The Future of Tuberculosis Control using Precision Medicine Technology and Big Data Concepts

Sanjay Satya-Akunuri Koka[1], Rohith Suba Koka[2], Purushotham Rudraraju[3], Kiran Kumar[4]

Precision Pundits–18417 Stoneridge Ct, Northville MI, USA

Abstract: ***Mycobacterium Tuberculosis, TB, is one of the deadliest diseases in the world, it causes over 1.3 million deaths per year. There has been constant need for innovative research methods on a clinical basis to combat the spread of TB and ultimately lead to the eradication the disease all over the world, these efforts have proved futile as the disease is still active. Therefore, it is imperative to study TB through the uses of Big Data Analytics and precision medicine/predictive intelligence. This project utilized vast amounts of public domain data from the World Health Organization to organize past data and build accurate predictive models based on various key performance indicators to determine the incident rate of TB from the years 2000 to 2030. The project resulted in discovery that TB incident rates in the world would still occur at a high rate in 2030. Due to the amount of data available, it was possible to organize, refine, and implement previous data setsin order to examine trends that may leadfurther action taken in order to incorporate a new public health model for TB-vulnerable countries such as India, China, and Indonesia.***

Keywords: Big Data, Tuberculosis, Precision Medicine, Public Health

1. Introduction

The World Health Organization predicts that by 2050, there will be over 10 billion people in the world. Therefore, the services required by global healthcare systems will have to adapt and advance. Mycobacterium Tuberculosis (TB) is responsible for 1.3 million deaths annually; therefore, it is essential to study present and historical data in order to predict future trends [4]. Predictive analysis is driven by unprecedented access to Big Data and a greater involvement by the healthcare consumer to shape predictive services for greater benefit. The benefits of precision medicine are limitless; unlocking the specific values of patient data allow for efficient decision-making and the overall improvement of care.

Predictive analytics is key to the outcome of this project; techniques involving statistical modeling, machine learning, and data mining were utilized to analyze historical and current data sets from the year 2000 to 2014 to develop accurate predictions regarding TB rates till the year 2030. Key performance indicators used for this project analysis included Gender-wise data, TB with HIV data, Country impact data and age impact data [1]. The subsequent models allowed for deeper studies for future projections of data sets. Exploiting patterns within the data to find risks and opportunities as well as identifying insights for preventing, detecting, and curing TB was reached by using public World Health Organization data on global tuberculosis rates [6]. The major objective of this project was to understand TB and gather information regarding its' demography and prediction capability by using Machine Learning analysis.

2. Literature Survey

Throughout the last decade, data analytics has been constantly evolving to the point where Big Data Analytics has become a staple of business models all over the world. In this project, Big Data Analytics is used to study, enrich, and incorporate petabytes of data into a line-regression algorithm in order to build a predictive model to illustrate the incident rate of Tuberculosis from the years 2000 to 2030 and encourage further growth of public health practices in areas affected. Tuberculosis is a dangerous, deadly bacterial disease that affects a large portion of the population in developed and developing countries. The bacteria will target essential organs in the body, such as the lungs and it can easily be spread through inhalation. There have been many strives in research to quicken diagnosis procedures and develop more effective drugs, but the disease has continued to spread.

There are many research scientists working in a clinical setting to combat and reverse the disease through the use of engineered drugs [4]. However, the best way to combat the disease is to put each community in a public health position in which they can avoid ever contracting the disease. Therefore, there is an immediate need to study the projection of the disease over time and in the future in order to discover if potentially affected committees have the needed resources and education to avoid contracting the disease in the future [6]. This project's results will encourage health agencies around the world to make the needed changes (on a clinical or educational level) in each community in present time in order to eradicate the disease and avoid prolonging the spread of the disease around the world.

