all 109 variables contained many relationships among variables, and several of these relationships did not directly influence COPD. The authors focused on the nodes that directly modulate COPD, the so-called Markov Blanket of COPD (Fig 1c). These nodes consist in the nodes with edges directed to COPD (i.e., parents of COPD), nodes with edges originating from COPD (i.e., children of COPD), and nodes with edges directed to children of COPD (i.e., parents of children of COPD). Model robustness was tested via a fivefold cross-validation in which each of five non-overlapping data subsets, obtained by randomly splitting the original dataset, is used as an independent dataset while the remaining four subsets are used to quantify the network dependencies. The odds ratios of single variable effects were calculated using median-unbiased estimation in R.[27]

Predictive Validation

A future independent set of patients, with respect to the time interval of patients' initial observation, was used for the predictive validation of the model (992 patients; 46 cases, 946 controls). COPD was predicted in each patient of the inde-

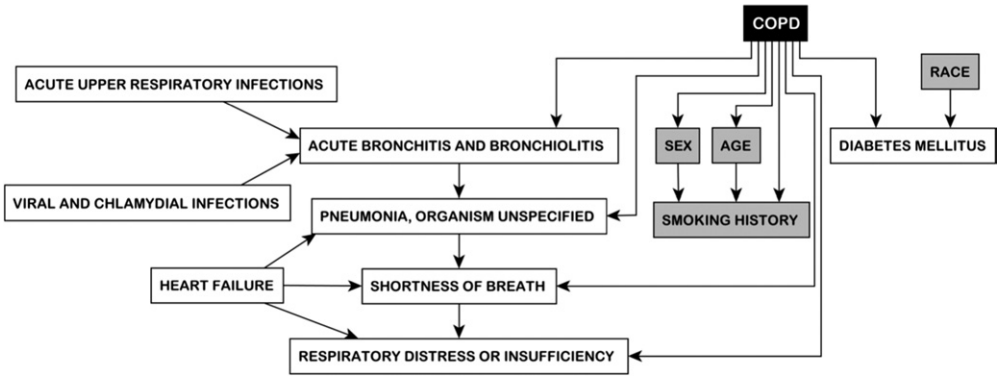

Figure 2. Predictive network of Chronic Obstructive Pulmonary Disease (COPD).

pendent set, and predicted COPD was compared to observed COPD. The probability of COPD given the comorbidity and demographic profile of an individual subject was calculated using the Clique algorithm implemented in Bayesware Discoverer.[28] The performance of the predictive model was evaluated with receiver operating characteristic (ROC) curves (Fig 1d). Predictive accuracy was measured as the area under the ROC curve (AUROC), and significance for this accuracy was obtained by comparing the classification ability of models obtained to random classification. The standard error (SE) for AUROCs and for the difference between AUROCs of two curves were estimated using the nonparametric asymptotic method proposed by DeLong et al., 1988[29] and described in Lasko et al., 2005.[30] For ROC plots, convex hulls were estimated. All ROC analyses were performed in R.[27]

Results

Medical data for 104 comorbidities, age, sex, race, BMI, and smoking history from 9,349 asthma patients was extracted from the i2b2 data mart. A predictive model of COPD was created with this medical data using Bayesian networks to find which asthma patients develop COPD (843 cases, 8,506 controls). Because the authors were interested in predicting COPD, the authors focused on the variables in the network that directly modulate it: age, sex, race, smoking history, and 8 comorbidities (Fig 2). In Fig 2, each box is a node that represents a variable, and each arrow between nodes is an edge that represents a dependency of the node receiving the arrow on the originating node. For example, the arrow between "COPD" and "smoking history" indicates that "smoking history" is dependent on "COPD." Similarly, "Acute bronchitis and bronchiolitits," "pneumonia, organism unspecified," "shortness of breath," "respiratory distress or insufficiency," "diabetes mellitus," "sex," and "age" are dependent on "COPD." Slightly more complex relationships are indicated by three nodes connected with two arrows. For example, "diabetes mellitus" is dependent on both "COPD" and "race." These two relationships imply that "COPD" and "race" are dependent on each other when the state of "diabetes mellitus" is known. Analogously, "acute upper respiratory infections," "viral and chlamydial infections," and "heart failure" are related to COPD through other nodes. It is important to note that the arrows in the network do not encode causal relationships. Although causal relationships may be found in a Bayesian network, proof that such a relationship exists requires further study with isolated variables. However, probabilistic dependencies represented by arrows suggest legitimate relationships that may lead to novel findings and can quantitatively assess the strength of known relationships. The distribution of the COPD network variables among cases and controls is shown in Tables 1 and 2. The model is robust to sampling variability, as demonstrated by a fivefold cross-validation AUROC of 0.83 (SE 0.01).

