International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 6, December 2018, pp. 5245~5252
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i6.pp5245-5252

Analysing Tuberculosis Trends in South Asia

Kumar Abhishek, M. P. Singh, Md. Sadik Hussain
Department of Electrical and Computer Science and Engineering, National Institute of Technology, India

Article Info

Article history:

Received Sep 5, 2017
Revised Jul 28, 2018
Accepted Aug 11, 2018

Keyword:

Data wrangling
Regression
Tuberculosis

ABSTRACT

Tuberculosis (TB) has been one of the top ten causes of death in the world. As per the World Health Organization (WHO) around 1.8 million people have died due to tuberculosis in 2015. This paper aims to investigate the spatial and temporal variations in TB incident in South Asia (India, Bangladesh, Pakistan, Maldives, Nepal, and Sri-Lanka). Asia had been counted for the largest number of new TB cases in 2015. The paper underlines and relates the relationship between various features like gender, age, location, occurrence, and mortality due to TB in these countries for the period 1993-2012.

Corresponding Author:

Kumar Abhishek,
Department of Computer Science and Engineering,
National Institute of Technology, Patna
Patna-800005, Bihar, India.
Email: kumar.abhishek@nitp.ac.in

1. INTRODUCTION

Tuberculosis, abbreviated as TB, is an infectious disease caused by Bacteria Mycobacterium tuberculosis (MTB). The most common route of transmission is airborne or droplet infection.People with active Pulmonary TB spread infectious aerosol droplets of 0.5 to 5.0 μm in diameter while coughing, sneezing, speaking, singing, or spitting. Each one of these droplets may transmit the disease, since the infectious dose of tuberculosis is very small (the inhalation of fewer than 10 bacteria may cause an infection).

The lungs are usually affected by TB (Pulmonary TB) but other body parts such as Lymph nodes, kidney, bone, brains (Extra Pulmonary TB) can also be affected. It may be fatal if not treated properly as it is a stigmatizing disease. People in some countries still believe that it is an incurable condition. It is very difficult to treat the patients if they do not follow the regimen tightly. There are only few antibiotics that can be used against it and resistance emerges readily. The vaccines of TB are not completely effective. MTB spreads by respiratory droplets and dies in sunlight. Transmission does not happen immediately, but requires some prolonged exposure. We may infer that it is associated with poor dark and dinghy living conditions. The person who is at higher risk of TB includes:

- A person with weak immune system.
- A person with poor nutritional status.
- Drug addicts.
- A person suffering from HIV infection or diabetes.

TB is one of the major factors of death and disability in the world. According to a report by WHO, there are approximately 9.6 million cases of TB , in which 2.2 million cases are in India only. The TB cases in India accounts for $340 billion to the Indian economy. TB is also a leading killer of people affected by HIV resulting for about 35% of the deaths. The main causes of TB epidemic are illiteracy, improper resource distribution, improper diets and lack of proper medication.

According to a report, about 49 million lives of TB patients were saved between 2000-2015 through TB diagnosis and treatment. The objective of this paper is to analyze TB trend in South Asia because this can help by providing proper knowledge of the cases, their reasons, geographical distribution, etc..

2. RELATED WORK

Many similar works have been done on Tuberculosis prediction. Many scholars adopted different methodologies.Noodchanath *et. al*. modeled the incidence of TB cases in Thailand [1]. Their aim was to design a model which analyzes the trend, seasonal and geographical incident of TB cases in southern Thailand from 1999 to 2004. They have used Negative Binomial Distribution for gender, age, location and for other variable Log-linear Distribution was used. After getting the distribution they have applied Linear regressing to best fit according to distribution. After the analysis, they realized that both male and female has the same risk of TB having age less than 25. With the increase in the age, in both males and females risk also increases. There is not any seasonal trend in the TB disease cases. There are long-term seasonal changes in TB cases from 1999 to 2004. The Geographical locations also affect TB cases. In upper western and lower southern there is a high risk of TB cases.

