* Why do the sponsors want the project in the first place? What do they lack, and what do they need?
* What are they doing to solve the problem now, and why isn’t that good enough?
* What resources will you need: what kind of data and how much staff?
* Will you have domain experts to collaborate with, and what are the computational resources?
* How do the project sponsors plan to deploy your results? What are the constraints that have to be met for successful deployment?
* What data is available to me?
* Will it help me solve the problem?
* Is it enough?
* Is the data quality good enough?

You can think of the null model as being “the obvious guess” that your model must do better than. In situations where there’s a working model or solution already in place that you’re trying to improve, the null model is the existing solution. Since this is the simplest possible model, its error rate is called the base error rate. The limit on prediction accuracy due to unexplainable variance is known as the Bayes rate. You can think of the Bayes rate as describing the best accuracy you can achieve given your data.

<https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0681-4>

[1] https://www.researchgate.net/publication/327722009\_A\_Review\_on\_Heart\_Disease\_Prediction\_using\_Machine\_Learning\_and\_Data\_Analytics\_Approach  
[2] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174944

**Introduction**

Cardiovascular disease or in-short CVD is a type of diseases with the involvement of the heart or blood vessels. CVDs includes a wide variety of types such as myocardial infarction (heart attack), stroke, abnormal heart rhythms, stroke etc. The cause of CVD varies according to the disease, in general the main causes are diabetes, high blood cholesterol, high blood pressure, excessive alcohol consumption, smoking and physical inactivity. It was being said that, 80% of CVD deaths for males and 75% of females are accounted by coronary artery disease and stroke [1]. Cardiovascular diseases are also one of the leading causes of death globally [1].

[1] <https://www.who.int/cardiovascular_diseases/publications/atlas_cvd/en/>

According to the Deputy Health Minister Dr Lee Boon Chye, for 13 years from 2005 to 2017, cardiovascular disease (CVD) remains to be the leading cause of death among Malaysians. The issue is that CVD is expected to increase in Malaysia in the near future due to the increase of Malaysian aged 65 in the population to 14.5% of the total population. With a 54% increase of mortality rate due to heart disease over 10 years amounting to 13,503 deaths compared to 8,776 in 2007, the future prospect is indeed worrying.

[2] <https://www.thestar.com.my/news/nation/2019/01/25/heart-disease-leading-cause-of-death>

Currently, screening is the most popular way in CVD identification. But there is a catch, screenings such as ECGs, myocardial perfusion imaging, cardiac stress testing and echocardiography are not recommended to be done among those with no CVD symptoms or at low risk [3][4]. With assumptions stated, it will be too late to detect CVD. Additional to that, biomarkers can be used to predict the risk of future CVD, but the biomarkers result are controversial [5]. The present cardiovascular disease detection in the medical field is yet mature enough and can be costly and time consuming in undergoing the procedure.

[3] <https://www.bmj.com/content/353/bmj.i2416>

[4] <https://www.ncbi.nlm.nih.gov/pubmed/25775317>

[5] <https://www.semanticscholar.org/paper/Multiple-biomarkers-for-the-prediction-of-first-and-Wang-Gona/baa494e8782f6c1c6180677a8ed47d25ac479ca5>

Hence, there is a need to come up with a system to early detect or identify CVD among people using general medical data so that early treatment and preventive measures can be done in an efficient manner. With the advancement of technology, computational power, storage and memory improved drastically. Additional to that, statistical algorithms for machine learning is developing and substantial amount of medical record data is available. By using computers to undertake machine learning on the data we have, accurate predictions ability on CVD can be made.