

# **Improving the Logistic Regression Model built from the Framingham Heart Study Dataset**

## **Problem Statement**

Cardiovascular diseases (CVDs) are the leading cause of death globally, and has been for the last 15 years (WHO, 2016). In the US, 1 in every 3 deaths is attributed to CVDs (American Heart Association, 2015), and is estimated to be costing \$1 billion/day in healthcare cost and lost productivity (CDC Foundation, 2015). Extensive clinical research and studies like the Framingham Heart Study have led to the identification of high blood pressure and high cholesterol as two important causal risks factors (Mahmood, Levy, Vasan, & Wang, 2014). Coronary heart disease (CHD) is the most dangerous and the most prevalent CVD. However, with proper screening and education CHD is preventable (Center for Disease Control and Prevention, 2013).

Despite decades of clinical and epidemiological research, estimating the risk of CHD remains challenging, with classification errors as high as 37% in some cases (Kones, 2011). Recently, Diabetes has been implicated as an additional risk factor, based on evidence of its comorbidity, but additional analysis could lead to further refinement of the screening programs (Fox, 2010). Clinical evidence confirms significant increases in blood glucose fluctuation in diabetic patients with CHD (compared to those without) and the important cardiovascular benefits of glycemic control agents (Cheng, Badreldin, Patel, & Bhatt, 2017; Xu & Rajaratnam, 2017). This implicates a role for blood glucose levels, as opposed to simply the presence or absence of diabetes, in the prediction model (Davidson, & Parkin, 2009; Zhang et al., 2014).

Such a model would be more accurate and improve efforts to screen for and educate about all CVDs, and particularly CHD.

### **Approach for the Study**

This quantitative research will involve a logistic regression analysis of the Framingham Heart Study data set (that includes a series of three time measures) on the 10-year eventual outcome Coronary Heart Disease.

#### **A Short History of the Framingham Heart Study**

The leading cause of death is, and has been for the last 20 years, cardiovascular disease. Unfortunately, heart disease and stroke has been a public health epidemic for many decades. In an effort to provide a solution, the National Heart, Lung and Blood Institute (NHLBI) in collaboration with Boston University began a research project almost 70 years ago (1948) in the town of Framingham, MA. This project is known as the Framingham Heart Study and it is credited with identifying the major risks factors that contribute to CVD.

Initially, over 5,000 participants were recruited for the Framingham heart study. These men and women agreed to have certain aspects of their behavior and their results from physical and clinical test monitored over an extended period. The data collected was analyzed and revealed commonalities relating to the development of CVD. These major risk factors identified by the researchers, include high blood pressure, high blood cholesterol, smoking, age, gender and more recently diabetes. This discovery is resulted to the advancement of medical practices and to effective treatment methods for the CVD and its related diseases.

The Framingham Heart Study has since expanded to include more people and other towns, and researchers continue to collect very important data. The advancement of new medical technologies and new collaborations have improved the protocols and expansions to included

ultrasound, echocardiograph and even genetic data. Data collected are also used for projects about other diseases like dementia, diabetes, osteoporosis. There is now a Framingham Heart Study risk calculator, that allows the calculation of the probability of developing Chronic Heart disease over the next ten years of your life.

### Framingham Risk Score<sup>1</sup>

Risk assessment tool for estimating a patient's 10-year risk of developing cardiovascular disease

Age:	<input type="text"/>	Years
Gender:	<input type="radio"/> Female <input type="radio"/> Male	
Total cholesterol:	<input type="text"/>	mmol/L
HDL cholesterol:	<input type="text"/>	mmol/L
Smoker:	<input type="radio"/> Yes <input type="radio"/> No	
Diabetes:	<input type="radio"/> Yes <input type="radio"/> No	
Systolic blood pressure:	<input type="text"/>	mm Hg
Is the patient being treated for high blood pressure?	<input type="radio"/> Yes <input type="radio"/> No	

This online assessment tool is intended as a clinical practice aid for use by experienced healthcare professionals. Results obtained from this tool should not be used for medical decisions.

**Calculate risk**

## Framingham Heart Study

A Project of the National Heart, Lung, and Blood Institute and Boston University

About Participants Our Investigators Risk Functions Bibliography

### Heart Age\* Predictor Using BMI

**Cardiovascular Disease**

- Cardiovascular Disease (10-year risk)
- Cardiovascular Disease (30-year risk)

**Atrial Fibrillation**

**Congestive Heart Failure**

**Coronary Heart Disease**

**Diabetes**

**Hypertension**

**Intermittent Claudication**

**Stroke**

Sex:  
☐ M ☒ F

Age (years):

Systolic Blood Pressure (mmHg):

Treatment for Hypertension:  
☐ Yes ☒ No

Current smoker:  
☐ Yes ☒ No

Diabetes:  
☐ Yes ☒ No

Body Mass Index:

**Calculate**

## Framingham Coronary Heart Disease Risk Score

The Framingham Risk Score estimates risk of heart attack in 10 years.