The generalizability of the network was tested by using it to predict COPD in an independent set of 992 asthma patients (46 COPD cases, 946 controls). When performing prediction, the information of all clinical variables, except for COPD, is used to infer the probability with which COPD is present in each patient based on the parameters learned by the network in the learning stage. This inference requires reversal of the relationships indicated by the edges in the network. For example, after learning that "sex" depends on "COPD," the authors have to know how "COPD" depends on "sex." The calculation of these inverted relationships is possible because of Bayes' theorem and is the basis of the prediction algorithm. The ROC curve corresponding to the classification of the independent subjects is shown in Fig 3. The corresponding AUROC is 0.83 (SE 0.03), which suggests the generalizability of the network and its ability to predict COPD.

The marginal effects and predictive accuracy of individual network variables are shown in Table 3. The marginal effect of each variable is represented as the odds ratio for a state of that variable being associated to COPD. The information of single variables in the independent group of subjects was used to predict COPD with the network. Five of the 12 variables in the network have AUROCs that are significantly different from random classification of the patients ($p < 0.01$) (Fig 4). Age is the single variable with the largest AUROC (0.81, SE 0.04), and predicting COPD with age alone is not statistically different than using all variable information ($p = 0.21$). If age information is excluded and the remaining variable information is used to predict COPD, then the corresponding AUROC is 0.73 (SE 0.04), which is still significantly better than random ($p = 2.95E\text{-}08$).

Discussion

A well-known goal of the use of EMRs is to improve the quality and efficiency of patient care.[1] EMRs have been acknowledged as a source to identify large numbers of subjects for research studies. Such studies include those limited to data collected through EMRs (e.g., the understanding of individual disease courses and outcomes), but

Table 1 ■ Patient Characteristics

	Network		Independent	
	Controls (n = 8506)	Cases (n = 843)	Controls (n = 946)	Cases (n = 46)
Sex				
female	6354 (74.70)	530 (62.87)	698 (73.78)	31 (67)
male	2152 (25.30)	313 (37.13)	248 (26.22)	15 (33)
Age				
18–44	3709 (43.60)	78 (9.25)	535 (56.55)	3 (6.5)
45–54	3261 (38.34)	425 (50.42)	300 (31.71)	19 (41)
55–64	1035 (12.17)	235 (27.88)	79 (8.35)	18 (39)
75+	501 (5.89)	105 (12.46)	32 (3.38)	6 (13)
Race				
Asian	116 (1.36)	8 (0.95)	22 (2.33)	1 (2.2)
Black	1071 (12.59)	103 (12.22)	101 (10.68)	6 (13)
Hispanic	1105 (12.99)	71 (8.42)	144 (15.22)	3 (6.5)
White	6214 (73.05)	661 (78.41)	679 (71.78)	36 (78)
Smoking history				
negative	2385 (28.04)	25 (2.97)	299 (31.61)	2 (4.4)
positive	6121 (71.96)	818 (97.03)	647 (68.39)	44 (96)

For each variable category, Number (%) are reported.

also extend to those requiring additional data gathering (e.g., genetic studies of complex diseases). With these goals in mind, data has been extracted from EMRs at various medical centers for the identification of subjects with diseases including asthma,[6] diabetes mellitus,[31] and heart failure.[32] However, few studies have demonstrated that the extracted data can itself be useful for clinical studies. In this work, the authors used data extracted from EMRs with tools available in the i2b2 asthma data mart to characterize and predict which asthma patients develop COPD.

