PROBABILISTIC MODEL FOR CHRONIC OBSTRUCTIVE PULMONARY DISEASE DIAGNOSIS AND PHENOTYPING USING BAYESIAN NETWORK

Amos Olwendo
Department of Health Informatics
Jomo Kenyatta University of Agriculture and Technology
Nairobi, Kenya
email: aolwend@jkuat.ac.ke

Dr. Leila Shahmoradi and Dr. Khosrow Agin
Department of Medical Informatics
Tehran University of Medical Sciences
Tehran, Iran
email: lshahmoradi@farabi.tums.ac.ir

ABSTRACT
World Health Organization (WHO) reports Chronic obstructive pulmonary disease (COPD) as a leading cause of death worldwide with 90% of COPD deaths realized in middle and low-income countries. COPD prevalence is expected to increase by more than 30% in the next decade unless urgent actions are taken to reduce the associated risk factors. Reports from Canada, Australia, and United Kingdom show that COPD is prone to misdiagnosis for Asthma yet such countries have hospitals equipped with tools and staff to handle COPD. A spirometer is the main diagnostic tool used to diagnose and separate COPD and Asthma. However, the use of spirometry comes with a number of challenges that make them less useful. This contrast triggers the desire to know the level of COPD prevalence in developing countries in addition to the design a diagnostic device that could be used in resource constrained societies. Consequently, this research seeks to provide a model for COPD diagnosis and classification of cases into phenotypes using a Bayesian Network. Model construction was achieved through developing the Bayesian Network structure and instantiating the parameters for each variable. Model performance is validated using neural network application based on the Levenberg-Marquardt algorithm. Results show that a Bayesian Network has successfully differentiated COPD from Asthma and classified COPD cases into phenotypes. In addition, this study seeks to determine the software requirements for the design and development of a medical device for COPD diagnosis. Such a tool may be used for targeted case finding in resource constrained communities for early identification and treatment of COPD.

KEY WORDS
Bond Graph Modeling, Mathematical Modeling, Bayesian Network, COPD

1 Introduction
Chronic obstructive pulmonary disease (COPD) is major public health problem in both developed and developing countries yet the disease is both treatable and preventable. COPD is the 3rd leading cause of mortality worldwide and is projected to rank 5th in terms of the burden of the disease by year 2020. As a result, COPD is lately receiving a lot of attention from public health stakeholders globally. However, in low income countries, COPD is relatively ignored or unknown and very few hospitals use spirometers for COPD diagnosis. A number of COPD cases experience misdiagnosis, a situation that results in patients being subjected to inappropriate health management strategies [1, 2, 3, 4, 5].

COPD is a respiratory system challenge primarily associated with cigarette smoking, air pollution (both indoor and outdoor), aging, alpha-1-antitrypsin (A1At) among other risk factors for its development and progression. Moreover, COPD is not a single disease but an umbrella term used to describe progressive lung diseases including emphysema, chronic bronchitis, refractory (non-reversible) asthma, and some forms of bronchiectasis. COPD is therefore a composite of lung diseases; obstructive pulmonary diseases that result into blocking small airways and/or parenchyma destruction (due to oxidative stress) hence limited airflow in and out of the lungs. The projected increase in COPD prevalence is first due to the increasing; aging population which is a result of better health care and nutrition. Secondly, the projected increase in COPD prevalence will result from the increasing sources of air pollution from industrial wastes and chemical fumes. Finally, the projected increase in COPD prevalence will also be as a result of the increasing number of female and teenage smokers [6, 7, 8, 9, 10].