Sampurna *et. al.* also worked on a similar project [2]. They have modeled TB incident in Nepal. In Nepal, the TB case rate is very high, so their objective was to model incident of TB in Nepal from 2003 to 2008. They used gender and location as a parameter to analyze the TB cases. To define the distribution, they used the Negative Binomial technique. After the distribution, they used linear regression to get the trend of TB cases and predict the result with two multiplicative components as predictors. They found that there are major drops in TB cases during these five years in Nepal. Average TB cases in male were 1.31 per 1000 population, which is very less and in the case of a female, it is 1.81. In Nepal region play a vital role in the case of TB. In Terrain region, it is higher followed by Hill and Mountain region. There is a decrease in trend, but still, a total number of TB cases is very high. The Higher rate is in Terrain region and urban area. They analyze TB case on a long term basis.

Sampurna *et. al*. analyzed the temporal and spatial variation of TB cases in Nepal [3]. Data was collected from 2003 to 2010. Data is modeled by using gender, age, location, year using linear regression model with log transformation of the rate of the parameter. They removed some outlier to get a good fit of the data in the model. HIV was also considered for getting the variation in the model. It is seen that HIV case leads the increase in TB cases. The rate of TB cases varies higher in Terrain region and TB cases, variation of location and years.

A work on data acquisition and analysis of solar energy generation was carried. The parameters considered for analysis were namely (1) average (2) maximum and (3) total amount of electricity generated on daily and monthly basis. Prediction was performed after analysis using ANN model [10].

The above analysis is focused on a particular country with location, age , sex as parameters. They also consider the seasonal and long-term trend. The proposed work analyses TB trends in all countries of South Asia. The parameters considered are Age, Gender, Location and HIV Cases as HIV reduces a person's immunity resulting in an increased chance of TB . To fill missing values general statistics like mean and median are not used instead a machine learning model is trained to predict the missing values.

3. DATA ANALYSIS

The data for analysis has been obtained from a WHO report on TB for all the countries. The raw data of six South Asian countries viz. India, Bangladesh, Sri Lanka, Pakistan, Maldives, and Nepal have been scrutinized. Some data were predicted using the method of Linear Regression. The raw data was processed through the following processes.

a. Data cleaning
b. Data validation
c. Data wrangling
d. Linear regression

The whole data set was classified methodologically to determine the country wise new pulmonary cases based on sex, location, and HIV positive cases.

4. PROBLEM FORMULATION

The dataset available have some missing values, negative values, and Outliers. So the dataset is cleaned from such unwanted values. The linear model is prepared and analysis is done by R language. The work is done step by step which follows:-

a. Find out negative value and fill them with null values.
b. Split dataset into an available dataset and missing dataset.
c. Prepare a linear model
d. Apply outlier removal algorithm
e. Predict missing values
f. Filter relevant data
g. Analyze dataset by R query

Some algorithms are applied to the dataset to analyze and obtain a meaningful result. First of all the negative value is found out and replaced by null value. Thereafter dataset is split into two parts. The first part is correct data and the second part is missing data. Algorithm 1 is used for replacing missing values with null value

Algorithm1 : fillNegativeValue
Input: dataset
Output: dataset without negative value

```
FOR row IN dataset:
    FOR column IN row:
        IF(dataset[column]<0):
```

From the correct dataset, a model is designed so that the missing values can be predicted. Some related values which were available in missing value row are passed as an input to the prepared model and we get data for missing values. This predicted value is not actual value, but we try to obtain values with maximum efficiency. To test the accuracy of the model from corrected dataset, two sub-datasets are found. One is training data and the second one is test data. Train data is used to train the model and test data is used to test the accuracy of the model. Some variations are done in the parameters of the model, so that a good datamodel can be obtained. Alogirtm2 is used for this purpose

Algorithm 2 : trainModel
Input: dataset
Output: linear data model

```
data=read_csv(dataset)
reg=linear_model.LinearRegression()
RETURN reg
```

The prepared model is a Linear Regression model. This is the model which uses some existing features by which a linear equation is obtained representing the predicted value. The equation is obtained with less Residual Square Sum (RSS) which is sum of the square of the difference between actual value and predicted value.RSS is used to predict continuous values. In this project, all values are continuous real numbers. Thus, the linear model is really helpful.

