**CSC 529**

**Sarah Cummings**

**Winter 2016**

# Case Study 3: Using Bayesian Networks to Analyze Factors Contributing to Disease

## Introduction:

According to the National Center for Biotechnology Information, Bayesian Network models are often some of the best models in predicting disease. Their systematic review of Bayesian Networks in this application, "Applying Naive Bayesian Networks to Disease Prediction", concludes Bayesian Networks are fundamental to the health field, allowing us to understand the relationships between factors relating to different diseases like never before (NCBI). By relying on the probabilities of different events occurring simultaneously, Bayesian Networks allow us to discover the dependency and independency of variables and conclude causation among variables despite their potentially complex relationship. The review, which include 23 studies, also found that predicting diseases based on a Bayesian Networks had the best performance in most diseases in comparison with other algorithms, with the highest reported accuracy.

For this case study, I will be using a hospital patient dataset that contains information on 10,000 patients who have been tested for lung cancer and tuberculosis. Both tuberculosis and lung cancer are serious diseases affecting the lungs, and the distinction between the two is difficult without first running a medical examination. In the data set, we also have information on whether or not the patient is a smoker, whether or not he or she has traveled to Asia, and an indication of if they are experiencing dyspnea or bronchitis. The goal is to model and visualize the relationships between these Boolean variables, and develop a better understanding of both lung cancer and tuberculosis.

**Data Description:**

This data set, first detailed by Lauritzen and Spiegelhalterin in 1988, has 10,000 rows of patient data and contains the following eight Boolean variables:

1) Smoker: indicator variable for if the patient is a smoker
2) Lung Cancer: Indicator variable for lung cancer
3) Visit to Asia: Indicator variable for if the patient has recently visited Asia
4) Tuberculosis: Indicator variable for tuberculosis
5) Tuberculosis or Cancer: A variable indicating if the patient has lung cancer or TB
6) X-Ray: Indicator variable for an abnormal lung x-ray
7) Bronchitis: Indicator variable for bronchitis
8) Dyspnea: indicator variable for dyspnea; whether or not they have difficult or labored breathing

**Data Cleaning:**

In this dataset, there are no missing values so we do not need to perform any preprocessing or data cleaning. Despite the fact that are variables are unbalanced, we won't want to resample to correct for this because that would affect the Bayesian network results. The only modifications I made to the dataset was renaming the columns with single letters so I could more easily reference them and so that the attribute names would fit appropriately on the graph. The abbreviations are as follows: A= Recent visit to Asia, S= Smoker, L= Lung Cancer, B= Bronchitis, T= Tuberculosis, E= Either Lung Cancer of TB, D= Dyspnea. This renaming schematic is only used in the experimental results section.

**Data Analysis:**

In an initial analysis, we will examine the variance of the different variables. In the tables below, we see the distribution of the categorical variables.

Patients who smoked:

```
> table(df$Smoker)

  no  yes
5031 4969
```

Patients with lung cancer:

```
> table(df$LungCancer)

  no  yes
9438  562
```

Patients who had recently been to Asia:

```
> table(df$VisitToAsia)

  no  yes
9906   94
```

Patients with TB:

```
> table(df$Tuberculosis)

  no  yes
9903   97
```

Patients with TB or Lung Cancer:

```
> table(df$TuberculosisOrCancer)

  no  yes
9349  651
```

Patients with abnormal xrays:

```
> table(df$X.ray)

  no  yes
8897 1103
```

Patients with Bronchitis:

```
> table(df$Bronchitis)

  no  yes
5549 4451
```

Patients with Dyspnea:

```
> table(df$Dyspnea)

  no  yes
5716 4284
```

As seen in the tables, we have 562 patients with lung cancer and 9438 without. We also have 97 patients with tuberculosis, and 9903 without. This leads to a total of 651 patients with tuberculosis or cancer. These features are unbalanced, but we will keep them unbalanced as that will be an important feature in the bayesian network model. Below, we examine the differences in these proportions across several other variables.

