STEM Fellowship Big Data Challenge 2017

How the Immunization Of Youth Influences the Mortality Rates Of Tuberculosis Around The World

Yasha Farhan[1,2], Mikerah Quintyne-Collins[1,3], Jessie Zheng[4], and Adam Skudlarski[5]

[1]Department of Statistics, University of Toronto
[2]Department of Cell and Systems Biology
[3]Department of Mathematics, University of Toronto
[4]Department of Physical and Environmental Science, University of Toronto
[5]Department of Computer Science, University of Toronto

August 31, 2017

Abstract

Millions of individuals around the world are currently living with Tuberculosis (TB), and while there have been many efforts to eliminate it's spread, the TB epidemic will not be eradicated anytime soon. However, timely administration of vaccines for diseases other than TB (such as Measles, Polio and Tetanus) to the youth may bring us one step closer to the eradication of this disease. Correlations between multiple vaccinations of the youth, TB mortality rates, and the deaths of children under 5 years old were obtained. The resulting data implied that, as the number of 1 year old children getting vaccinated for diseases other than TB increases, the mortality rate of TB decreases. Continuing this decrease in TB mortality rates is the key to eradicating TB - the disease that kills millions, but silently affects millions more.

Keywords
Epidemiology, Immunization, Youth, Mortality, Tuberculosis

1 Introduction

With an estimated 10.4 million new cases and 1.4 million deaths reported globally in 2015, Tuberculosis (TB) continues to be ranked as one of the top 10 causes of death worldwide [1], despite vigorous efforts made by global organizations to eradicate it. [2]

TB is an infectious airborne disease caused by the bacterium *Mycobacterium tuberculosis* (*M. tuberculosis*) that is released into the air when an infected person sneezes or coughs. If inhaled, the bacterium may become active, germinate in the lungs and spread to the rest of the body. From the lungs or throat, there is a possibility of it being expelled into the air and infecting others. [3]

To prevent the spread of TB, it is best to catch and treat it as early in life as possible. The World Health Organization's *End TB Strategy* aims to eradicate TB globally, seeking to reduce the number of deaths by 95% from 2015 to 2035. Their strategies involve an immunization effort, with the most modern vaccine proving effective only about 20% of the time. [2] Currently, Bacille Calmette-Guérin (BCG) is the only licensed vaccine for TB and is recommended to be administered at birth, especially in low-income and middle-income countries, which constitute 95% of the attributable deaths by TB. However, several lines of research indicate that low socioeconomic status is related to the delay in the administration of the BCG vaccine, typically due to the WHO's Open Vial Policy. [4][5]

The purpose of this research was to investigate how the administration of various types of vaccines is related to the number of deaths caused by TB, and to the number of deaths of children, infants and neonates in general. To support this investigation, the global death patterns caused by TB in 2015 were observed and several prediction models were tested using data from 2000 to 2014. TB mortality rates, child mortality numbers and immunization data were used as pre-

dictors in these models.

2 Materials & Methods

The primary tools used for the analysis were R (Version 3.3.2) [6] (running in a Jupyter notebook on Apache Spark 2.0.), and RStudio (Version 1.0.143).

The data sets were obtained from the World Health Organization and consisted of TB mortality, child mortality and immunization data, per country, from 2000 to 2015. Immunization data for the following vaccines were obtained: Bacillus Calmette–Guérin (BCG), Measles (MCV1), Haemophilus Influenzae Type b (Hib), Hepatitis B (HepB), Polio, Tetanus and Diphtheria-tetanus-pertussis (DTP3).

The data sets were first cleaned and merged into a master data frame which was inspected for patterns of missing data. Data for BCG, Hib, HepB, and Tetanus were subsequently removed. Imputation was then carried out on the remaining data. The imputation was done using the R package "mice", using predictive mean matching with 50 iterations and 20 multiple imputations.

To test the strengths and significance of the correlations between the mortality rates of TB, the number of deaths of children under the age of 5, and the three types of vaccines, a pairwise scatter plot for these variables was examined.

The algorithm used to predict the number of deaths of TB was linear regression. The models used were a full model that included all predictors (number of deaths of children under the age of 5, infants and neonates, as well as the three types of vaccines) and a null model that included no predictors. Each model was trained using a training set consisting of all the data from 2000 to 2014 and the test set consisting of data for the year 2015. To determine which model performed the best, using the test set, the mean squared error (MSE) was calculated for each model.