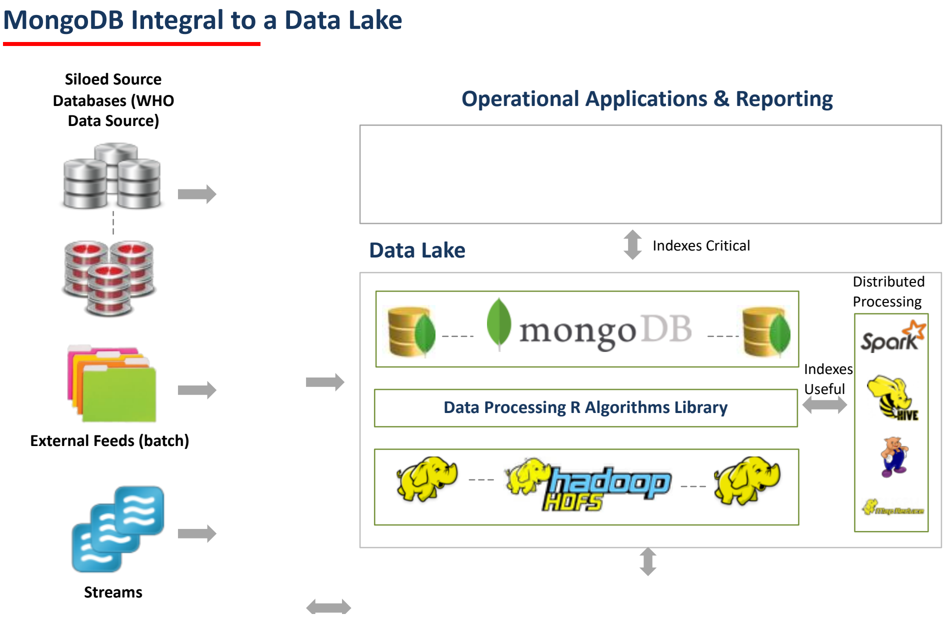
Problem Definition

Based on the World Health Organization's stance regarding the deadly Tuberculosis disease, this project's aimed to utilize variable key performance indicators in order to build statistical models to show the present rates of Tuberculosis and utilize predictive technologies to predict the incident rate of Tuberculosis in the near future of a given sample size [6]. The calculated predictive incident rate of Tuberculosis in the world discovered in this project will most definitely highlight the importance of eradicating this disease through changes in the present-day public health system.

3. Methodology/Approach

Building the prediction algorithm used for predicting TB prevalence rates involved the utilization of a linear regression algorithm. The Linear regression is a statistical method that applies to our data to predict future values [6]. While compiling data, it is essential to obtain a line of best fit, a line in which the total prediction errors are as minimal as possible.

To efficiently utilize Big Data analytics in this project, there was a need to establish the correct architecture solutions in order to realize the potential of the data gathered. The functionality of the program would be to Aggregate cross-silo structured and unstructured data, control data quality, transform data as needed, and report on the data with consistent performance. To ensure high quality, a data lake was used to simplify the end to end data management process [3]. A Data Lake is a flexible data platform for aggregating cross-siloed data in a single location. The data lake allows the user to “mine” and derive insight from the data across the enterprise and from various third-party sources. This project used chose Mango DB integrates with Hadoop distributed processing layers [3]. The architecture is explained in the below diagram:


MongoDB Integral to a Data Lake
Siloed Source Databases (WHO Data Source)
External Feeds (batch)
Streams
Operational Applications & Reporting
Data Lake
Indexes Critical
mongoDB
Data Processing R Algorithms Library
hadoop HDFS
Indexes Useful
Distributed Processing
Spark
HIVE

The data extraction process included gathering all available WHO data from their official website in the form of .csv form files [1]. To effectively utilize big data and create the predictive models, the data was converted into a JSON format and uploaded successful into Mongo database. From there, the data was standardized and enriched using the enterprise data quality concepts such as Deduplication and Data matching. Transformation rules were implemented, and SQL queries were created to extract the refined data from MongoDB for the project’s analytics [3]. To successfully extract and publish the data, application program interfaces were developed to fetch the data from MongoDB using Node [3]. To visualize the subsequent data extraction, d3.js library was used to build a predictive dashboard. Applied the Machine Learning Algorithm (Linear Regression Algorithm) to the existing data and predicted the future data using “R” [6].

