Preventing Cardiovascular Disease using Data Science

TAN HUAN XI, GREGORY 3S1(27)
OOI XUAN SHAN 3S1(24)
NEO SOUW CHUAN 3S2(17)

Introduction

Cardiovascular disease (CVD) is the number one cause of death globally, taking the lives of around 17.9 million people each year, accounting for about 31% of all deaths.
Our project's objective is to gain insight on this issue by researching the factors behind cardiovascular disease and the reasons for its prevalence using data science and machine learning. We aim to do this by closely analysing patient profiles and demographics.Using the processed data, we hope to spread awareness about cardiovascular diseases to the public, so as to prevent it from claiming more lives.

Case Studies

1.Predicting 30-year Risk of Cardiovascular Diseases

This case study focuses on 30-year long-term prediction algorithms for CVDs as opposed to 10-year risk prediction to better understand public health burden and the true need for intervention. The study followed 4506 participants of the Framingham Offspring cohort aged 20 to 59 years and free of CVD and cancer at baseline examination in 1971–1974 for the development of 'hard' CVD events (coronary death, myocardial infarction, stroke). A modified Cox model that allows adjustment for competing risk of noncardiovascular death was used to construct a prediction algorithm for 30-year risk of hard CVD. Cross-validated survival C statistic and calibration $\chi 2$ were used to assess model performance. The 30-year hard CVD event rates adjusted for the competing risk of death were 7.6% for women and 18.3% for men. Standard risk factors (male sex, systolic blood pressure, antihypertensive treatment, total and high-density lipoprotein cholesterol, smoking, and diabetes mellitus), measured at baseline, were significantly related to the incidence of hard CVD and remained significant when updated regularly on follow-up. The study found that standard risk factors remain strong predictors of hard CVD over extended follow-up and that 30-year risk prediction functions offer additional risk burden information that complements that of 10-year functions.

Considering the extensive length of follow-up and the potential bias due to the competing risk of noncardiovascular mortality in the prediction of long-term risk, the researchers employed a modified Cox model to adjust the risk estimates for the competing risk of non-CVD mortality. The standard Cox model, similar to the standard Kaplan-Meier estimator, may provide biased estimates of absolute long-term risk because it fails to treat those who die of noncardiovascular causes as ineligible for development of CVD events. The competing risk model corrects this shortcoming by calculating the cumulative incidence of CVD in the following manner:

$$\hat{I}_{CVD}(30) = \sum_{t_i<30} \hat{\lambda}_{CVD}(t_i)\hat{S}(t_{i-1}).$$

The quantities under summation denote the instantaneous hazard of CVD at event time ti and survival rate from both CVD and noncardiovascular death past event time ti1. Five-fold cross-validation was used to account for the fact that we evaluated the model on the same data on which it was developed; in this way, the researchers were able to utilize all data available while correcting for potential overoptimism in the assessment of model performance. Additionally, they performed internal validation by randomly splitting the sample 2:1 and developing the function on the first two thirds and evaluating its performance on the remaining third.
Our main takeaway from this case study were the methods of analysing data and statistics, as well as evaluating the effectiveness of the prediction algorithm. We learnt about the Cox model, which is a regression model used to explore the relationship between the survival time of the patients and several explanatory variables. The C statistic gives the probability of a randomly selected patient who experienced an event (disease or condition) having a higher risk score than a patient that did not. We also learnt about cross validation, also known as out-of-sample testing statistical method used to estimate the skill of machine learning models.

2. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015

This case study integrated data on disease incidence, prevalence, and mortality to produce consistent, up-to-date estimates for cardiovascular burden.CVD mortality was estimated from vital registration and verbal autopsy data. CVD prevalence was estimated using modeling software and data from health surveys, prospective cohorts, health system administrative data, and registries. Years lived with disability (YLD) were estimated by multiplying prevalence by disability weights. Years of life lost (YLL) were estimated by multiplying age-specific CVD deaths by a reference life expectancy. A sociodemographic index (SDI) was created for each location based on income per capita, educational attainment, and fertility.

In 2015, 19.9 million CVD deaths occurred (one-third of all global deaths), and 423 million people had prevalent CVD (1 in 17 of the global population). In contrast to the conventional wisdom that CVD remains mainly a condition of wealthy nations, it was found that, adjusted for age, far more cases of CVD are now occurring in countries with the lowest sociodemographic levels than with highest levels; with most CVD at middle sociodemographic levels in men and at middle and low sociodemographic levels in women. Estimated age-standardized CVD prevalence was highest in certain African and Middle Eastern nations; and lowest in several high-income Asian, South American, and Western nations. Among CVD subtypes, estimated ischemic heart disease mortality was highest in Central Asia and Eastern Europe, and lowest in high-income Asia Pacific nations (e.g., Japan). Estimated stroke mortality was highest in

Oceania and central Sub-Saharan Africa. Estimated age-standardized CVD mortality remained relatively stable in Sub-Saharan Africa and Southeast Asia, and increased in Bangladesh and the Philippines. In contrast, significant declines occurred in all high- and some middle income countries. In high-income Western nations, this decline appeared to plateau in more recent years—perhaps a harbinger of the advancing harms of the obesity and diabetes epidemics in these nations.