Bayesian networks were used to create a predictive model of COPD using the following extracted variables: age, sex, race, BMI, smoking history, and 104 comorbidities. Of these, age, sex, race, smoking history, and 8 comorbidities modulate the risk of COPD. The model has good predictive accuracy, as indicated by an AUROC of 0.83 (SE 0.03) when using the model to predict COPD in an independent set of patients. The ability of single variables to predict COPD was assessed using the information from one variable at a time to predict COPD with the network (Table 3). The strongest single

Table 2 ■ Distribution of Chronic Obstructive Pulmonary Disease (COPD) Network Comorbidities in Patients

		Network		Independent	
ICD-9 Code	Description	Controls (n = 8506)	Cases (n = 843)	Controls (n = 946)	Cases (n = 46)
0799	viral and chlamydial infections				
	0	8274 (97.27)	819 (97.15)	926 (97.89)	44 (96)
	1	232 (2.73)	24 (2.85)	20 (2.11)	2 (4.4)
250	diabetes mellitus				
	0	8338 (98.02)	796 (94.42)	938 (99.15)	45 (98)
	1	168 (1.98)	47 (5.58)	8 (0.85)	1 (2.2)
428	heart failure				
	0	8131 (95.59)	693 (82.21)	927 (97.99)	40 (87)
	1	375 (4.41)	150 (17.79)	19 (2.01)	6 (13)
465	acute upper respiratory infections				
	0	8158 (95.91)	785 (93.12)	917 (96.93)	45 (98)
	1	348 (4.09)	58 (6.88)	29 (3.07)	1 (2.2)
466	acute bronchitis and bronchiolitis				
	0	8259 (97.10)	784 (93.00)	929 (98.20)	43 (93)
	1	247 (2.90)	59 (7.00)	17 (1.80)	3 (6.5)
486	pneumonia, organism unspecified				
	0	7776 (91.42)	546 (64.77)	888 (93.87)	35 (76)
	1	730 (8.58)	297 (35.23)	58 (6.13)	11 (24)
78605	shortness of breath				
	0	8057 (94.72)	623 (73.90)	905 (95.67)	38 (83)
	1	449 (5.28)	220 (26.10)	41 (4.33)	8 (17)
78609	respiratory distress or insufficiency				
	0	8173 (96.09)	678 (80.43)	932 (98.52)	40 (87)
	1	333 (3.91)	165 (19.57)	14 (1.48)	6 (13)

For each variable category, Number (%) are reported.

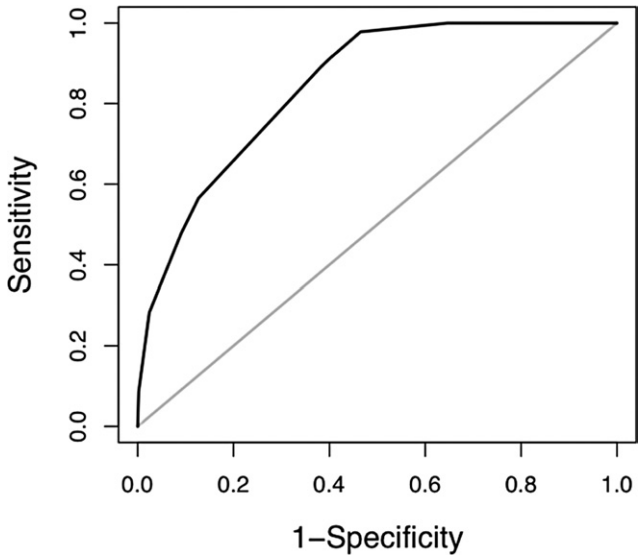

Figure 3. Receiver Operating Characteristic (ROC) curve corresponding to prediction of Chronic Obstructive Pulmonary Disease (COPD) in an independent group of patients.