4. Results & Discussion

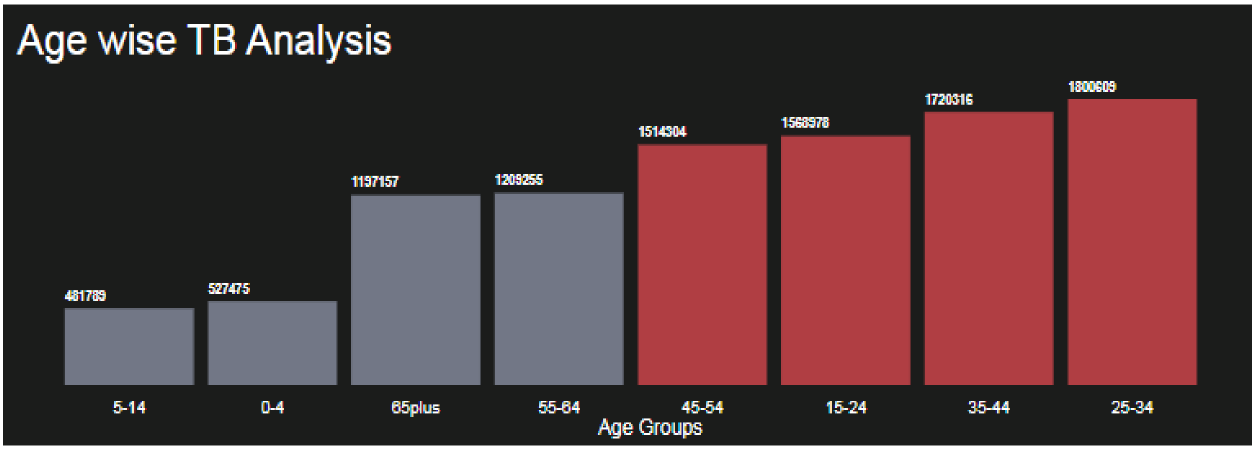


Figure 1: Age Group TB Analysis

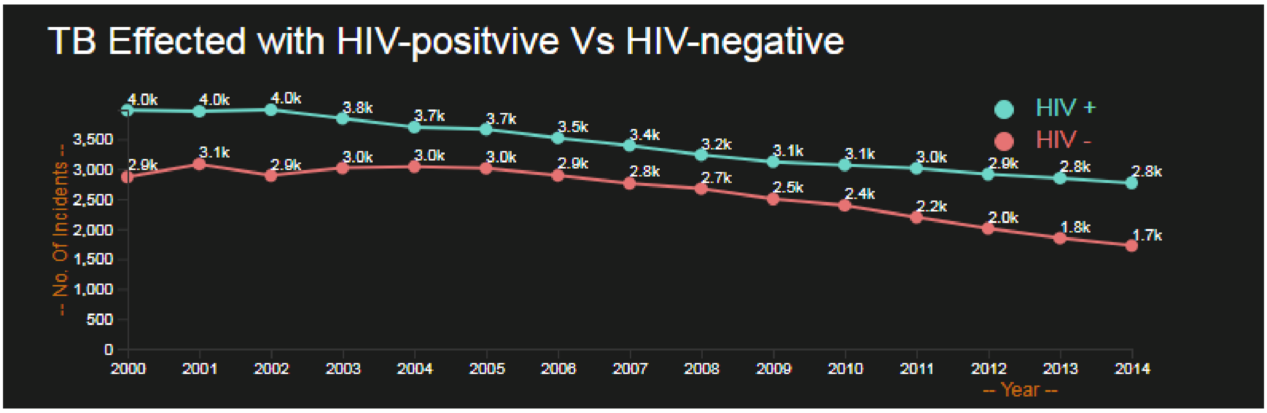


Figure 2: TB with HIV Analysis

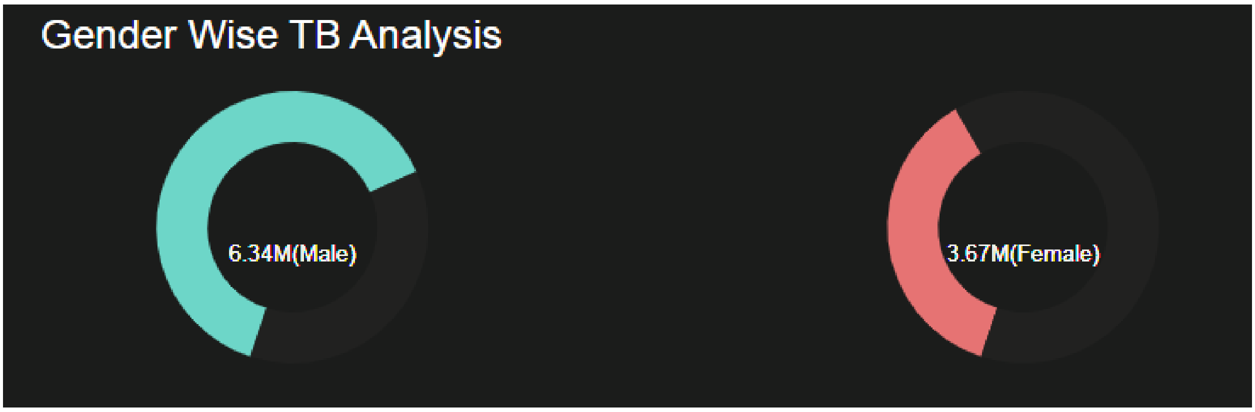


Figure 3: TB Analysis in regard to Gender

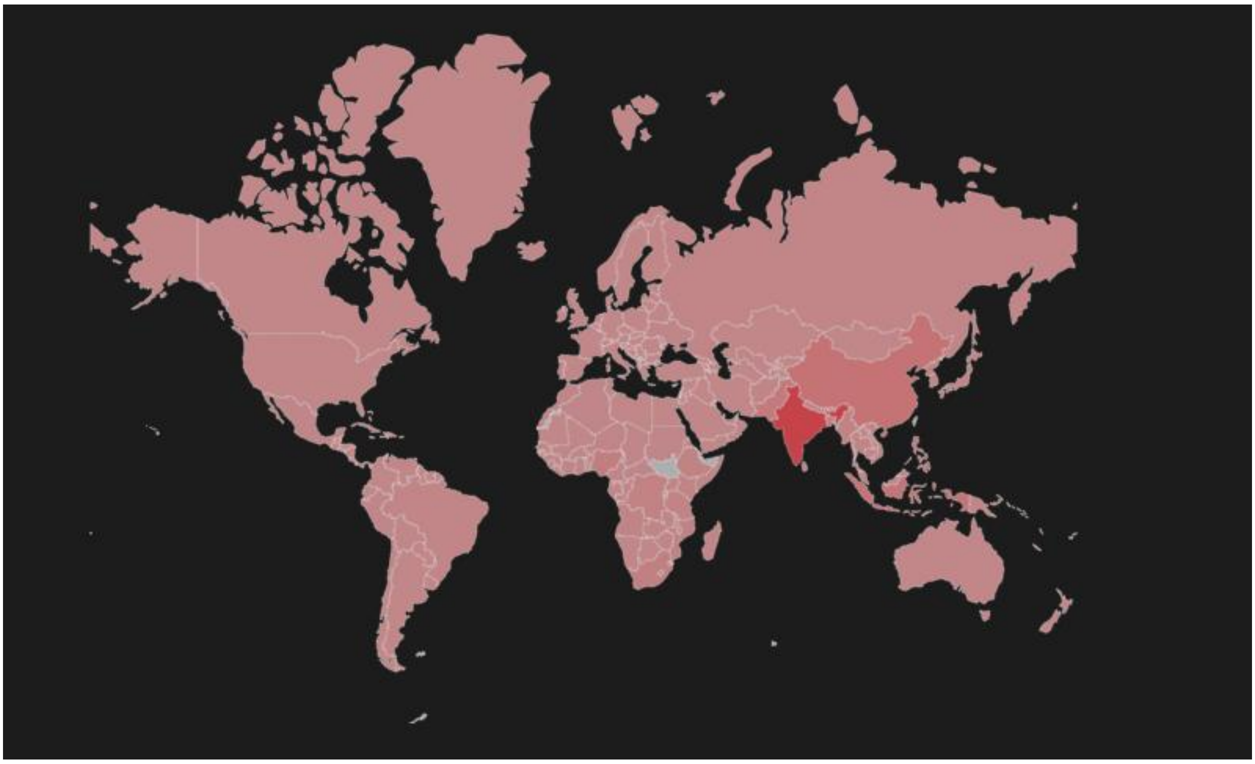


Figure 4: Country wise Analysis

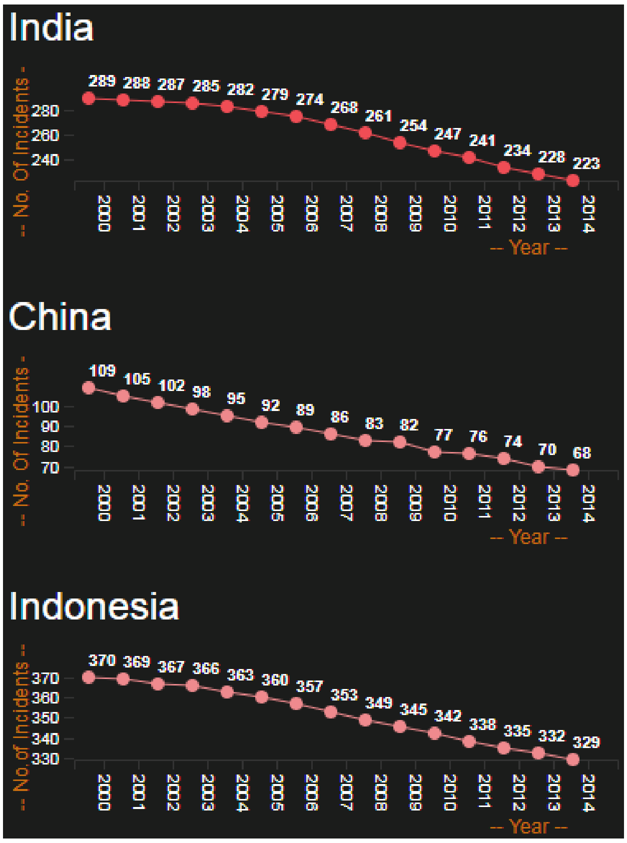


Figure 5: Severity of TB by Country Analysis

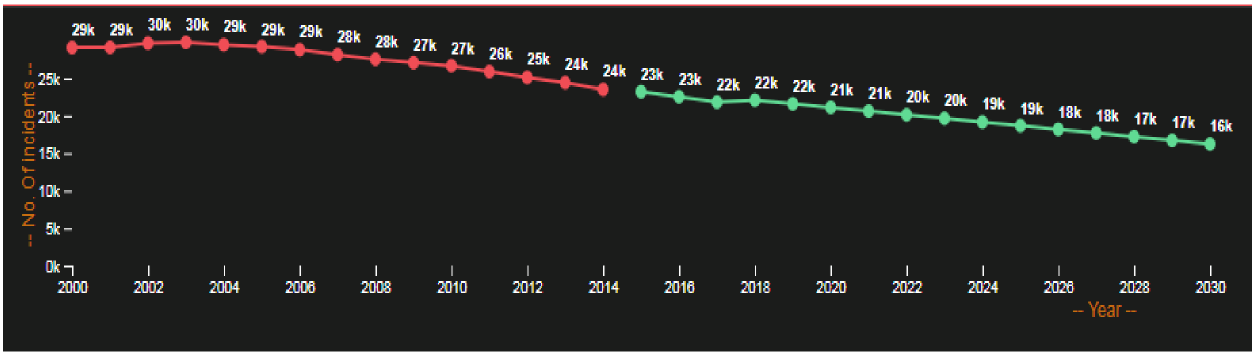


Figure 6: Predictive Analysis using Machine Learning methods

The models were built based on data collected by the World Health Organizationfrom the years 2000 to 2014; the models illustrate historical data that was utilized in our line regression analysis in order to report on TB's future incident rates per 100,000 sample in the world [1]. The project successfully developed accurate data graphs to convey the importance of each key performance indicator. Although, WHO provided an extensive library of data points in relation to TB rates around the world, the data points used in this study were chosen based on WHO's official stance on relevant data in regard to TB prevalence in the world [1]. The key performance indicators used in this study include age demography analysis, gender analysis, Tuberculosis cases with HIV analysis, and Country severity analysis [6].