This case study tells us that CVD rates are increasing globally in low-income nations, as well as nations that suffer from obesity and overconsumption epidemics. The study shows the demographic risk factors leading to CVD, such as wealth and region. The data provided is comprehensive and reliable, tracking the burden of CVD globally over the course of 25 years. The information in this study is highly useful as it provides the prevalence of CVDs by country, helping our group to determine which countries to focus on.

3. Prevention of Cardiovascular disease

Cardiovascular disease has levelled of and a decline has started. This decline starts earlier in some European countries earlier than others. However the rate of CVD decline had decreased in the US. This might mean that the decrease is merely temporary and a postponement of events rather than 100% prevention, furthermore, because of rising healthcare costs we must be more invested in caring and preventing CVD rather than curing it. There are many factors which affect the chance of the person getting CVD, with smoking or tobacco being the biggest factor. Most of the time, encouragement, motivation and advice would be ineffective at discouraging smoking thus drug therapies such as including nicotine replacement therapy (NRT), bupropion or varenicline should be considered early on. Smoking cessation pharmacotherapy may double or triple quit rates, and combined with counselling improves quit rates further.

At the personal level, strategies that help to improve patient self control and to induce sustainable behaviour change. Many apps and devices are available that provide data that can be useful for lifestyle changes.

The following is the list of diet changes that affect CVD risk:
Consume an increased amount of fruit, seeds, nuts, vegetables; 2 to 3 servings of each daily.
Limit consumption of saturated fatty acids to <10% of energy intake daily by replacing with poly-unsaturated fatty acids (PUFA).
Vegetable oils rich in PUFA and soft spreads based on soybean oil, canola oil work as good replacements, or oils for cooking. Limit the consumption of refined grains/white rice and sugar;

aim at 30-45 gr of fibre daily, preferably from wholegrain products such as Long Grain Brown Rice,Black Rice and Purple rice.
At the individual level, physical activity should be advised; it should become regular life from childhood onwards. Children and adolescents should spend 30 to 45 minutes daily in exercise activities. This should be maintained for as long as possible.
Prevention of CVD has been a success story; however, challenges remain related to residual CVD risk, environmental factors, the ageing of the population and poor control adhering to recommendations regarding CVD prevention. Some of these factors are related to human behaviour and self discipline and to socio-economic features.

Methodology

We decided to use open data websites such as Kaggle, data.world and UCL Machine Learning Repository to collect patient data. For further information, we plan to use data from government and international organisations such as https://data.gov.sg/ , www.who.int/gho/en/ , www.moh.gov.sg/resources-statistics and www.cdc.gov/heartdisease.

After collecting data, we cleaned and preparated the data for visualisation and the prediction algorithm using R. We extracted specific columns and rows of data and also merged different datasets in order to obtain meaningful data.

Tableau was used to visualise and present the processed data in an easy to understand manner. The data was visualised in the form of heatmaps, pie charts and bar graphs to understand CVD trends. All of the patient datasets were visualised and compared to All the completed data visualisations will be uploaded onto a website for public viewing.

For the prediction model, we combined 2 of the patient datasets with similar variables in order to obtain a wider spread of data. The model was trained to make predictions on 70% of the data. The remaining 30% was used to test the accuracy of the model.

In order to obtain a prediction algorithm with high accuracy, we experimented with different prediction models until a satisfactory result was obtained.

The first model used was a Decision Tree, which uses a tree-like model of decisions and their possible consequences to classify data. At first, a very complex decision tree was obtained with an accuracy of 74.2%. As such, pruning was used to reduce complexity and the chance of overfitting, which in turn increases the accuracy of the model. Pruning was done by plotting out the cross-validated error against different values of the complexity parameter of the decision tree in order to find the optimal complexity parameter. This value was then applied to the decision tree, which produced a much simpler decision tree with increased accuracy of 72.93%

The next model used was a Random Forest, which outputs a prediction based on the collective decision of multiple decision trees. We found that the optimal number of trees was 100. This model gave a slightly higher accuracy of 74.67%