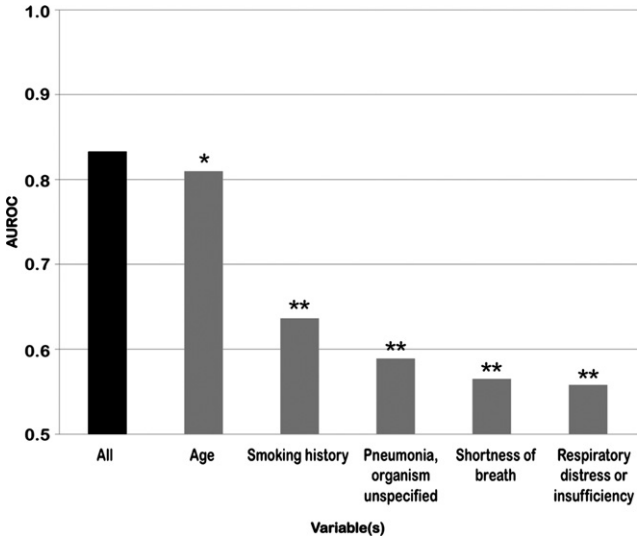

Figure 4. Predictive accuracy of individual network variables that perform better than random. *P value comparing AUROC of all variables to using Age only: 0.21. **P value comparing AUROC of all variables to using other single variables shown <1E-6.

variable predictor in the network is age (Fig 4). Surprisingly, this variable alone predicts COPD with an AUROC of 0.81 (SE 0.04) in the independent subjects, which is not significantly different than the area obtained with all variables. The relationship between age and COPD is well known. Because COPD is a chronic disease that worsens over time, it is characteristically present in older adults.[7] Emergency room (ER) visit and hospitalization rates for COPD among U.S. adults have been estimated[33] and are consistent with our findings. The 65–74 and 75+ age groups have the highest rates of ER visits and hospitalizations, while the youngest groups have the lowest. Although the importance of age in COPD is known, it is not obvious that it should be the best predictor in our model. For example, smoking is the most important cause of COPD[34] and being male is also traditionally associated with a higher likelihood of having COPD.[7] Besides age, most other single variables are unable to predict COPD better than at random, and those that are able to predict COPD better than at random have AUROCs that are significantly lower than that using all variables or age (Fig 4). If age information is suppressed while the remaining variable information is used to predict COPD, the corresponding AUROC is 0.73 (SE 0.04). This demonstrates

Table 3 ■ Strength and Predictive Ability of Individual Network Variables

Variable	Effect, OR	Referent Group	Single Variable AUROC (SE)	p Value
Acute bronchitis and bronchiolitis	2.52 (1.86, 3.36)	0	0.52 (0.04)	0.013
Acute upper respiratory infections	1.74 (1.29, 2.30)	0	0.50 (0)	1
Age			0.81 (0.04)	1.22E-15
45–64	6.19 (4.87, 7.97)	18–44		
65–74	10.78 (8.30, 14.13)	18–44		
75+	9.95 (7.32, 13.57)	18–44		
Diabetes mellitus	2.94 (2.09, 4.06)	0	0.51 (0.02)	0.18
Heart Failure	4.69 (3.82, 5.75)	0	0.50 (0)	1
Pneumonia, organism unspecified	5.79 (4.93, 6.80)	0	0.59 (0.06)	1.86E-06
Race			0.50 (0)	1
Asian	0.66 (0.29, 1.27)	White		
Black	0.91 (0.72, 1.12)	White		
Hispanic	0.61 (0.47, 0.77)	White		
Respiratory distress or insufficiency	5.97 (4.87, 7.30)	0	0.56 (0.05)	2.59E-08
Sex	1.74 (1.50, 2.02)	female	0.53 (0.05)	0.17
Shortness of breath	6.34 (5.28, 7.59)	0	0.57 (0.05)	3.33E-05
Smoking history	12.67 (8.68, 19.43)	negative	0.64 (0.01)	4.34E-05
Viral and chlamydial infections	1.05 (0.67, 1.58)	0	0.50 (0)	1