3 Results

The pairwise scatter plot obtained is shown in Figure 1. As observed from the plot, there is strong evidence (p-value$<$ 0.001) of a positive correlation ($r > 0$) between the TB mortality rate and the number of deaths of children under the age of 5. Conversely, the TB mortality rate shows strong evidence of a negative correlation ($r < 0$) with each type of vaccine. And while the evidence of a negative correlation between each of the types of vaccine and the number of deaths of children under the age of 5 is very strong, the correlation coefficient is only weakly negative. Finally, each type of vaccine shows a very strong positive correlation with the other, but only with weak evidence.

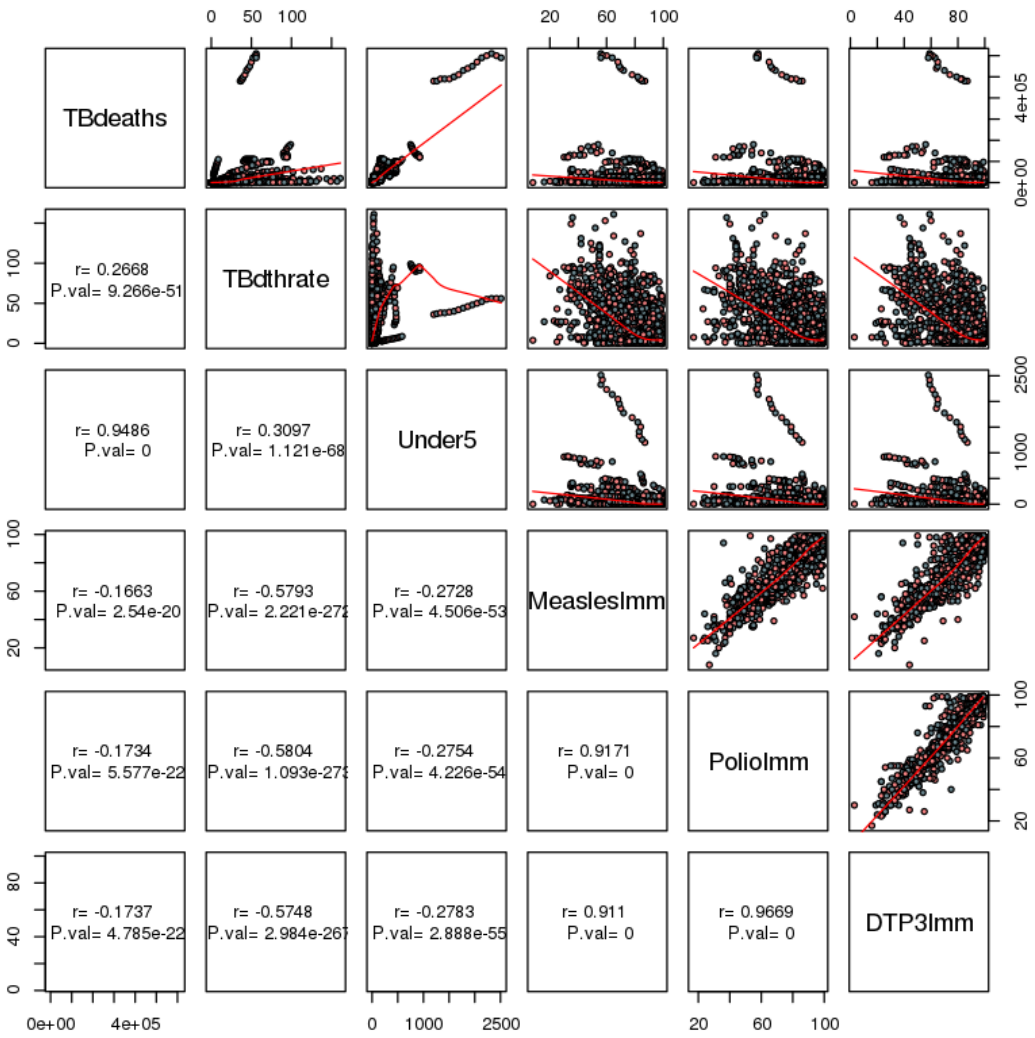
From the results of the model selection, the model that contained all predictors displayed the lowest Mean Squared Error and was chosen as the best model to predict the number of deaths caused by TB. The actual and predicted number of deaths for 2015 were used to obtain choropleth maps. (Figures 3 & 4)