The final model we tested was k-Nearest Neighbours, which makes a prediction on a datapoint based on the data points closest to it. Proximity of the datapoints is calculated using euclidean distance. The optimal value of k (number of neighbouring datapoints to be considered) was found to be 9 by plotting out the accuracy of the model against different k-values. This model yielded a very high accuracy of 99.56%, thus we decided to use the k-Nearest Neighbours model for the algorithm

Job Distribution

Gregory Tan	Coding of prediction algorithm, data preparation
Ooi Xuan Shan	Coding of website, data collection
Neo Souw Chuan	Data visualisation, data collection

Timeline

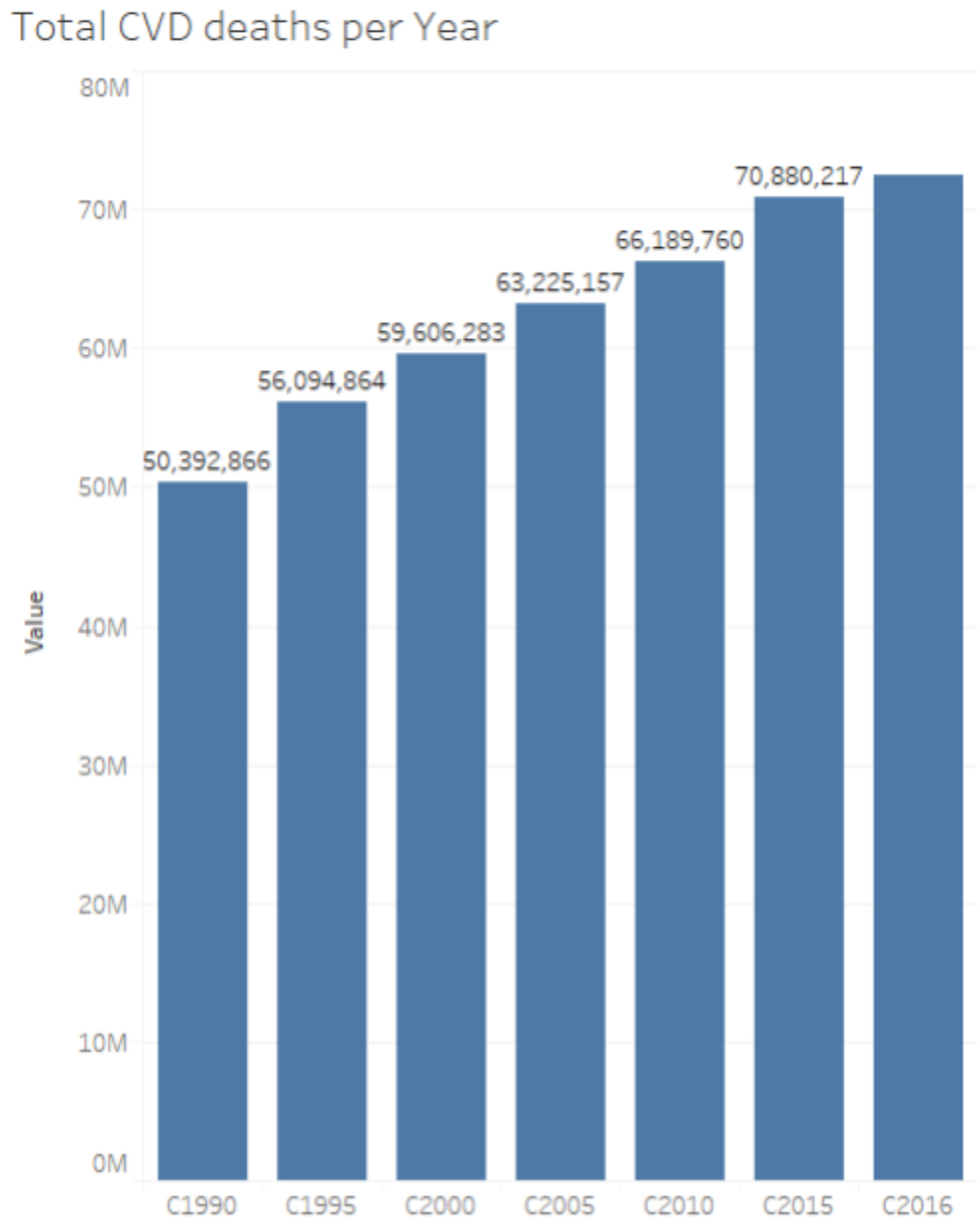
Timeframe	Objective/Event
Term 1 Week 2	● Introduction to DS ● Group Allocation ● Started Research
Term 1 Week 4	● Preparation of proposal slides ● Writing of proposal
Term 1 Week 5-9	● Collection of data ● Editing of proposal
Term 1 Week 10	● Data Science Sabbatical Training ● Acquired skills to clean and present data ● Learnt how to use machine learning prediction algorithms
March Break	● Finalize Proposal Idea
Term 2	● Preparation and cleaning of data