There are some outlier values in the dataset, which can not be corrected by linear regression. Outlier produces more RSS value, due to which, linear regression is away from a good fit. So an algorithm is applied to increase the accuracy of the model. Algorithm 3 is applied to a training dataset and each time some outlier is removed and the remaining dataset model is re-trained. This process is repeated until maximum accuracy is obtained .

Algorithm 3: removeOutlier
Input: prediction, feature, target
Output: outlier removed dataset

```
test=np.array([])
x=np.append(test,features)
x=x[:90]
y=np.append(test,target)
y=y[:90]
z=np.append(test,predictions)
z=z[:90]
temp=np.array([x,y,abs(z-y)])
result=temp.transpose()
result.sort()
length=len(result)
```

```
start=math.floor(length*0.1)
cleaned_data = result[start+1:]
RETURN cleaned_data
```

From above process, a better accurate model is obtained. Now, all missing data are predicted using algorithm4.

Algorithm4 : prediction
Input: dataset,input
Output: predicted value

```
train,test=train_test_split(dataset,test=3)
lm= linearModel(dataset)
    WHILE(accuracy<expected_accuracy)
    DO
        clean_data=removeOutlier(lm.predict(test[features]))
        lm.fit(clean_data[features],clean_data[target])
RETURN lm.predict(clean_data[input])
```

The result shows that the prediction of missing values using the machine learning model are more precised than the general statistical model. After getting the missing values, the dataset becomes complete and accurate. Since analysis is done only for South Asian Countries with some features (Year, Age, Region, and New TB Cases etc.), so from the whole dataset required dataset is obtained. Then all the data sets are analyzed to grab the information using R Language.

For different countries, we get a different linear equation and according to that linear equation, the missing values are predicted. Some of them are mentioned as follows -

For Bangladesh,

$$y = 5683q - 11317460$$

where y is new smear-positive case and q is the year for which values are going to be predicted. Similarly,

For India,

$$y = 25749q - 51139722$$

For Maldives,

$$y = -5q + 10091$$

For Nepal,

$$y = -30(q - 2010)^2 + 1556$$

To convert the above equation into the linear equation, let $(q - 2010)^2 =$T

so, $y = -30T + 1556$

For Pakistan,

$$y = (698(q - 1998)^2) + 2567$$

let $(q - 1998)^2 = T$

so,

$$y = 698T + 2567$$

For Sri Lanka,

$$y = -5(q - 2009)^2 + 4764$$

let $(q - 2009)^2 = T$

so,

$$y = -5T + 4764$$

Similarly, after finding a linear equation for all variable, missing values are predicted. Now, on these data sets analysis work is done and some facts are found out which are explained in the result analysis section.

5. RESULTS AND ANALYSIS

The analysis is focused on TB data for South Asian Countries (mainly India, Pakistan, Nepal,Bangladesh,Maldives) from WHO for a period of 1993-2012. The analysis shows that India has the highest cases registered for new smear-positive TB where as the Maldives has the least cases reported for new smear-positive TB(depicted in Figure1). Considering gender wise data for new smear-positive TB cases, male between the age group of 35 to 44 are most to have recorded for new cases of TB as shown in Figure 2. Females between age group 15-24 years recorded the highest cases for new smear-positive TB as shown in Figure 3.