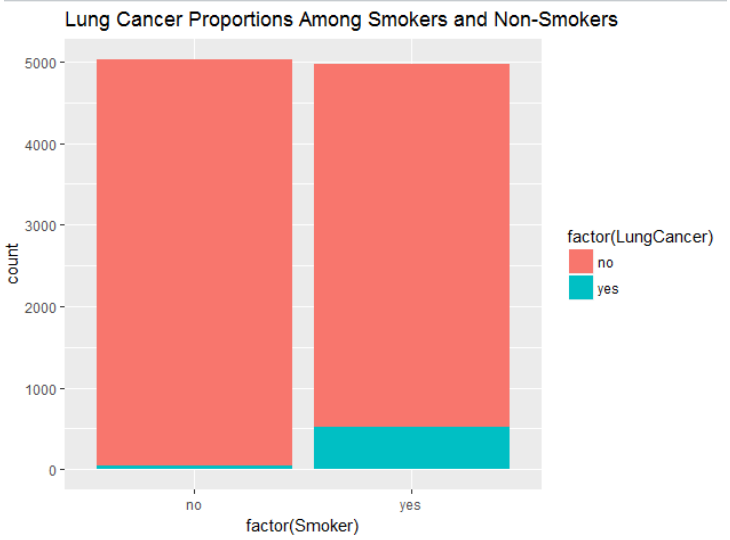



**figure 1**

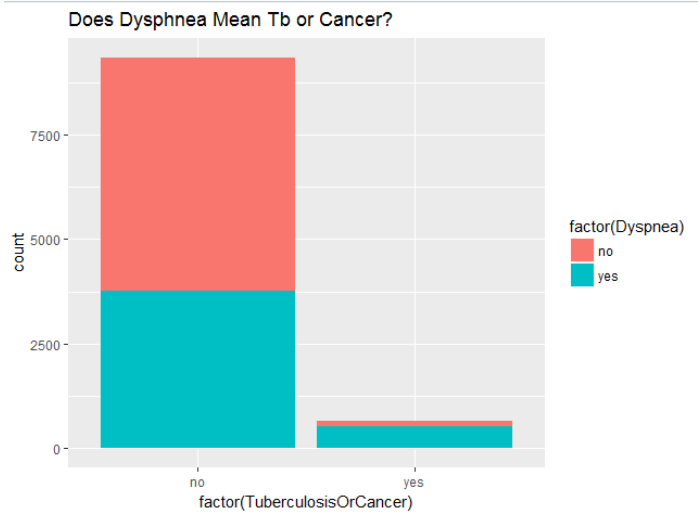



**figure 2**

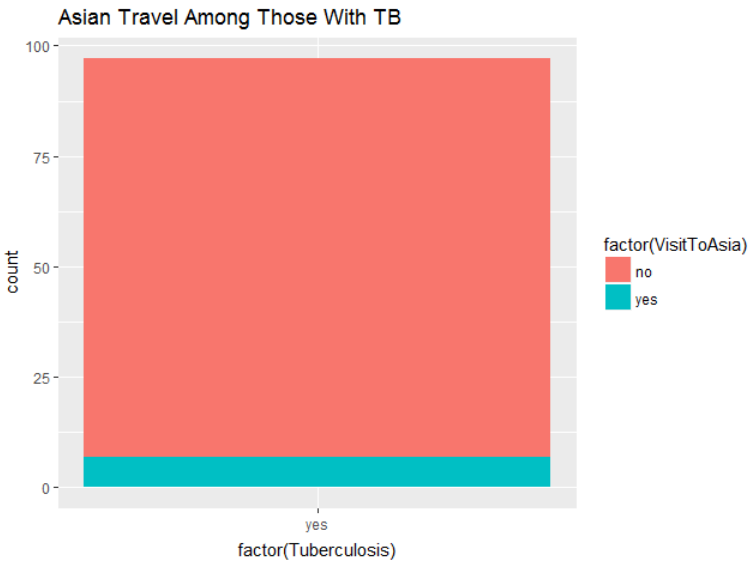



**figure 3**

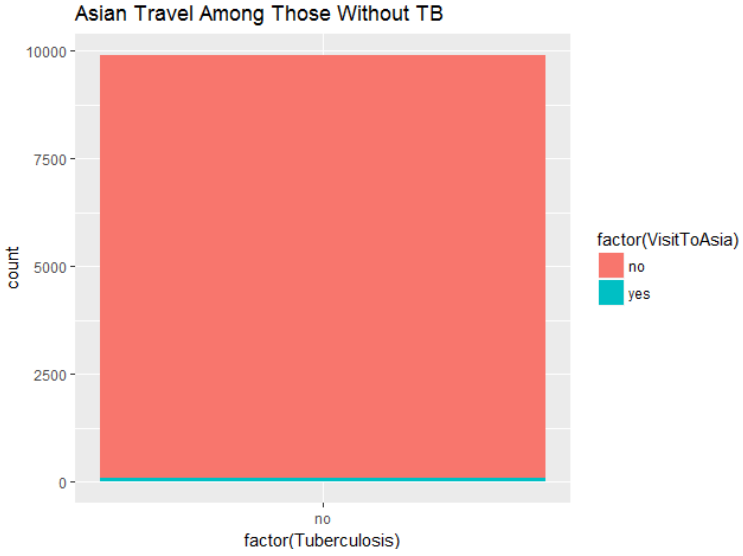
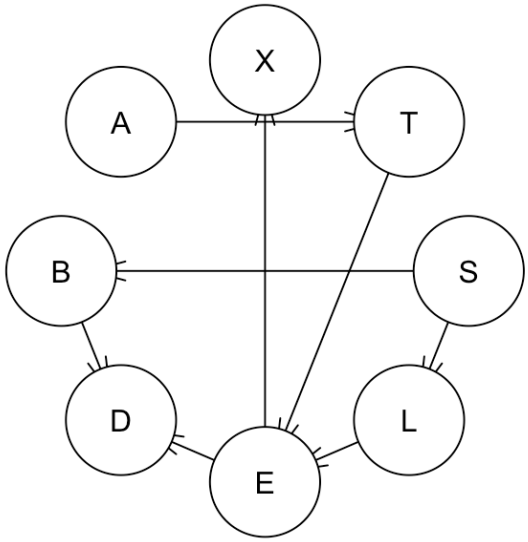



**figure 4**

As seen in figure 1, a significantly higher proportion of smokers are diagnosed with lung cancer as compared with non-smokers. As expected, smoking appears to be a risk factor for lung cancer. Figure 2 shows us that while a significant proportion of tuberculosis and lung cancer patients experience Dyspnea, the symptom is not always indicative of something as serious as lung cancer or TB.As seen in figures 3 and 4, those with tuberculosis are more likely to have gone to Asia than those without Tuberculosis, indicating that going to Asia is potentially a risk factor for TB.