June Holidays Week 1-2	● Start of prediction algorithm and data visualisation
June Holiday Week 3-4	● Creation of Website to present data
Term 3 Week 1-3	● Finishing of algorithm and visualisation
Term 3 Week 2-5	● Complete written report
Term 3 Week 7	Final Evaluation

Outcomes, Analysis and Discussion

Total Deaths

From the same dataset, the ever increasing threat of CVD can also be seen. The graph below shows the number of deaths per year due to CVD. It is shown that more and more people die each year due to CVD, meaning that the threat and importance to predict and prevent CVD is always increasing.
Figure on the right.



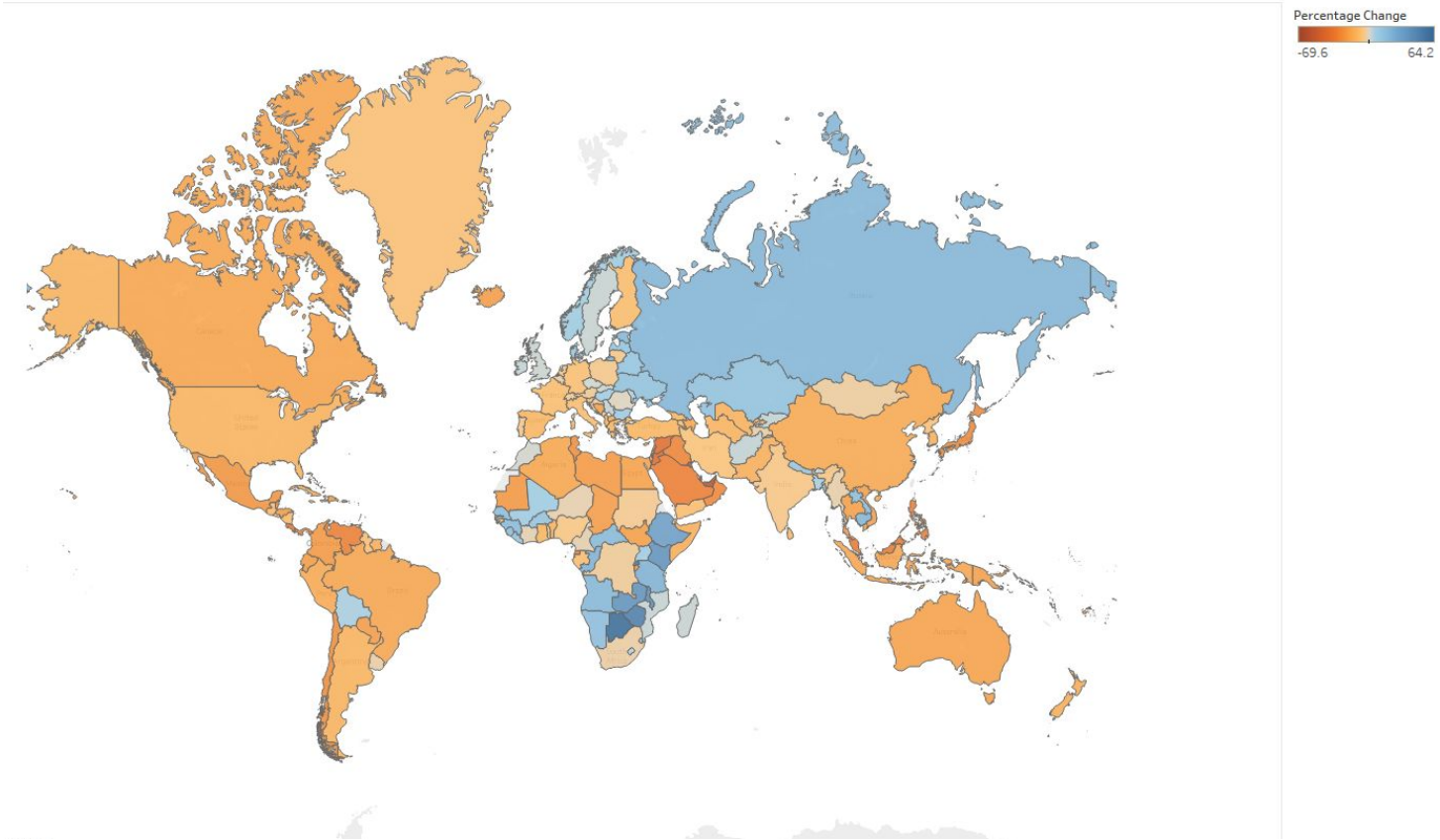
C1990, C1995, C2000, C2005, C2010, C2015 and C2016.