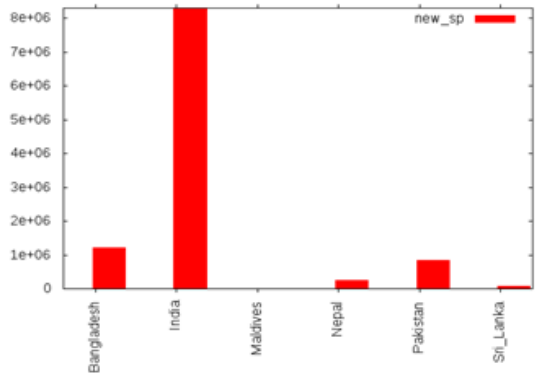

Figure 1. New smear-positive case country wise

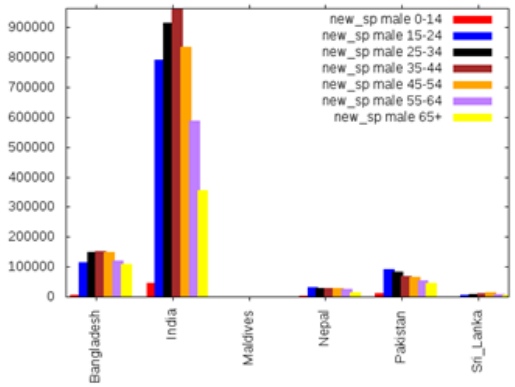

Figure 2. New smear-positive case country wise for male age wise

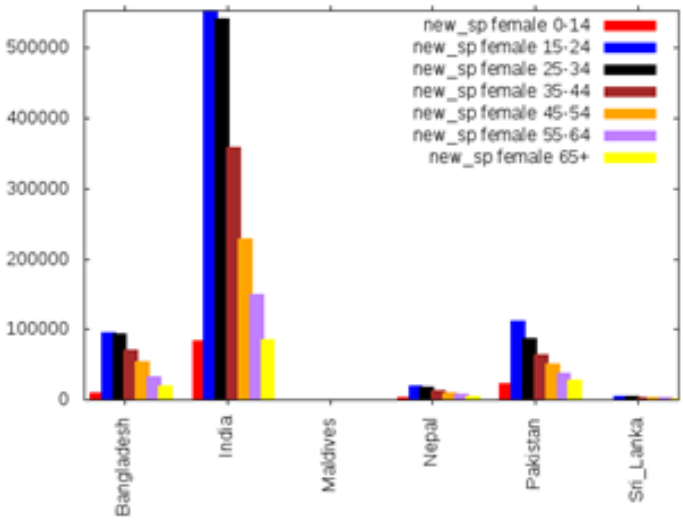

Figure. 3. New smear-positive case country wise for female age

With respect to Figure 1, Figure 2 and Figure 3 India has the highest number of new smear-positive TB cases collectively as well as gender wise also India recorded the highest cases, followed by Bangladesh, Pakstan, Nepal, Sri Lanka. Maldives has the least number of new smear-positive TB cases collectively for the whole population as well as gender-wise.

Figure 4 depicts that there has been a constant increase in the number of patients registered under new smear-positive TB cases every year despite the efforts of WHO and respective countries toward effective implementation of the Stop TB Strategy.

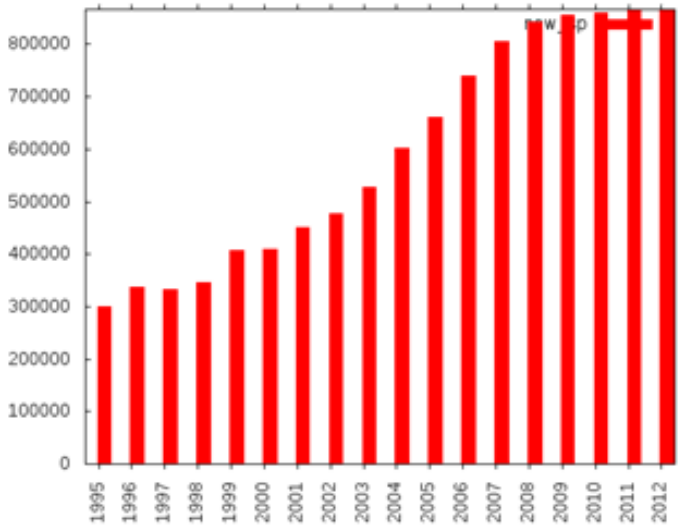

Figure 4. New smear-positive case year wise

According to Figure 5, India counts for the highest number Return Relapse cases followed by Bangladesh, Pakistan, Sri Lanka, Nepal. The return Replase case indicates those patients who are identified as smear-positive tab and they have missed their treatment regime. The Patients identified as Return relapse cases have a high chance of MDR (Multi-Drug Resistant) TB which in recent year has been increasing.