1.1 GOLD
Global Initiative for Chronic Obstructive Lung Disease (GOLD) is a worldwide organization mandated to develop strategies for prevention of COPD. GOLD provides guidelines for COPD diagnosis and assessment, treatment and management strategies, and COPD comorbidities. COPD diagnosis is not a simple task that requires a differential diagnosis of a number of diseases. Due to the complex nature and variations among COPD cases, GOLD offers three main recommendations for managing COPD. The first addresses treatment objectives directed towards relieving and reducing the effects of symptoms and objectives that are tailored towards reducing the risks of undesirable future
health events. The second recommendation addresses the need for classifying COPD severity which is a predictor of future undesirable health events such as exacerbation, hospitalization, and death. Third, GOLD recommendation addresses development of simple and reliable questionnaires to be used in assessing patient symptoms and the risks for future adverse health events. These patient management strategies are to be applied in a clinical settings to assists in providing the right therapy to patient based on individual needs. However, GOLD has not yet come up with any efforts to identify obstructive lung diseases that make up COPD as phenotypes [7, 8, 9].

Moreover, diagnosing COPD is not an easy task due to slow progression of its associated symptoms hence COPD is prone to inadequate diagnosis that eventually results into inaccurate morbidity and mortality data. Misdiagnosis of COPD also results into policy implications given that an appropriate management of COPD improves the patient’s quality of life hence reduced exacerbation and need for hospitalization. COPD management strategies are formulated based on phenotype. Clinically, identifying phenotypes of a disease means differentiating the disease cases based on appearance, function, and/or behavior. That is, COPD phenotypes would classify patients based on distinct attributes. A given patient may be suffering as a result of a single or a combination of one or so COPD phenotypes. However, to this day, there are no agreed upon COPD phenotypes. Instead, GOLD has guidelines through which clinicians should arrive to their own COPD patient management strategies. Such guidelines acknowledge that a COPD patient may be suffering due to a combination of asthma and COPD (a condition known as asthma COPD Overlap), or a single or a combination of the other constituent diseases as shown in Figure 1. Currently, COPD diagnosis requires a combination of spirometry results which measures lung functioning, symptoms and challenges patients experience during their daily lives. To effectively diagnose COPD, it is recommended that clinicians consider a review of the underlying risk factors the patient might have been exposed to the time of diagnosis. Surprisingly, a number of spirometer users usually end up making less confident interpretation of results due to infrequent use [1, 7, 8, 11, 12, 10].

1.2 COPD in Developed Nations

COPD is a burden experienced worldwide. In the United States of America (USA), COPD is the 3rd leading cause of death. However, cases of COPD misdiagnosis are not a common phenomena in USA. On the other hand, countries like Canada, China, UK, Australia, New Zealand, Spain, India, South Africa, Brazil, and Mexico show high prevalence coupled with the challenge of misdiagnosis. The high COPD prevalence in developed countries is not only as a result of high smoker populations but also high levels of air pollution. Additionally, COPD is also considered as an old-age disease as it is mainly associated with people 40 years or older. Therefore, high prevalence of COPD in developed countries is also attributed to high aging populations which is a result of better standards of living and health care services. However, reports also show that under-utilization of spirometers leads to lack of confidence in spirometer calibration, use and interpretation of results thus misdiagnosis of COPD in such developed nations [7, 8, 3, 5, 13].

1.3 COPD in Africa

To date, COPD has not attracted much of the attention of health care policy makers and stakeholders in the African continent. However, studies raise concerns for the lack of and/or under-utilization of spirometers in the public health care systems to accurately diagnose COPD and asthma among sub-Saharan African countries. This is probably due to the costs associated with acquisition and use of spirometers. However, in the last couple of years, there have been increased interests in investigating the prevalence and burden associated with COPD in a few sub-Saharan Africa. Such studies report that COPD and asthma prevalences are not only on the rise but also the high level of neglect attributed to COPD. The few studies conducted were in Cape Town, South Africa, where COPD prevalence and its possible association between pulmonary tuberculosis (PTB) were investigated. Other studies have also been conducted in various populations in Tanzania, Malawi and Rwanda [14, 15, 2, 16, 17, 18, 19].