Deaths per Year due to CVD, from 2000 to 2016 with Interval of 5 years

Location

The cardiovascular-disease-deaths-by-age dataset which contains number of CVD deaths per year from 2000 to 2016, along with country and year shows the percentage change of deaths by CVD in most countries. The graph shows the increase in percentage of CVD deaths through a stronger blue colour while a decrease in percentage of CVD deaths in red. It shows that most countries have decreased in percentage of CVD deaths but in certain places, mostly centered around the African continent and near Russia have instead increased. This might mean that more developed countries have better healthcare making tm more able to prevent and deal with CVD, decreasing CVD death. Figure below

A more reddish colour means increase in percentage of deaths by CVD in that country, a more blueish colour means decrease.



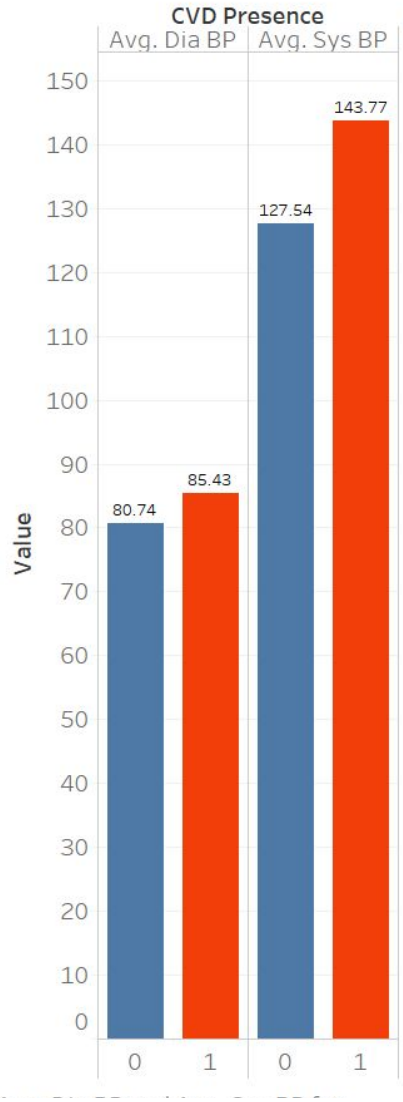
Location

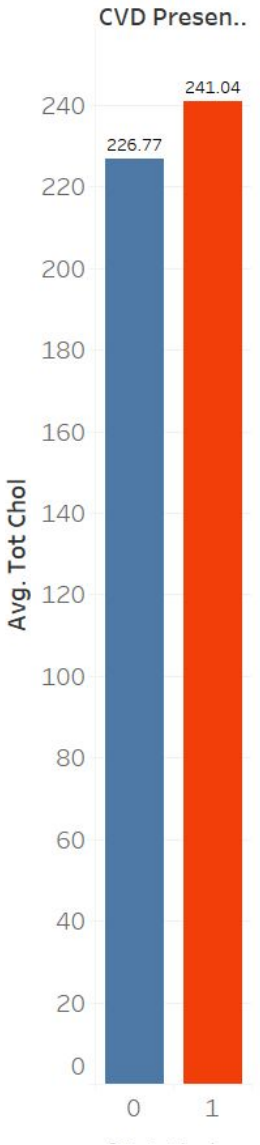
The same dataset also shows us the percentage of deaths per country due to CVD, with a darker red colour meaning a higher percentage of deaths due to CVD. It can be seen that the highest percentage of CVD deaths is in Eastern Europe, in countries like Russia and Uzbekistan. Although we cannot conclude precisely why these countries have such a high prevalence, it might be due to the food they consume. Figure below

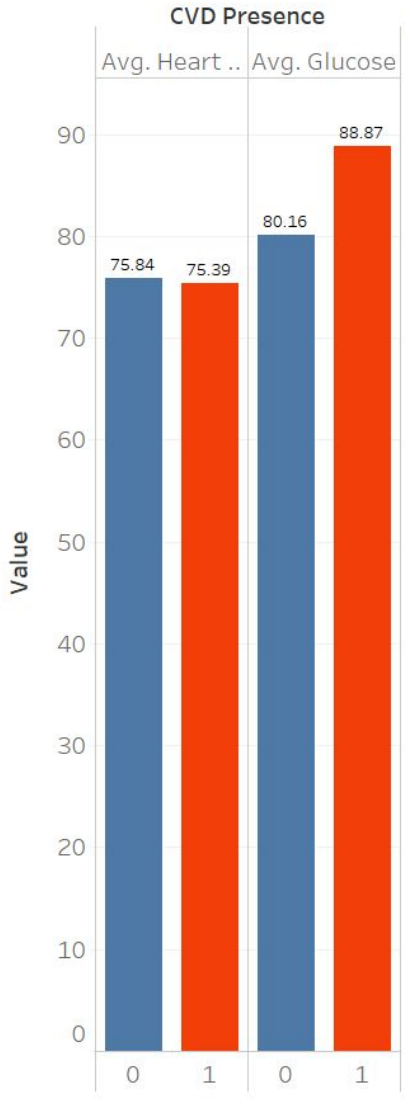
A more reddish colour means greater percentage of deaths by CVD in that country.