Figure 1 depicts the age analysis KPI modeland it shows a wide range of ages that correlated with high TB rates. TB rates were highest in the 25-34 age group and lowest in the 5-14 age group as shown by WHO data [1]. Figure 1 was created by analyzing all of the incidents from 2000 to 2014 across the world and organizing each TB incident by age group.

Figure 2 clearly states the impact HIV had on TB rates during 2000 to 2014. HIV positive shows higher rates of TB than HIV negative. As shown in the figure, HIV patients are more susceptible to TB infection, and the number of incidents has gone from 4,000 to 2,800 for HIV+ patients from 2000 to 2014. The number of incidents for HIV-patients from 2000 to 2014 has also gone down from 2,900 to 1,700 incidents [1]. HIV positive patients do have a higher number of TB cases in comparison to HIV negative patients in accordance to our sample, thus suggesting that HIV factors into the susceptibility of the TB disease.

Figure 3 illustrates the gender analysis and clearly states that there were more men affected by TB than woman in the world [6]. The figure shows that there were 6.34 million men compared to the 3.67 million women that were affected by the TB epidemic in accordance with data collected by WHO during 2000-2014 [1].

Figure 4 shows the world-demographic of the TB disease in terms of highest incident rate in countries. According the analysis and the WHO database, the countries Indonesia, India, and China had the highest incident rate of TB per 100,000 sample [1]. Subsequently, figure 5 shows the severity of TB by year in a 100,000 sample from the WHO database in India, Indonesia, and China [1].

Figure 6 depicts a trend line graph of the disease in terms of incident rate by year by a 100,000 sample [1]. Clearly, the rate of TB incidents has and will continuously go down, but by 2030, the epidemic will not be completely eradicated [2]. During the year 2000, there was 29,000 incidents out of the 100,000-sample studied from the WHO database. Our predictive analysis on the 100,000 sample data shows that there would be 16,000 cases of Tuberculosis in the year 2030. Since the disease is still active by 2030, there is a need to reinvent public health policies and strategies; there must be further research conducted to find efficient methods to raise awarenessfor prevention and complete eradication of Tuberculosis by 2030.

The usage of line-regression analysis as a form of predictive analysis fit the best for the types of variable data sets that were utilized in this project. Since, the project used high volumes of data and had to organize and refine the data, line regression analysis was the most efficient and simple predictive algorithm for the creation of the predictive model as shown in figure 6 [5]. Previously, other researchers have also completed predictive research projects but have opted to use different predictive methods based off their own data used in the study. The utilization of the line-regression analysis has successfully allowed this project to show the number of incident rates in a 100,000-person sample size as a prediction for the near future. This project's aim and results are unique because of the methodology and data sets used; the goal of this project is to showcase the state of the Tuberculosis disease in the world and how the disease is affected by different variables. The predictive modeling result now highlights the need for change in present day in order to avoid the predicted trend and eradicate Tuberculosis in the world.