1.4 Probabilistic Reasoning

Advancements in artificial intelligence (AI) and machine learning (ML) research have led to a greater percentage of human life being channeled towards the control and monitoring through automated processes [20]. Such advances are minimizing the need for human input in a number of applications. Probabilistic reasoning has a lot to offer in
medical diagnosis given the nature of the many challenges encountered. The main task in probabilistic reasoning is to identify all relevant variables $x_1, \ldots, x_n$ in a given environment and create a model $P(x_1, \ldots, x_n)$ that mimics the interactions of such variables. Inference is then performed by introducing evidence and then calculating probabilities of interest based on the given set of evidence. As a matter of fact, medical diagnosis is prone to incomplete vague observations made by patients that are sometimes exaggerated yet the physicians are expected to achieve the best outcomes for the patient [21]. In addition, a physician’s decisions about a patient’s next test follows from observed current status of events. Probabilistic reasoning does tend to offer an ability to capture and adequately represent patient symptoms, risk factors and other important variables to be used for disease diagnosis by assigning a degree of belief to every observation.

2 Methodology

2.1 Study design

A cross-sectional study conducted at the respiratory unit at Loghman Hakim Referral and Teaching Hospital, south of Tehran. Simple random sampling was used to identify 100 known COPD and Asthma cases respectively and were confirmed using GOLD “Key Indicators for Considering a Diagnosis of COPD” and “Considerations in Performing Spirometry” guidelines as shown in tables 2.1 and 2.3 respectively[7]. The data set comprised of cases of patients suffering from asthma, asthma COPD overlap syndrome, chronic bronchitis, emphysema, and general COPD. Data collection was achieved through interviews using a self administered questionnaire. Questions were organized in a structured linguistic form to make it easy for patients to be able to describe their health conditions. For example, a question on shortness of breath, was categorized as: High, Moderate, Low or None. To ensure that the patient’s choice was as close as it could get to the above choices, patients to state the time period over which they had experienced such a condition and assert a number value in the range 0 to 10. Data analysis began by randomly assigning 60% of the data set for training the network and 40% for testing. To ensure an even distribution and random assignment of each patient cases to either the training and the testing sets, the data set was grouped based to the disease and assigned each case a unique identification number. Thereafter, generated random numbers for each group and obtained fractions for training and testing sets using the percentages specified above.

2.2 Building the Bayesian Network

The process of developing our model was iterative and was started by consulting the expert and also studying through the relevant literature. Given the fact that we wanted to achieve the best results, we had to conduct interviews for the COPD patients. This involved developing a structured questionnaire based on our network representation. So, we ended up with a questionnaire of ten questions in addition to patient age and a Bayesian network of five risk factors (Smoking, Exposure, Age, Allergy, and Family History of either Asthma or COPD), four diseases (COPD, Asthma, Chronic bronchitis, and Emphysema), and lastly six symptoms (Dyspnea, Cough, Sputum, Chest pain, Body Activity, and Wheezing). Also, each of the variables has constituent parameters distributed in a representation based on degrees of belief.

2.3 Micro Models

One of the major challenges you may encounter when dealing with Bayesian networks involves dealing with large probability distributions. The first solution to such a problem lies in the use of micro model techniques which include the Noisy-Or model, deterministic CPDs, tree CPDs, sigmoid CPDs among others. One of the micro models above that stands out in medical diagnosis problems is the Noisy-Or model given its simplicity [20, 21]. Using the Noisy-Or, a number of diseases $X$ can give rise to some symptom $y$ provided at least one of the diseases is present. We can therefore build a conditional probability $P(e \mid C_1, C_2, \ldots, C_n)$ as $P(e \mid C) = 1 - \prod_{i=1}^{n} 1 - P(e \mid C_i)$