Heart Rate, Blood Pressure, Glucose Level and Cholesterol

From the data collected based on the Framingham heart study, data analysis shows that there seems to be a correlation with CVD presence and increase in Diastolic and Systolic Heart Pressure. Figure on the left, red meaning presence of CVD
It can also be seen that CVD presence also leads to increase in Total Cholesterol and Glucose level in blood, however there does not seem to be a correlation between CVD presence and heart rate. Figures Below. *From left to right, Graph of Cholesterol, Graph of Heart rate and Glucose,*
Graph of Diastolic and Systolic Blood pressure. Red means presence of CVD



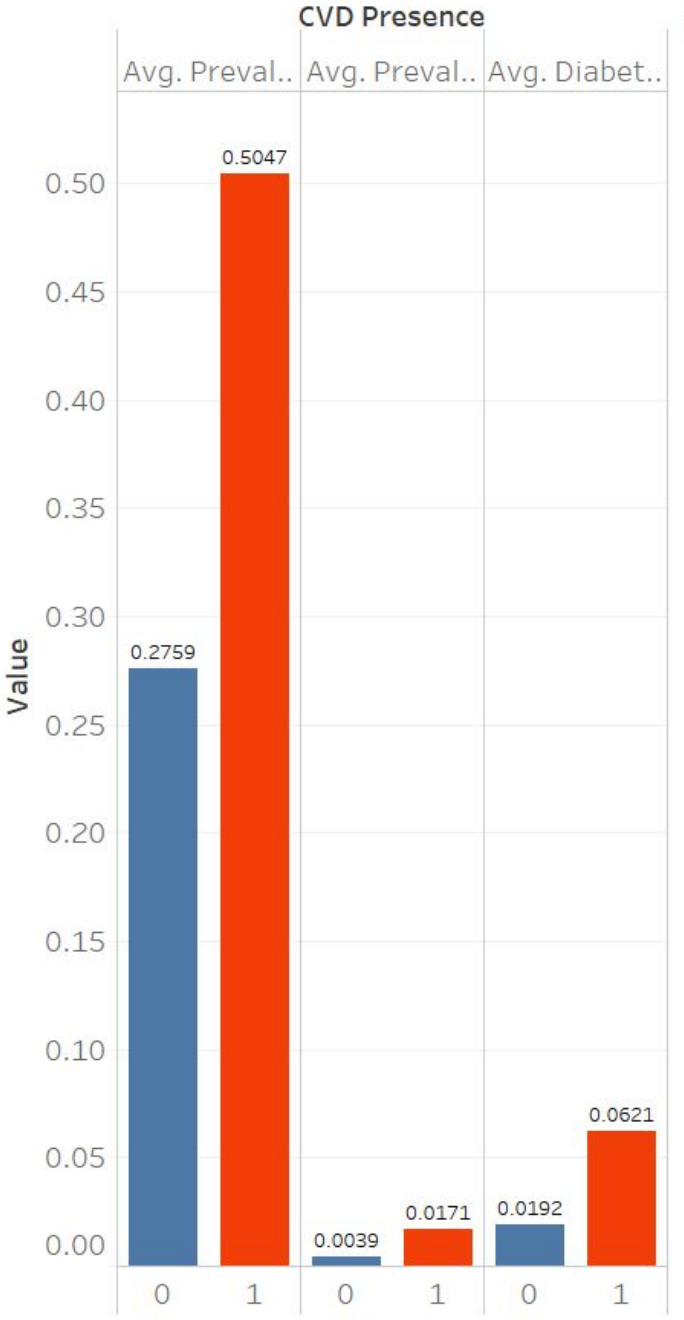




Stroke, Hypertension and Diabetes

The Framingham heart study also shows that there is probably a strong correlation between presence of CVD and Prevalence of Stroke, Hypertension and presence of Diabetes.