**BEFORE USE**

There are several distinct Framingham Risk Calculators. MDCalc uses the 'Hard' Coronary Framingham outcomes model, which does not take other co-morbidities into account. For other Framingham Risk Calculations, go to the 'About' section for a link to the Official Framingham Website.

Pearls/Pitfalls

Age:  years

Sex: ☐ Female ☐ Male

Smoker: ☐ No ☐ Yes

Total Cholesterol:  mg/dL Norm: 150 - 200

HDL Cholesterol:  mg/dL Norm: 40 - 80

Systolic BP:  mm Hg Norm: 100 - 120

## Framingham Risk Score Calculator for Coronary Heart Disease

This Framingham risk score calculator estimates the 10-year coronary heart disease risk based on certain criteria like gender, age, cholesterol and systolic pressure. You can discover more about the scoring system and about all the cardiovascular risk factors involved below the form.

Gender:\*

Age:\*

Total cholesterol (mg/dL):\*

HDL cholesterol (mg/dL):\*

Under hypertension treatment?

Systolic blood pressure (mmHg):\*

Smoker?

**CALCULATE** **RESET**

The Ten Year Chronic Heart Disease (TenYearCHD) Risk Factor calculator is available as and software application online. It requires the input of four factors: “age”, “gender”, “Cholesterol” and “Blood Pressure”. Recently, some forms of the application now include a diagnosis of “diabetes” response. The calculation of the probability of CHD is based on a model built from the Framingham Heart Study dataset. It is my contention that the measure of diabetes is not as good a risk factor as the actual measure of the blood glucose level. I have used a subset of the FHS dataset to build two models, one with and without “glucose” and evaluated the comparable quality of the two. I contend that including “glucose” specifically blood glucose levels, improve the accuracy of the model.

### **Future Endeavors**

The results of this research indicate that there is a lot more that can be done to improve the accuracy of the model. In particular, teasing out the correlation between the different variables and find any potential confounders as well as conducting Bayesian studies to deal with the missing data. In addition, looking at the models separately for each gender. The Framingham Heart Study dataset now includes, longitudinal data (three separate time measures of all the variables for over 1.5 million patients. Modeling this data could provide very important and useful insight for dealing with this deadly epidemic.

### **References**

- American Heart Association. (December, 2015). New statistics show one of every three U.S. deaths caused by cardiovascular disease. Retrieved from <http://newsroom.heart.org/news/new-statistics-show-one-of-every-three-u-s-deaths-caused-by-cardiovascular-disease>
- CDC Foundation. (April 29. 2015). Disease and Stroke Cost America Nearly \$1 Billion a Day in Medical Costs, Lost Productivity. Retrieved from

<https://www.cdcfoundation.org/pr/2015/heart-disease-and-stroke-cost-america-nearly-1-billion-day-medical-costs-lost-productivity>

Center for Disease Control and Prevention. (September, 2013). (Preventable Deaths from Heart Disease & Stroke. Retrieved from <https://www.cdc.gov/vitalsigns/HeartDisease-Stroke/>

Cheng, J. W., Badreldin, H. A., Patel, D. K., & Bhatt, S. H. (2017). Antidiabetic agents and cardiovascular outcomes in patients with heart diseases. *Current Medical Research and Opinion*, 1-26.

Davidson, J. A., & Parkin, C. G. (2009). Is hyperglycemia a causal factor in cardiovascular disease? *Diabetes care*, 32, S331-S333.

Fox, C. S. (2010). Cardiovascular disease risk factors, type 2 diabetes mellitus, and the Framingham Heart Study. *Trends in cardiovascular medicine*, 20, 90-95.

Kones, R. (2011). Primary prevention of coronary heart disease: integration of new data, evolving views, revised goals, and role of rosuvastatin in management. A comprehensive survey. *Drug Des Devel Ther*, 5, 325-80.

Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The lancet*, 383, 999-1008.

World Health Organization (WHO). (September, 2016). Cardiovascular diseases (CVDs)- Fact sheet. Retrieved from <http://www.who.int/mediacentre/factsheets/fs317/en/>

Xu, J., & Rajaratnam, R. (2017). Cardiovascular safety of non-insulin pharmacotherapy for type 2 diabetes. *Cardiovascular Diabetology*, 16, 18.

Zhang, X. G., Zhang, Y. Q., Zhao, D. K., Wu, J. X., Zhao, J., Jiao, X. M., ... & Lv, X. F. (2014). Relationship between blood glucose fluctuation and macrovascular endothelial dysfunction in type 2 diabetic patients with coronary heart disease. *Eur Rev Med Pharmacol Sci*, 18(23), 3593-3600.