Figure 2. Noisy-Or design

Figure 2 an example sketch diagram showing the design of the Noisy-Or micro model. In a Noisy-Or CPD, each parent $C_i$ is capable of establishing an effect on the effect $E$ except under some unusual circumstance by the suppressor $Q_i$. In addition, the leak variable is meant to represent all other causes of the effect $E$ that are not included in the model. However, we can simplify the Noisy-Or model by eliminating all the network variables that have single children as long as their absence will not compromise network accuracy. The act of eliminating such variables is known as bypassing and it leaves us with a simplified design as shown in the Figure 3 below.
2.4 Levenberg-Marquardt algorithm

To validate BN results, we developed a neural network application trained using the Levenberg-Marquardt (LM) algorithm. The LM algorithm serves as an intermediate optimization algorithm between the Gauss-Newton method and gradient descent algorithm and address the limitation of each of those techniques.

2.5 Model evaluation

We used precision, recall and F1-Score to evaluate the model, where precision is defined as (1) and recall defined as (2) and F-Score defined as (3).

\[
\text{Precision} = \frac{\#\text{Correct}}{\#\text{Guessed}} \quad (1) \\
\text{Recall} = \frac{\#\text{Correct}}{\#\text{Actual}} \quad (2) \\
F_1 \text{ Score} = 2 \left( \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \quad (3)
\]

Figure 4. Precision, Recall, and F1 Score

Precision is a measure of a classification system’s ability to present only relevant items. Recall is the measure of the system’s ability to correctly classify all relevant items. Precision and recall results of classification systems are usually inversely proportional. That is, the higher the precision, the lower the recall and vice versa. However, our main objective of determining precision and recall is to achieve a compromise of the two measures in order to avoid presenting results that may be biased. F measure, also known as F-score or F1 score is used to get a trade-off; the harmonic mean of precision and recall. F1 score of 0 is said to be very low and therefore performing poorly and an F1 score of 1 is as well achieving the maximum performance. Therefore, an F1 score of precision and recall attempts to achieve an acceptable average for the system performance with regards to data classification. In other words, with precision and recall, we attempt to find a balance where a classification algorithm achieves a balance of both the true positive and true negative results.

3 Results

In this section, the result obtained using both the Bayesian network model and neural network model based on LM algorithm using the test data set are presented. The results have obtained through the model a relatively close to the results obtained for each case through medical diagnosis.

3.1 Results using the Bayesian Network

The Bayesian network classified 40 Asthma and COPD test cases respectively as summarized below.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Number of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>34</td>
</tr>
<tr>
<td>Asthmatic</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 1. BN Classification of 40 Asthma test cases

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Number of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>1</td>
</tr>
<tr>
<td>Asthmatic</td>
<td>13</td>
</tr>
<tr>
<td>Chronic bronchitis</td>
<td>5</td>
</tr>
<tr>
<td>general COPD</td>
<td>21</td>
</tr>
</tbody>
</table>

Table 2. BN Classification of 40 COPD test cases

3.2 Results using Neural Networks

Also, the neural network model classified the same 40 Asthma and COPD test cases respectively as summarized below.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Number of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>34</td>
</tr>
<tr>
<td>Asthmatic</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 3. NN Classification of 40 Asthma test cases

3.3 Model classification Accuracy

We used precision and recall to calculate the F1 score which is used as metric to evaluate classification algorithms. F1 score values (0, 1) and the closer the score is
to 1 the better. Both techniques achieved the same precision results. However, the accuracy and F1 score for the Bayesian network are much better compared to the LM algorithm.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Number of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>2</td>
</tr>
<tr>
<td>Asthmatic</td>
<td>2</td>
</tr>
<tr>
<td>Chronic bronchitis</td>
<td>25</td>
</tr>
<tr>
<td>general COPD</td>
<td>10</td>
</tr>
<tr>
<td>None</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5. Model accuracy, precision, recall, and the F1 score

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN</td>
<td>0.99</td>
<td>1</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>NN</td>
<td>0.96</td>
<td>1</td>
<td>0.93</td>
<td>0.96</td>
</tr>
</tbody>
</table>