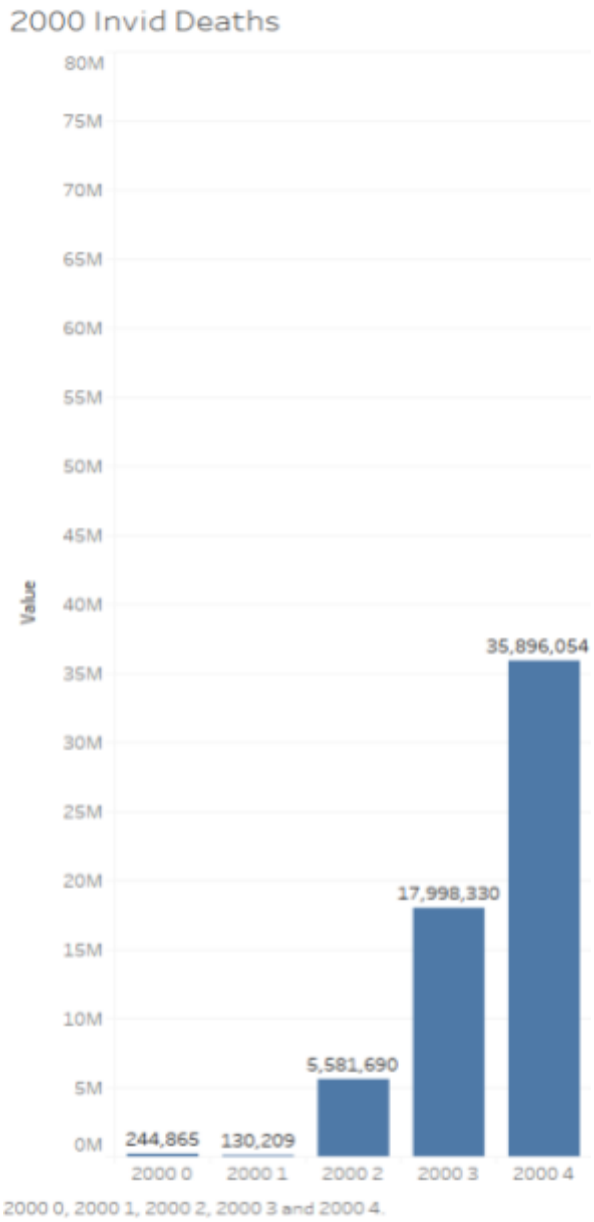
Figure on the right, red indicating presence of CVD. *From left to right, the columns are Prevalence of Hypertension, Prevalence of Stroke, Presence of diabetes, respectively/*



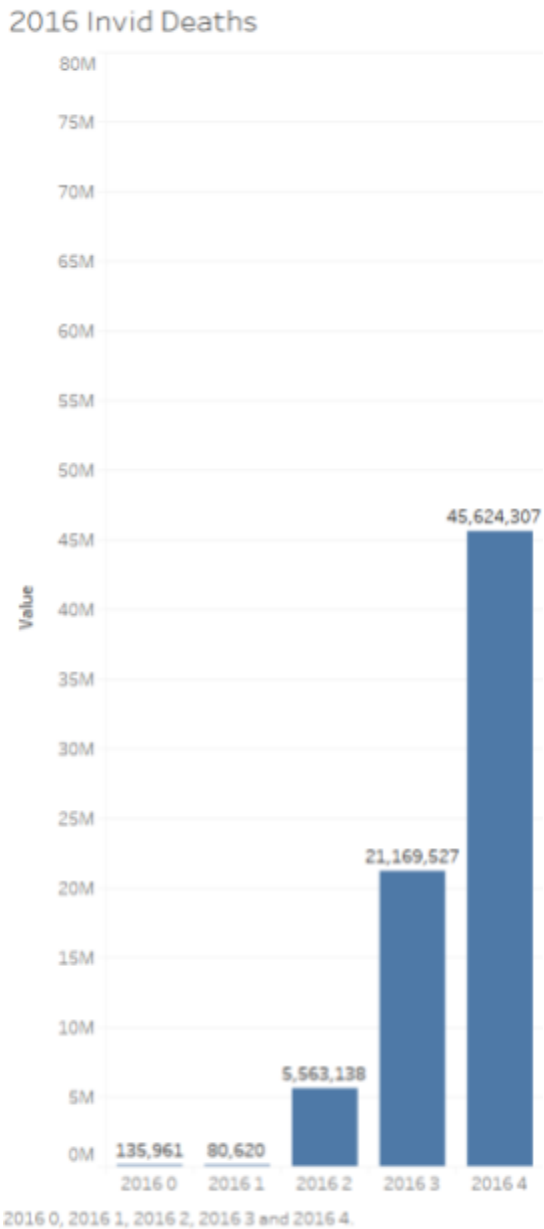
Age

The same data from the maps shows that there is a higher number of deaths to CVD as the age ranges closer to 70 and above. This seems to show that age is an important factor of presence of Death.