5. Conclusion

The social and economic impact of Tuberculosis is devasting. Tuberculosis afflicted individuals go through poverty, social stigma and severe discrimination. This project successfully integrated the usage of Big data and predictive analysis to analyze and ultimately support WHO's mission to eradicate the TB epidemic by 2030. Predictive analysis used in this project states that TB rates around the world are higher in men, more prevalent in in HIV afflicted patients, and the age wise distribution is between the ages 25 years and 60 years. Our predictive modeling was built through regressive analysis on historical data; the analysis showed that the Tuberculosis infection will persist beyond 2030 [6]. The only way to deviate from this path is for governments and public health organizations, especially in high risk countries like India, China and Indonesia, to approach the TB epidemic with higher emphasis on the development of healthy living, increased concentration on high-risk HIV areasand more emphasis on medical education programs for communities stricken by the deadly epidemic.

6. Future Scope

Although the present study is completed, it has opened the path to many possibilities in terms of future research. The data compiled from WHO database and the usage of predictive technologies clearly indicate the severity of the Tuberculosis disease in the world. Due to the organization of the data using Big Data principles, the three main areas in the world that are affected by TB have identified as Indonesia, China, and India. The usage of predictive technologies confirmed that TB would continue to persist in 2030. This research study used Big Data software in order to compile, organize, and use large volumes of public data taken from the WHO database [4]. From there, the line-regression analysis built a concise prediction model for the study. The line-regression analysis is only one of the possible methods that could be used for a predictive analysis paper. To improve upon future studies, it may be beneficial to utilize multiple methodologies on a common data sample set in order to find which methodology results in the most efficient and organized predictive models [4]. This study's use of line-regression analysis and Big Data platforms such as MongoDB provide the best possible conclusions for a predictive analysis project with the data available [3]. However, it is critical for researchers to continue gathering more data points and find different avenues of clinical support to help eradicate the disease completely. Obtaining more precise data points can allow for more efficient conclusions regarding a larger sample size [6]. This project only had access to a certain set of data that was compiled by WHO, this data was not organized, and this caused minor problems during initial stages of the project. The usage of organized and concise data sets is integral to improving future predictive analysis projects.

References

[1] "Download Data as CSV Files." World Health Organization, World Health Organization, 17 Oct. 2019, https://www.who.int/tb/country/data/download/en/.

[2] Fofana, B. K. (2019). Tuberculosis Surveillance Data Analysis, Volta Region, Ghana. ACTA Scientific Medical Sciences, 3(4), 4–9.

[3] Jain, V., & Upadhyay, A. (2017). MongoDB and NoSQL Databases. International Journal of Computer Applications, 167(10), 16–20. Retrieved from https://www.ijcaonline.org/archives/volume167/number10/jain-2017-ijca-914385.pdf

[4] Sandhu, G. (2011). Tuberculosis: Current situation, challenges and overview of its control programs in India. Journal of Global Infectious Diseases, 3(2), 143. doi: 10.4103/0974-777x.81691

[5] Scally, G. (2004). The importance of the past in public health. Journal of Epidemiology & Community Health, 58(9), 751–755. doi: 10.1136/jech.2003.014340

[6] Wang, J., Wang, C., & Zhang, W. (2018). Data Analysis and Forecasting of Tuberculosis Prevalence Rates for Smart Healthcare Based on a Novel Combination Model. Applied Sciences, 8(9), 1693. doi: 10.3390/app8091693

Author Profile

Sanjay Koka is a Bacc2MD Sophomore student at the University of Toledo studying Biochemistry/Pre-med. As a member of the university, he is a part of the Lambda Sigma Honors Society, Sigma Phi Epsilon Fraternity, Biochemistry research, and editorial publications. He has learned and actively engaged in BIG DATA/Precision Medicine Seminars. He has been fortunate to understand the importance of these two concepts as well as learning about Artificial Intelligence. Moreover, Sanjay worked with Big Data and Medical professionals to understand more about the concepts of Big Data and helped design a “Patient 360” technology model that is associated with global technology companies.