PROBABILISTIC GRAPHICAL MODELS IN COPD

Ivy Shi1, Y. Zhang2, D.V. Manatakis1, J. Bon2, D. Chandra2, C. Karoleski2, F.C. Sciurba2, P.V. Benos1
1University of Pittsburgh - Pittsburgh, PA/US, 2University of Pittsburgh School of Medicine - Pittsburgh, PA/US

INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is characterized by persistent airflow limitation, often caused by a narrowing of the airways that makes it difficult for those with the disease to breathe. The third leading cause of death worldwide, COPD is an incredibly heterogeneous disease, with multilevel factors impacting the disease (environmental, clinical, biological, genetic) including diverse co-morbidities such as heart failure, lung cancer, and diabetes mellitus [1]. Therefore, COPD is associated with significant variation in imaging results as well as symptomatic and physiologic presentation, and exhibits variability in progression. This makes understanding of COPD on a system-wide level particularly relevant and compelling. Predicting those patients who will remain stable from those who will exhibit disease progression is critical in defining prognosis and selecting patients for potentially expensive interventions. Currently, there is no satisfactory method for progression prediction that accounts for the multifaceted disease aucters.

One possible way to discern how the diverse factors interact and affect COPD is through probabilistic graphical modeling. Graphical modeling represents causal relationships between measured variables, such as weight and presentation of diabetes, based on the conditional (in)dependencies of the data. From the probability distributions of each of the variables, causal relationships can be established and used to form a network representation of our graphical model. Since graphical models provide a fast, accurate computerized method for predicting causal relationships over an enormous number of potentially relevant variables, it is particularly suited for the challenge of analyzing the multitude of potential factors in COPD [2]. A number of methods for learning directed graphs have been developed in the past, but they typically assume that all variables are of the same distribution type: categorical (gender, ethnicity, smoking status, etc.) or continuous (drug dose, age, etc.) [3]. Our lab recently presented a method for learning mixed graphical models (MGM) that can accommodate both discrete and continuous data types, which is especially relevant for COPD, which has a diverse array of potential factors of both data types.

OBJECTIVE

The purpose of this study is to use MGM to learn a directed graph over the multitude of potentially relevant variables affecting disease progression of COPD. The information from this directed graph will then be used to form a mathematical predictive model of COPD disease progression. Predicting patients who will exhibit disease progression from those who will remain stable is critical in selecting patients for potentially expensive and intensive treatment options, and can be used to form more personalized treatment options.

HYPOTHESIS/SUCCESS CRITERIA

We hypothesize that variables directly connected to the COPD progression variable in the graphical model can stably predict the at-risk COPD patient subpopulation. A mathematical model of the following form can be developed to predict COPD disease progression:

\[ c_1X_1 + c_2X_2 + \cdots + c_nX_n = COPD \text{ Progression} \]

where \( c \)'s are constants and \( X \)'s are the values of the variables directed connected to the COPD progression variable. This study will be considered successful if the graphical network can be learned computationally in less than 360 hours, and if the predictive mathematical model can be validated using 10-fold cross-validation (model is learned over 9/10th of the data and used to predict the remaining 1/10th of the data) with < 5% error.

METHOD

A COPD cohort data set was obtained from the UPMC Pulmonary Physiology Lab, derived from 449 patients with COPD, who also returned for a 2-year follow-up visit. The dataset included over 350 variables derived from surveys, physiologic and quantitative radiologic measurements, spirometry measurements (pre/post bronchodilator effects), and blood biomarkers (creatinine, cholesterol, etc.). COPD disease progression was measured through an FEV1 Decline variable, which represented how the patient’s forced expiratory volume over 1 second declined over the 2 year period between visits. As MGM is dependent on a sparsity parameter, an optimal sparsity parameter was found using a Stability Approach for Regularization Selection (StARS): many networks were learned from subsets of the data, and the sparsity parameter that yielded the most stable graphs (edges are conserved over graphs of all subnetworks) was selected as optimal [2]. The mathematical predictive model was then found using linear regression of the variables directed connected to FEV1 Decline, and a 10-fold cross-validation was performed, where the predictive model was learned over 9/10th of the data (405 patient samples) and used to predict the remaining 1/10th of data (44 patient samples).