Figure below, red indicating presence of CVD. *From left to right, the graphs are number of deaths by CVD in 2000, and deaths by CVD in 2016 respectively.*



From left to right, deaths due to CVD from age 0-5, 5-14, 15-49, 50-69, 70+ in year 2000`



From left to right, deaths due to CVD from age 0-5, 5-14, 15-49, 50-69, 70+ in year 2016

Implications and Recommendations

Our project was limited in terms of the amount of data we collected as most of the patient data came from the same countries (UK and USA), which meant it was not representative of the global trends. The prediction algorithm only included 7 different variables due to the limited data that we had. The types of visualisation we used for the data was also restricted, as we mainly used bar graphs to compare different features between CVD and non-CVD patients

Possible future improvements for the project would be to get more data and to include more features in the prediction algorithm to do a more in depth study on CVD trends. The data could also include follow-ups on the patients to see if they contracted CVD in the years after the data was recorded, rather than simply indicating whether the patient had CVD at the time of the medical examination. The data visualisation could also be expanded upon, by including charts with multiple variables to understand correlations between CVD factors.

Conclusion

From the graphs above it is evident that there are many factors affecting the presence of CVD. Factors include:

- Poverty level: According to a graph shown above, there is a huge increase in the percentage of deaths due to CVD in the developing countries(Whole Africa Continent and India), while the number of deaths decreased in the developed countries(e.g. United States of America and Australia). This might possibly be due to more developed healthcare infrastructure and better public education of affluent countries meaning citizens are more aware of their health and there is faster response when there is a case of CVD
- Age: As age increases, chances of having CVD and deaths due to CVD increases
- Gender: Males have a higher chance of having by CVD as compared to females.
- Education levels: More educated people(those who studied till university) have a lower chance of having CVD as compared to less-educated people(those who dropped out at education in primary schools or high schools)
- Smoking levels: Those who smoke more have a higher chance of having CVD as compared to those who smoke less or do not smoke at all.
- Other diseases: Those who are infected with Hypertension, Stroke or Diabetes have a higher chance of having by CVD.
- Other factors include: People with higher blood pressure(Diastolic and Systolic), higher cholesterol, and blood sugar have a higher risk of CVD. There is no trend in heart rate.
- Exercising: Those who exercise more have a lower chance of being infected by CVD.

How can Singaporeans lower the risk of CVD?

- Monitor blood pressure
    - It is important for those above the age of 65 to frequently monitor their blood pressure as they are at highest risk of CVD
    - If the blood pressure is too high, losing weight, limiting alcohol and caffeine, and reducing stress helps to keep it low. Low blood pressure = lower risk of getting CVD
- Lowering cholesterol and blood sugar
    - Cholesterol(especially LDL) can be lowered by taking more olive oil, nuts, beans, and high fiber fruit
    - Blood sugar can be reduced by drinking more water, reducing stress and eating less carbohydrates
- Exercise more
    - Exercising can help to lower blood sugar and blood pressure, and keep your body fit and healthy and thus lowering the chances of CVD
- Do not smoke
    - Smoking increases your blood pressure and heart rate, and is also one of the causes for many diseases such as stroke

Bibliography

- Predicting the 30-Year Risk of Cardiovascular Disease. (n.d.). Retrieved January 25, 2019,from https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.108.816694
- Halm, V. P., Nurmohamed, M. T., Twisk, J. W., Dijkmans, B. A., & Voskuyl, A. E. (2006). Arthritis Research & Therapy, 8(5). doi:10.1186/ar2045 URL:https://arthritis-research.biomedcentral.com/articles/10.1186/ar2045
- Orri, J. C., Thompson, C. J., & Sellmeyer, D. E. (2009, April 01). Case Study: Aerobic Exercise Training Improves Cardiovascular Disease Risk in a 71-Year-Old Woman With Type 2 Diabetes. Retrieved January 19, 2019, from http://clinical.diabetesjournals.org/content/27/2/88
- Top strategies to prevent heart disease. (2019, January 09). Retrieved January 19, 2019, from https://www.mayoclinic.org/diseases-conditions/heart-disease/in-depth/heart-disease-prevention/art-20046502
- P. (n.d.). Prevention of cardiovascular disease: Recent achievements and remaining challenges. Retrieved January 24, 2019, from https://www.escardio.org/Journals/E-Journal-of-Cardiology-Practice/Volume-15/prevention-of-cardiovascular-disease-recent-achievements-and-remaining-challeng

- Roth, G. A., Johnson, C., Abajobir, A., Abd-Allah, F., Abera, S. F., Abyu, G., . . . Murray, C. (2017). Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *Journal of the American College of Cardiology,70*(1), 1-25. doi:10.1016/j.jacc.2017.04.052