The second column reports odds ratios and 95% confidence intervals in parentheses.
The referent group refers to that used to compute the odds ratio.
The fourth column reports the AUROC of the network using the single variable data and corresponding standard errors (SE).
Fifth column p values correspond to comparison of single variable AUROC to a random classifier.
The network AUROC using all variables is 0.83 (p = 8.88E-15 compared to random).
AUROC = Area Under the Receiver Operating Characteristic curve.

that the model contains significant predictors of COPD besides age, and that interesting interactions among these variables are able to predict COPD, albeit with lower accuracy than age alone. Consistent with the Dutch hypothesis of COPD, these results suggest that some subjects with asthma develop COPD as they age regardless of their smoking status and independently of other network variables. Further study of the relationships among the network's variables is required to confirm this premise, and incorporation of other variables into the model is necessary to understand what alters the progression to COPD among patients with asthma as they age.

Some of the comorbidities that were found to modulate the risk of COPD are general symptoms, such as "shortness of breath" and "respiratory distress or insufficiency." These symptoms alone are insufficient to indicate COPD, but in the context of the model are helpful to predict COPD. Infections are known to be related to COPD exacerbations,[35] which supports some of the other variable relationships with COPD (e.g., "pneumonia, organism unspecified", "acute bronchitis"). For each patient, the authors used all ER visit primary diagnoses and hospitalization admission diagnoses available during the observation period to infer comorbidities. The authors did not differentiate comorbidities based on the order in which they occurred in time. Therefore, the authors are limited in knowing whether the above comorbidity predictors and COPD are in causal relationships. In future studies, more careful evaluation of the time course of comorbidities may help to establish whether there are early predictors of COPD. The current results are still useful to indicate how COPD is related to other comorbidities. For example, patients who have hospital admissions/ER visits for COPD are likely to have separate hospital admissions/ER visits for shortness of breath, but shortness of breath events are also related to pneumonia, heart failure, and respiratory distress or insufficiency. Though one might intuitively expect that COPD and shortness of breath be related, the network demonstrates that the relationship is influenced by several other variables, and relationships between variables that would be intuitively thought to be directly related are not always present (e.g., acute upper respiratory infections and shortness of breath). Further, the network provides a quantitative measure of relationships among variables, which is stronger than having an intuition that relationships should exist.

Though the model performs well at predicting COPD in the independent set of asthma patients, there is clear room for improvement in predictive accuracy. The ROC curve in Fig 3 shows that, in the independent group of patients, the predictive model is highly sensitive (100/90/80%) for thresholds at which the specificity is lower (45/60/67%). Some of the factors that affect the predictive accuracy are errors in data extraction, the inherent limitations of the medical record data, and the challenge of determining which patients have COPD.

There is virtually no error in extracting age, sex, race, and BMI as these variables are structured data in the RPDR. However, there are limitations to the ways in which temporal age and BMI measures can be represented as a single value for this study. Because the authors condensed longitudinal data to establish which patients were affected by comorbidities, the authors selected single values for age and BMI for each patient. The age that was used was that of the patient at the initial ER visit or hospitalization recorded. The range of observations for each patient had to be at least five years but was never greater than 10 years. Therefore, the age ranges that the authors used to categorize patients tended to be much greater than the amount of aging that each patient underwent during the observation period (i.e., most patients remained in the same age category across observations). The BMI that the authors calculated was based on average height and weight values. Because subjects were adults during the times of observation, height is expected to remain constant. However, a person's weight can fluctuate over time. The authors assumed that the average weight was the best representative weight over the course of observation because it would best account for the weight that was most often observed for each patient. In the vast majority of patients, there was minimal change in weight over the extracted values.

Extracting smoking history poses greater challenges than the extraction of other demographic variables because it requires the use of NLP methods on free text portions of EMRs. Previous work describes some of the problems involved in extracting smoking status from EMRs, and how NLP has been used to successfully determine it.[4,6] The HITEx methodology that was used to extract the smoking data for this paper has been shown to have an accuracy of 90% compared to expert classification,[6] which is very good reliability. In addition to the challenge of extracting smoking status by NLP, there can be uncertainty in the veracity of patient reporting. The authors chose to define a negative smoking history as one where a patient had "never smoked" in at least 90% of the smoking histories extracted. A positive smoking history could contain a mixture of "current-smoker," "past-smoker," and less than 90% "never-smoked" in the extracted smoking histories. Most often, positive smokers had a large percentage of "current-smoker" and "past-smoker" as extracted smoking histories. Because the authors looked at records from patients who were observed for at least 5 years, it was deemed more significant to differentiate patients with a positive smoking history from those with a negative history, than to differentiate current smokers from past smokers. Therefore, the authors chose a definition of nonsmokers that would attempt to ensure that this group truly contained patients with a negative history of smoking.