4 Discussion

The results obtained from the pairwise scatter plot offer valuable insights into the correlations between several variables that might otherwise be overlooked during the analysis of TB mortality rates. For clarification, infant and neonatal deaths were excluded from the scatter plot because preliminary correlation analysis showed that they were perfectly correlated with the number of deaths of children under the age of 5. As seen from the scatter plot, the mortality rate of TB is strongly negatively correlated with the number of 1 years olds that are vaccinated against Measles, Polio, Diphtheria, Tetanus and Pertussis. This fact implies that as the number of vaccinated 1 year olds increases, the rate of deaths by TB decreases, thereby emphasizing the importance of the administration of vaccines for diseases other than TB at birth. These implications appear to oppose the observation made by Cheong et al. of a fatal case of TB in a 14 year old girl following measles vaccination in Korea. [8]. The authors of that study, however, stated that no definitive evidence of activation of TB after measles vaccination have been found. While no causal relationships can be drawn from the conclusions of our analysis due to the data being observational, the present study does indeed highlight the importance of timely vaccinations for children under the age of 5.

The greatest limitation of this research was the absence of complete data. Figure 2 displays the proportions and pattern of missing data. Due to the large volume of data missing from the Tetanus, Hib, BCG and HepB3 datasets, these vaccines were removed from the data frame. Without complete data, correlations may not be very accurate and there is an increased room for error. Simply removing the samples with missing data would bias the analysis. Hence, to overcome this, multivariate imputation by chained equations was used. Since all the data are continuous, one imputation model, namely predictive mean matching, was used for all variables. Predictive mean matching was

Figure 1: Pairwise Scatterplot for the data.



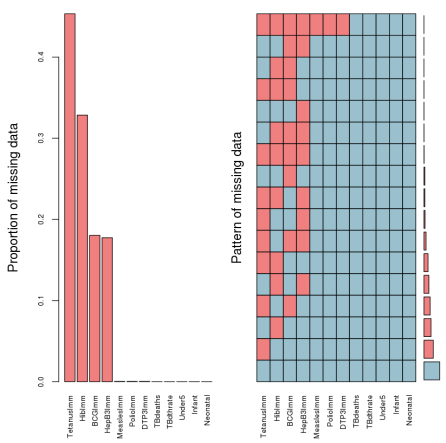


Figure 2: The pattern and proportions of missing data. The taller the red bars, the more data missing.

chosen due to the fact that most of the variables were weakly to moderately correlated with strong evidence and since it works well with variables that do not necessarily follow a normal distribution.

Conclusions

Though the presence of incomplete data poses a challenge to develop an accurate representation and extrapolation of correlations between existing data sets, the findings brought forth from this investigation are not counterproductive. In order to eradicate TB globally, the obstacles that keep TB present must be eliminated. By highlighting the importance of vaccinating children at a very young age, this research provides the world with a method of preventing unwanted diseases in the future. The findings in this investigation indicate the possibility of a wider range of vaccines against TB, other than relying solely on BCG. To extend the study at hand, the testing of vaccines against Measles, Polio, Diphtheria, Tetanus and Pertussis on TB patients and investigating more means of prevention can be done. This will not only solidify the findings from this study, but also provide more options against TB infection and disease in the future.

References

[1] Global Tuberculosis Report 2016 [Internet]. 1st ed. Geneva: World Health Organization; 2016 [cited 22 May 2017]. Available from: http://apps.who.int/iris/bitstream/10665/250441/1/9789241565394-eng.pdf?ua=1

[2] WHO | WHO End TB Strategy [Internet]. Who.int. 2017 [cited 25 May 2017]. Available from: http://www.who.int/tb/post2015_strategy/en/

[3] CDC | Tuberculosis (TB) [Internet]. Cdc.gov. 2017 [cited 23 May 2017]. Available from: https://www.cdc.gov/tb/default.htm

[4] Thysen S, Byberg S, Pedersen M, Rodrigues A, Ravn H, Martins C et al. BCG coverage and barriers to BCG vaccination in Guinea-Bissau: an observational study. BMC Public Health. 2014;14(1). DOI: 10.1186/1471-2458-14-1037

[5] Akmatov MK, Mikolajczyk RT: Timeliness of childhood vaccinations in 31 low and middle-income countries. J Epidemiol Community Health. 2012, 66 (7): e14-10.1136/jech.2010.124651.