4 Discussion

Diagnosis and phenotyping of COPD is a difficult task that requires the application of the laid down guidelines and procedures for one to obtain reliable results. A Bayesian network (a probabilistic graphical modeling technique) has proven to provide a model for medical diagnosis. A number of studies have attempted to create models for COPD diagnosis. However, none of such studies has attempted to classify COPD cases into phenotypes. This study presents a model of COPD diagnosis and phenotyping. Results show a successive relation and overlap of asthma between COPD. The Bayesian network and the NN applications achieved 100% classification of the 40 Asthma test cases. Also, BN classified 40 COPD test cases as 39 COPD and 1 Asthma. On the other hand, NN classified 37 out of the 40 test cases as COPD, 2 as Asthma and 1 as neither Asthma nor COPD. The variation in the classification for both techniques is related to their reasoning approach. For instance, we used a probability of 0.20 as the threshold for the BN. Therefore, for a given case with the BN indicating a probability of 0.87 for COPD, 0.77 for Chronic bronchitis and 0.18 for Asthma, using 0.20 threshold, we would declare such a case as belonging to the chronic bronchitis phenotype. However, the LM algorithm classified some cases with probabilities as low as 0.12 as asthmatic COPD. One of such cases if one with an identification ID of 200. According to BN, the network shows that COPD has a probability of 0.845 with chronic bronchitis being 0.685, and 0.124 Asthma. Therefore, by lowering the threshold for the BN to lets say 0.10, such a case would be adequately diagnosed.

4.1 Observations

Results show that Bayesian network could be a better tool for modeling medical diagnosis for overlapping cases. Some of the observation we have made include; (1) the Bayesian network is can be employed not only to diagnose COPD and Asthma but also to classify COPD cases based on phenotypes; general COPD, chronic bronchitis, emphysema, and asthmatic COPD. (2) cardDiag application results confirm that we can go beyond classifying COPD cases to observe the overlaps in COPD cases. (3) COPD is mainly common among smokers; 94 out of 100 COPD cases were active cigarette smokers. We also observed that 93 out of 100 COPD cases were associated with patients 40 years of age or older. For the COPD cases with the patient aged less than 40, we observed that such patients were involved in heavy smoking or high exposure to at least one of the risk factors. COPD is also associated with cough that starts scantily and worsens with the progression of the disease. Cough is most likely associated with chronic bronchitis and may not be observed in patients with emphysema. Also, spumum is also common in COPD cases especially for those suffering from chronic bronchitis. Sputum normally starts out as colorless and becomes yellowish-brown and bloody based on the severity of the disease with blood being the most severe. However, may not be observed in patients with emphysema and asthma cases usually show colorless sputum unless it is accompanied with an infection. Finally, COPD is greatly associate with dyspnea also known as shortness of breath that results into inactivity that worsens over time. Dyspnea is common in all COPD cases may be more severe among those suffering from emphysema.

5 Conclusion

This research provides a model for diagnosing and classifying COPD based on phenotypes; General COPD, Chronic bronchitis, Emphysema, and the Asthmatic COPD using a Bayesian network (BN). A BN is a probabilistic modeling tool composed of random variables and the relationships of such variables is based on probabilities that maximize certain outcomes. We validated our BN model using a neural network model based on the Levenberg-Marquardt algorithm. Results show that the BN model achieved an overall classification of 98.75% for our test cases with an F1 score of 0.99. Furthermore, the Levenberg-Marquardt algorithm-based neural network application also had acceptable results based on the expectation of 95% confidence level with an interval of 5%. Furthermore, this study seeks to develop a medical device that could be used for targeted case finding and diagnosis of COPD. Specifically, such a device would be found useful in resource constrained societies that may not have lung specialists to handle COPD cases. Moreover, case finding would help in identifying COPD cases or persons at risk of developing COPD early in time.
References