As for most demographic variables, comorbidities have virtually no error in being extracted by the i2b2 data mart as these variables were inferred directly from ICD-9 billing codes corresponding to ER primary diagnoses and hospitalization admission diagnoses. However, the assignment of diseases with ICD-9 codes is subject to error. In previous work, principal diagnosis classification using ICD-9 billing codes was measured to have accuracy between 72 and 80%, depending on the amount of data available per record, when compared to expert classification.[6] This low accuracy was nearly identical to that using HITEx (73–82%), but using ICD-9 codes resulted in higher specificity values (85–91% compared to 82–87%). Thus, most subjects classified as having a diagnosis according to ICD-9 codes, would also be classified as such by an expert. In addition to being limited

in their ability to describe individualized disease presentations, the use of ICD-9 codes is often biased and subject to error when they are assigned without a thorough patient evaluation or by physicians inexperienced in their use. Additionally, some diseases are not classified using uniform criteria, leading to the labeling of patients with different pathological processes as having the same disease. Despite these limitations, the use of ICD-9 codes is standard in United States hospitals for billing and record keeping because they allow a finite and uniform set of diagnoses.[36] In our work, ICD-9 codes provide a satisfactory means of classifying patients with diseases, as evidenced by the relationships among variables found by and good predictive accuracy of our model.

A further limitation of our study is the definition of COPD. A complex disease process, COPD is not always diagnosed using uniform criteria. Several patients with COPD feature characteristics of emphysema and chronic bronchitis, blurring the distinctions between these diseases, while some subjects with emphysema and chronic bronchitis do not have COPD.[37] Therefore, our definition of COPD, based on ICD-9 codes from EMRs, is likely to classify some subjects as having COPD that would not be classified this way based on other standards (e.g., lung function measures). This sort of misclassification would only decrease the predictive accuracy of our model and hence bias our result in the direction of no effect. Nonetheless, having objective measures such as lung function incorporated in our data extracted from EMRs would be helpful to increase the accuracy of our model. The i2b2 data mart will be expanded to include such measures, although few will likely be available because lung function tests are not routinely ordered for most patients. In this sense, data obtained in epidemiological studies, which gather uniform measures for all participants, have a clear advantage over data extracted from EMRs. However, both sources of data are important and information extracted from them can be complementary. The fastest increase in clinical knowledge is likely to be achieved by integrating findings from multiple sources.

Despite all the limitations listed above, including those attributable to extracting data by NLP, the authors have shown that our EMR-extracted data are of good enough quality to create predictive models. Our model's AUROC of 0.83 for predicting COPD in an independent group of patients is comparable to that of routine clinical tests such as prostate-specific antigen tests (AUROC 0.62–0.86)[38] and mammography (AUROC 0.67–0.84).[39] Such clinical tests are evaluated prospectively, while our model was created and tested with retrospective data, which may be subject to bias in the assignment of comorbidity ICD9 codes. Despite the limitations of the retrospective data used, the current AUROC of our model suggests that the performance on prospective data will be good. Gathering prospective data to test our predictive model will provide a more objective assessment of its predictive accuracy.