RESULTS

MGM successfully learned the full graphical model in 126 hours. Figure 1 below shows the first and second neighbors to the FEV1 Decline variable. Variables directed connected (first neighbors) to FEV Decline are: Creatinine, Current Smoking, Diabetes, Functional Residual Capacity (FRC), Gastroesophageal Reflux (GERD), change in airway resistance with inspiration (PostR5 % Change), and Bronchodilator Reversibility (RV % Change). In 10-fold cross validation, a mathematical predictive model using these first neighbor variables was able to reliably predict FEV1 Decline with 1.7% error.
DISCUSSION
As MGM was able to learn the network in 126 hours, less than our 360 hour criteria, the task is shown to be reasonable in computational feasibility. The 1.7% error in 10-fold cross validation also indicates that this model can be reliably used as a predictive tool to identify patients at risk for COPD disease progression.

This network also offers face validity by identifying variables as direct connectors that are expected to be associated with lung function decline, that being continued tobacco exposure “Currently Smoke”, and bronchodilator reversibility “RV % Change”, currently the accepted clinical indicators of at-risk patients. Other connectors are more provocative, and offer unique perspectives. Most notably, three markers of non-pulmonary co-morbidities are direct connectors to FEV1 Decline, that being: “Creatinine” (a biomarker of renal dysfunction), “Diabetes” (history of diabetes) and “GERD” (history of gastroesophageal reflux disease). While each has been previously linked with COPD, such a dominant association with lung function decline is not well described. Such associations, however, are consistent with a systems biology mechanistic approach to disease, whereby activity and interaction in multiple organs rather than a single organ centric approach better defines the potential underlying mechanisms and impact on the patient.

Creatinine, for example, a direct connector to FEV1 decline, is also a hub in our network. The connections within this hub may offer further insights into the mechanistic associations between renal and lung disease. A recent publication proposes a mechanistic link between emphysema and renal dysfunction through RAGE (the Receptor for Advanced Glycation End Products), the receptor of which, rRAGE, is a direct connector to Creatinine in our network. The Creatinine hub is further linked to a number of other important variables and confounders, including the blood biomarker CCP (Clara cell protein) whose association to COPD has been previously reported. In fact, the interaction between CCP and RAGE identified in our network provides incentive to explore relationships between these molecular pathways. Other direct links to Creatinine, including “Anti Hypertensive” (taking medication for hypertension) and “Stroke” may be indicators of a common vascular mechanistic systems link.

Diabetes, another comorbid condition, is also directly linked to lung function decline. Diabetes is directly linked to “CAC” (Coronary Artery Calcium), and “Pulse Pressure” (another marker of peripheral vascular disease), each of which are highly relevant to diabetes and offers potential systemic mechanistic links to lung function decline that may warrant further investigation. In fact “OPG” (osteoprotegerin), a biomarker directly connected to diabetes in our network, is known to be associated with osteoporosis and has previously been implicated as a potential link between diabetes and coronary artery disease. Gastroesophageal Reflux, “GERD”, the final comorbidity variable linked to lung function progression, also has face validity, as lung function decline associated with lung hyperinflation can alter pressure gradients leading to reflux.

Finally, three pulmonary physiology variables are linked directly to COPD progression, namely “RV % Change” (a marker of bronchodilator reversibility), “FRC” (Functional Residual Capacity, a marker of pulmonary air trapping) and “PostR5 % Change” (a complex measure related to change in airway resistance with inspiration). While the association with bronchodilator response is expected from previous studies, the latter 2 associations offer novel insights.

Looking to the future, mechanistic links suggested by this network, such as that between RAGE and CCP, could be further explored with wet-lab experimentation. Independent COPD cohorts may also be found to validate the mathematical predictive models for more rigorous testing.

CONCLUSION
COPD is an incredibly heterogeneous disease, in which predictive models for COPD progression can be highly useful to identify at-risk patients. Our study successfully found a combination of factors that can determine and predict COPD progression. Additionally, for the first time we are able to build a causal network of COPD that combines heterogeneous types of information such as measurements of lung function, symptoms, systemic comorbidities and blood biomarkers with environmental exposures such as ongoing tobacco exposure. The associations found in this network are particularly notable in that they extend previous work describing the important link between non-pulmonary organ comorbidities and lung function impairment, supporting the systems biology paradigm in understanding lung disease activity.

ACKNOWLEDGMENTS
Thank you to Mr. Vineet Ragh and Dr. Dimitris Manatakis for assistance in the data processing.

REFERENCES