**Experimental Results:**

To model and better understand the relationships among these variables, I will be using various types of bayesian networks. First, I will use our knowledge from the data analysis and what we learned about the relationships among variables to create a random generated network. The results were as follows:

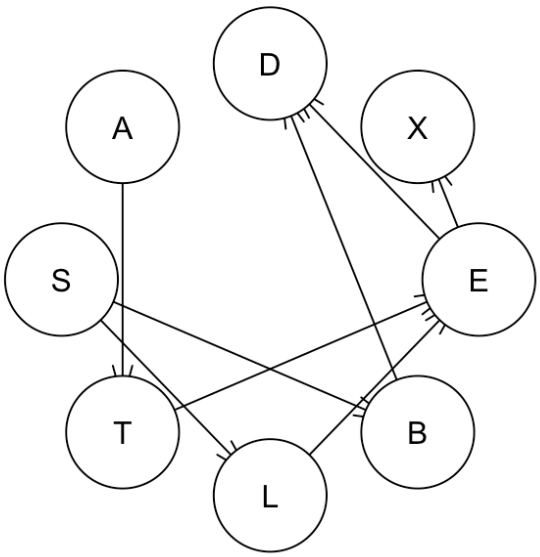



```
Random/Generated Bayesian network

model:
 [A][S][B|S][L|S][T|A][E|L:T][D|B:E][X|E]
nodes:                                  8
arcs:                                   8
  undirected arcs:                      0
  directed arcs:                        8
average markov blanket size:            2.50
average neighbourhood size:             2.00
average branching factor:               1.00

generation algorithm:                   Empty
```

AIC score:-22295.75

Recall: A= Recent visit to Asia, S= Smoker, L= Lung Cancer, B= Bronchitis, T= Tuberculosis, E= Either Lung Cancer of TB, D= Dyspnea.