[6] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[7] Imputation by Predictive Mean Matching: Promise & Peril [Internet]. Paul Allison; 2015 [cited 26 May 2017]. Available from: https://statisticalhorizons.com/predictive-mean-matching/

[8] Cheong, H.-K., Yang, B. -G., Sohn, Y. M., Kwon, I. -H., Kim, J C., Kim, H., and Kim, J. R. (2010). A Fatal Case of Disseminated Tuberculosis Coincident with Measles-Rubella Vaccination. Epidemiology and Health, 32, e2010002. http://doi.org.my.access.library.utoronto.ca/10.4178/epih/e2010002

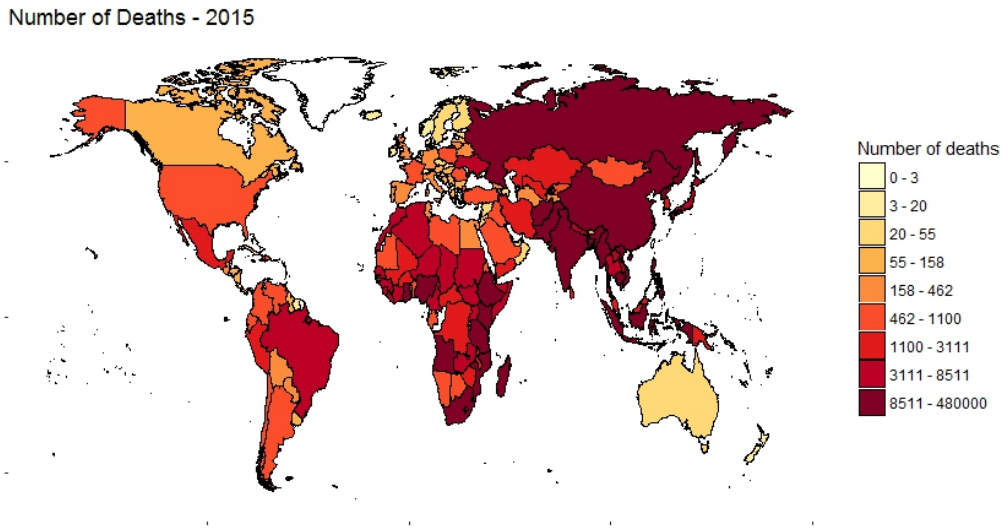


Figure 3: Actual deaths caused by TB, in 2015

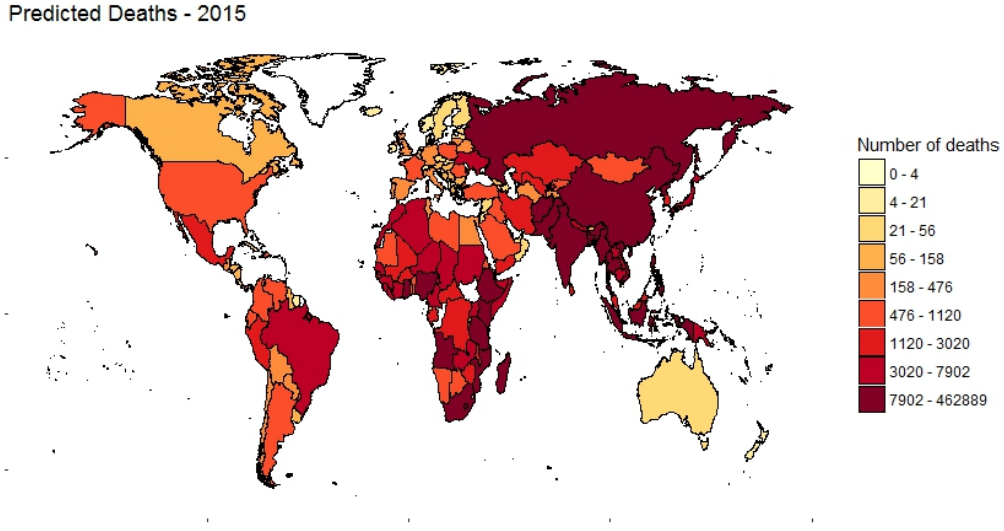


Figure 4: Predicted deaths for TB in 2015, using all the predictors