If the accuracy of NLP extraction of smoking status, which is 90% at this writing, were perfect, then the performance of our predictive model would increase significantly. Because smoking is known to be an important risk factor for COPD, the authors expect that the AUROC might improve by as much as 0.05. Similarly, if the errors associated with using ICD-9 codes were reduced, our model's predictive performance would increase. One way to accomplish this would be to improve principal/admission diagnosis assignment by combining NLP methods with the extraction of billing codes. Although our current accuracy of classification with ICD-9 codes is low (72–80%), the specificity is good (85–91%). Therefore, even though most comorbidities assigned in our data are accurate, the authors would likely obtain additional comorbidities per patient with better methods to extract primary/admission diagnoses from EMRs. This would likely strengthen some existing comorbidity relationships and perhaps introduce new ones, increasing the AUROC corresponding to prediction of COPD by 0.05–0.10. However, even if classification of smoking status and comorbidities from EMRs were perfect, the authors would still expect other errors mentioned (e.g., inaccuracy in patient reporting and limitations inherent in disease classification schemes), to keep the prediction accuracy of our model below 100%.

Our results demonstrate the promise of using medical records to create predictive models and attest to the utility of approaches like i2b2's in instrumenting the healthcare enterprise. Potential applications of predictive models include improved resource allocation for healthcare systems and more closely targeted individualized prevention/management programs. To this end, future studies will improve the NLP methodology used to extract data, expand our model to include more comorbidities and medication history, and consider the time course of patients' medical histories.

Conclusions

The authors have created a predictive model of COPD using comorbidities and demographic information extracted from medical records of asthma patients and used this model to predict COPD in an independent group of asthma patients with good predictive accuracy. In our model, age, sex, race, smoking history, and 8 comorbidities modulate the risk of COPD. The AUROC corresponding to prediction of COPD in the independent set of patients is 0.83. Age is the best individual predictor of COPD (AUROC = 0.81), but the remaining variables have notable ability to predict COPD in the absence of age information (AUROC = 0.73). Our results show that it is possible to use data extracted from medical records to create predictive models. With improvements in data extraction and inclusion of more variables, such models may prove to be clinically useful and serve to better understand disease trends.

References ■

1. Committee on Quality. Of Health Care in America IoM. Crossing the Quality Chasm: A New Health System for the 21st Century, Washington, D.C.: National Academy Press, 2001.
2. Nalichowski R, Keogh D, Chueh HC, Murphy SN. Calculating the benefits of a Research Patient Data Repository. AMIA Annu Symp Proc 2006;1044.
3. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc 2006;1040.
4. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008 Jan–Feb;15(1):14–24.
5. i2b2: Informatics for integrating biology and the bedside. Available at: http://www.i2b2.org. Accessed 1/5/09.