Next, I experimented with grow shrink, without feeding the algorithm the expected to-from relationships among the variables. The algorithm did not note any connection between going to Asia and getting TB. Using the IAMB learning algorithm produced the same model as the grow- shrink method, as did the fast.iamb method. I used a hillclimbing method next, with AIC as my scoring method. The hillclimbing method looked a lot more to what we would have expected:




```
Bayesian network learned via Score-based methods

model:
 [A][S][T|A][L|S][B|S][E|T:L][X|E][D|B:E]
nodes:                                  8
arcs:                                   8
  undirected arcs:                      0
  directed arcs:                        8
average markov blanket size:            2.50
average neighbourhood size:             2.00
average branching factor:               1.00

learning algorithm:                     Hill-Climbing
score:                                  AIC (disc.)
penalization coefficient:               1
tests used in the learning procedure:   84
optimized:                              TRUE
```

As seen in the network graph above, a recent visit to Asia increases likelihood of tuberculosis. Smoking leads to long cancer and bronchitis, either lung cancer or TB can cause an abnormal chest X-ray, and all three of lung cancer, TB, and bronchitis may cause dyspnea. The model summaries are detailed in the table below.

| Model | Nodes, Arcs | Avg Marov Blanket | Avg Neighborhood | Avg Branching Factor | Score |
|---|---|---|---|---|---|
| **Generated** | 8, 8 | 2.50 | 2.0 | 1.0 | -22295.75 |
| **Grow-Shrink/ iamb** | 8, 5 | 1.75 | 1.25 | 0.5 | Partial directed graph |
| **Hillclimbing** | 8, 8 | 2.50 | 2.0 | 1.0 | -22295.75 |

**Experimental Analysis:**

Based on our results, and what we know about our data, the hillclimbing method for creating a bayesian network was the most successful. As expected, visiting Asia may increase your chances of TB, and smoking most definitely increases your chances of lung cancer. It is difficult to determine what is causing a patient's lung troubles based on symptoms and chest X-rays alone, as Bronchitis, TB, and lung cancer all may cause dyspnea and an abnormal chest x ray. Our hill climbing model also allows us to generate some conditional probabilities, using cpquery. Our model generated the following probabilities, based on our 10,000 patients:

The probability of having tuberculosis: 0.0072

The probability of having lung cancer: 0.0558

The probability of having bronchitis: 0.4494

Thus, the majority of our patients were experiencing bronchitis, and lung cancer or tb or more rare. We can also use our model to find the likelihood of different diseases given a set of conditions. Suppose we have a patient who has recently visited Asia, and does not smoke, then:

The probability of him having tuberculosis:0.05882353

The probability of him having lung cancer: 0

The probability of him having bronchitis: 0.3125

We find that the patient is most likely experiencing Bronchitis, but she may also have TB. It is very unlikely that she has lung cancer.

Next, we will examine the difference in chances of being diagnosed with lung cancer based on whether or not the patient is a smoker:

The probability of having lung cancer as a smoker: 0.1092437

The probability of having lung cancer as a non-smoker: 0.008768434

As expected, patients who are smokers are much more likely to develop lung cancer, with 10% of them having the disease as compared to less than 1% for the non-smokers.

**Conclusion:**

In conclusion, we found that naive babes models are in fact useful for modeling health data, particularly symptoms and diagnoses. The hillclimbing model was the most successful, as it gave us a detailed understanding of the relationships between these diseases and their symptoms. It included all 8 nodes, 8 directional arcs, and had an AIC score of -22295.75. Though the initial generated model had the same AIC score, its model only included 5 arcs among the 8 nodes, not painting a full picture of our data. Finally, we found that going to Asia increased one's chances of contracting tuberculosis, smoking increased chances of lung cancer, and dyspnea and abnormal X-rays may be a result of bronchitis, lung cancer or tuberculosis. This model can help diagnosis patients when their symptoms may be a result of any of the three previously mentioned diseases. In the future, I would include additional symptoms for modeling these diseases, as well as include more data or test the model against new data.

**Citation:**

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5203736/