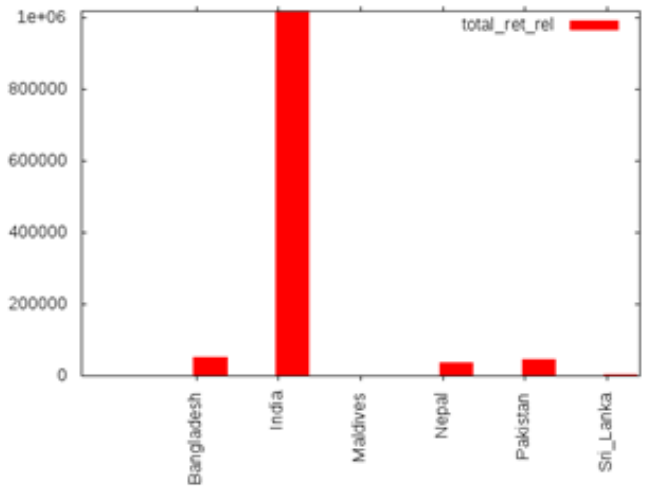

Figure 5. Return relapse case country wise

According to Figure 6, it can be seen that after 2011, Return relapse cases start decreasing due to awareness about medical cures for TB. WHO executes lots of awareness programs which pays positive effect on Return relapse cases.

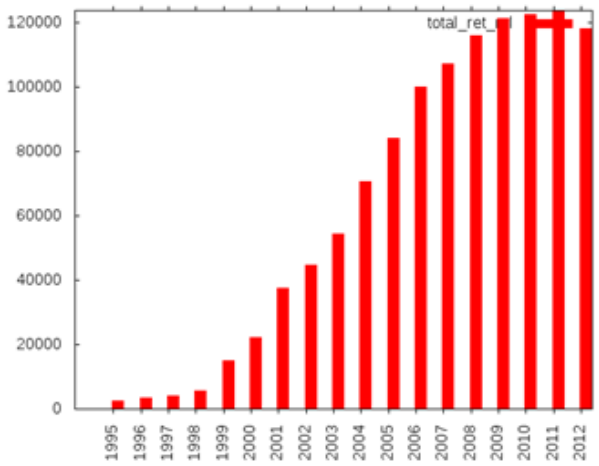

Figure 6. Return relapse case year wise

People suffering from HIV have 26 to 31 times greater chances of TB than people without HIV. With respect to WHO report in 2014 ,9.6 million new cases of TB were registered out of which 1.2 million were people with HIV. The people suffering from HIV have high chances of getting TB if they are in contact with TB suffering people because of their reduced immunity. The TB in HIV people can be cured with proper treatment regime and surveillance. With respect to Figure 7 India has more than 50 percent of HIV registered patients are affected by TB. According to Figure 8, the increase in the percentage of cases of TB/HIV is almost constant after 2009.