6. Zeng QT, Goryachev S, Weiss S, et al. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006;6:30.
7. Mannino DM, Buist AS. Global burden of COPD: Risk factors, prevalence, and future trends. Lancet 2007 September 1;370(9589): 765–73.
8. Global Strategy for the Diagnosis. Management and Prevention of COPD, global initiative for chronic obstructive lung disease (gold) 2007. Available at: http://www.goldcopd.org. Accessed 5/1/08.
9. Lange P, Parner J, Vestbo J, Schnohr P, Jensen GA. 15-year follow-up study of ventilatory function in adults with asthma. N Engl J Med 1998 Oct 22;339(17):1194–200.
10. Silva GE, Sherrill DL, Guerra S, Barbee RA. Asthma as a risk factor for COPD in a longitudinal study. Chest 2004 Jul;126(1): 59–65.
11. Lokke A, Lange P, Scharling H, et al. A 25 year follow up study of the general population. Thorax 2006 Nov;61(11):935–9.
12. Postma DS, Boezen HM. Rationale for the Dutch hypothesis. Allergy and airway hyperresponsiveness as genetic factors and their interaction with environment in the development of asthma and COPD. Chest 2004 Aug;126(2) (Suppl):96S–104S; Discussion:59S–61S.
13. Xu X, Rijcken B, Schouten JP, Weiss ST. Airways responsiveness and development and remission of chronic respiratory symptoms in adults. Lancet 1997 Nov 15;350(9089):1431–4.
14. Hospers JJ, Postma DS, Rijcken B, Weiss ST, Schouten JP. Histamine airway hyper-responsiveness and mortality from chronic obstructive pulmonary disease: A cohort study. Lancet 2000 Oct 14;356(9238):1313–7.
15. Tsui FC, Espino JU, Dato VM, Gesteland PH, Hutman J, Wagner MM. Technical description of RODS: A real-time public health surveillance system. J Am Med Inform Assoc 2003 Sept–Oct; 10(5):399–408.
16. Reynolds GM, Peet AC, Arvanitis TN. Generating prior probabilities for classifiers of brain tumours using belief networks. BMC Med Inform Decis Mak 2007;7:27.
17. Antal P, Fannes G, Timmerman D, Moreau Y, De Moor B. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. Artif Intell Med 2004 Mar;30(3):257–81.
18. Verduijn M, Rosseel PM, Peek N, de Jonge E, de Mol BA. Prognostic Bayesian networks II: An application in the domain of cardiac surgery. J Biomed Inform 2007 Dec;40(6):619–30.
19. Sebastiani P, Nolan VG, Baldwin CT, et al. A network model to predict the risk of death in sickle cell disease. Blood 2007 Oct 1;110(7):2727–35.
20. Sakai S, Kobayashi K, Nakamura J, Toyabe S, Akazawa K. Accuracy in the diagnostic prediction of acute appendicitis based on the Bayesian network model. Methods Inf Med 2007;46(6):723–6.
21. Ogunyemi OI, Clarke JR, Ash N, Webber BL. Combining geometric and probabilistic reasoning for computer-based penetrating-trauma assessment. J Am Med Inform Assoc 2002 May–Jun;9(3):273–82.
22. Sanders DL, Aronsky D. Detecting asthma exacerbations in a pediatric emergency department using a Bayesian network. AMIA Annu Symp Proc 2006:684–8.
23. Cunningham H, Humphreys K, Gaizauskas R, Wilks Y, GATE. A Tipster-Based General Architecture for Text Engineering. TIPSTER Text Program (phase III) 6 Month Workshop, California: Morgan Kaufmann; 1997.
24. Krogh A. What are artificial neural networks? Nat Biotechnol 2008 Febr;26(2):195–7.
25. Noble WS. What is a support vector machine? Nat Biotechnol 2006 Dec;24(12):1565–7.
26. Cooper G, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn 1992;9(4):309–47.
27. R Development Core Team. R: A language and environment for statistical computing. Vi 2007.
28. Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. Nat Genet 2005 Apr;37(4):435–40.
29. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 1988 September;44(3):837–45.
30. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform 2005 Oct;38(5):404–15.
31. Wilke RA, Berg RL, Peissig P, et al. Use of an electronic medical record for the identification of research subjects with diabetes mellitus. Clin Med Res 2007 Mar;5(1):1–7.
32. Pakhomov S, Weston SA, Jacobsen SJ, et al. Electronic medical records for clinical research: Application to the identification of heart failure. Am J Manag Care 2007 Jun;13(6 Part 1):281–8.
33. Mannino DM, Homa DM, Akinbami LJ, Ford ES, Redd SC. Chronic obstructive pulmonary disease surveillance—United States, 1971–2000. Respir Care 2002 Oct;47(10):1184–99.
34. Anthonisen NR, Connett JE, Kiley JP, et al. Effects of smoking intervention and the use of an inhaled anticholinergic bronchodilator on the rate of decline of FEV1. The lung health study. J Am Med Assoc 1994 Nov 16;272(19):1497–505.
35. Wedzicha JA, Seemungal TA. COPD exacerbations: Defining their cause and prevention. Lancet 2007 September 1;370(9589): 786–96.
36. Klabunde CN, Warren JL, Legler JM. Assessing comorbidity using claims data: An overview. Med Care 2002 Aug;40(8) (Suppl):IV-26–35.
37. Jeffery PK. Comparison of the structural and inflammatory features of COPD and asthma. Giles F Filley lecture. Chest 2000 May;117(5) (Suppl 1):251S–60S.
38. Punglia RS, D'Amico AV, Catalona WJ, Roehl KA, Kuntz KM. Effect of verification bias on screening for prostate cancer by measurement of prostate-specific antigen. N Engl J Med 2003 Jul 24;349(4):335–42.
39. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med 2005 Oct 27;353(17):1773–83.