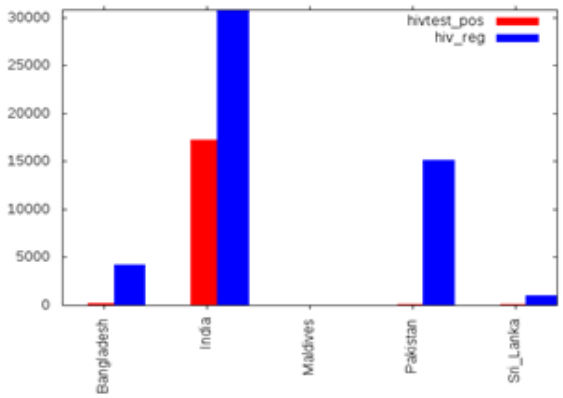

Figure 7. HIV and TB case country wise

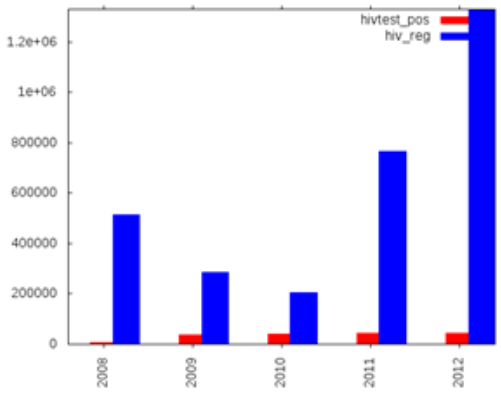

Figure 8. HIV and TB case year wise

6. CONCLUSION AND FUTURE ENHANCEMENT

The paper shows various variations in the TB cases based on the sex, location, year, and HIV-affected cases. It can be used as an aid for reconciliation and a rethinking of the approach adopted earlier and the further improvement needed to accomplish the goal of 1case per million per year of WHO by 2050. The presented variation can help in focusing and taking a targeted approach for the basement of the worst affected region. This paper provides an analysis of data related to South Asia with a limited number of parameters which have been which have been taken into consideration. In future dataset can be analyzed for all countries of the world with some more parameters. The same work can be done for analysis of other diseases.

REFERENCES

[1] Kongchouy N, Kakchapati S, Choonpradub C. "Modeling the incidence of tuberculosis in southern Thailand". *Southeast Asian Journal of Tropical Medicine and Public Health*. 1; 41(3):574, 2010.

[2] Kakchapati S, Choonpradub C, Lim A. "Spatial and temporal variations in tuberculosis incidence, Nepal". *Southeast Asian Journal of Tropical Medicine and Public Health*. 1; 45(1):95, 2014.

[3] Kakchapati S, Yotthanoo S, Choonpradup C. "Modeling tuberculosis incidence in Nepal". *Asian Biomed*. Nov 8;4(2), 2010.

[4] Golub, J.E., Saraceni, V., Cavalcante, S.C., Pacheco, A.G., Moulton, L.H., King, B.S., Efron, A., Moore, R.D., Chaisson, R.E. and Durovni, B. The impact of antiretroviral therapy and isoniazid preventive therapy on tuberculosis incidence in HIV-infected patients in Rio de Janeiro, Brazil. *AIDS (London, England)*, *21*(11), p.1441, 2007.

[5] Murray CJ, Salomon JA. "Modeling the impact of global tuberculosis control strategies". *Proceedings of the National Academy of Sciences*. Nov 10; 95(23):13881-6, 1998.

[6] Cohen T, Murray M. "Modeling epidemics of multidrug-resistant M. tuberculosis of heterogeneous fitness". *Nature medicine*. Oct; 10(10):1117, 2004.

[7] Castillo-Chavez C, Song B. "Dynamical models of tuberculosis and their applications". *Mathematical biosciences and engineering*. Sep 1; 1(2):361-404, 2004.

[8] Blower SM, Chou T. "Modeling the emergence of the'hot zones': tuberculosis and the amplification dynamics of drug resistance". *Nature medicine*. 2004 Oct 1; 10(10):1111.

[9] Dye C, Williams BG. "The population dynamics and control of tuberculosis". *Science*. May 14; 328(5980):856-61; 2010.

[10] Lee J, Kim SB, Park GL. "Data Analysis for Solar Energy Generation in a University Microgrid*". International Journal of Electrical & Computer Engineering*. Jun 1; 8(3), pp. 2088-8708, 2018.

[11] Ahmad T, Arif S, Chaudry N, Anjum S. "Epidemiological Characteristics of Poliomyelitis during the 21st century (2000-2013)". *International Journal of Public Health Science (IJPHS)*. Sep 1; 3(3):143-57, 2014.

[12] Zignol M, Hosseini MS, Wright A, Weezenbeek CL, Nunn P, Watt CJ, Williams BG, Dye C. "Global incidence of multidrug-resistant tuberculosis". *The Journal of infectious diseases*. Aug 15; 194(4):479-85, 2006.