Comparison between Radial Basis Function Network, Multi-Layer Perceptron and Cox Proportional Hazard Model in Survival Data

Mohammad Salehi Veisi

Department of Statistics, Faculty of Basic Science, Behbahan Khatam Alanbia University of Technology
Behabahan, Iran

Abstract
An interrelated group of artificial neurons using a mathematical model for information processing based on a connectionist approach to calculation is called an Artificial Neural Network (ANN). Performance of the Radial Basis Function (RBF) was also compared with the most commonly used Multi-Layer Perceptron (MLP) network model and the Cox Proportional Hazard (PH) model. Heart attack database was used for empirical comparisons and the outcomes show that RBF performs better than other models. The cox model is the most applicable method for finding the relationship between explanatory and the stable response variable or any other response variable. One of the limitations from this model is the hypothesis of proper dangers. This means that the amount of danger between two or more than two group of the explanatory variables must be constant over time.

Keywords:
Radial Basis Function Network, Multi-Layer Perceptron, Hazard Model, Survival Data

1. Introduction

The aim of this paper is to review the concept of cox model semi-parametric and cox parametric with well-known distributions; we also considered the Spline method in basic hazard assessment in the proportional hazards (PH) model. Several formulas have been studied for Radial Basis Function (RBF) method and some examples have been provided for better understanding. In the following, we used RBF method for estimating the PH model bases. In this paper, three different methods have been compared with each other: Cox proportional hazards (PH) model, Radial Basis Function (RBF), Multi-Layer Perceptron (MLP) network model unlike cox model that is a semi parametric model, and there is no hypothesis about its basic function model, but if we consider a parametric form like vibel, compretz and exponential and we will have a parametric model unlike cox model that has specified basic danger and more hypothesis (Pérez-Godoy et al., 2014).

The Cox PH model is broadly used for the study of time to event data in the presence of covariates. This Cox model is popular because of its simplicity, and not being based on any assumption about the survival distribution (Bernhart et al., 2015). The theoretical basis for the model has been solidified by linking it to the study of counting processes and martingale theory, given in the books of Fleming and Harrington and Andersen. These growths have led to the introduction of numerous new extensions of the original model. Neural computing is an information processing concept, motivated by biological system, framed of a large number of highly interrelated processing elements or neurons to solve particular problems (Dickman et al., 2015). An ANN is configured for a particular application, such as pattern recognition or classification of data, done a learning process. Learning in biological systems requires adjustments to the synaptic connections that exist among the neurons (Rusthoven et al., 2016). The McCulloh and Pitts' network had a fixed set of weights. In 1949 Hebb modernized the first learning rule, i.e., if two neurons are participating at the same time then the lastingness between them should be increased. In 1950 and 1960's Block, Minsky, Papert, and Rosenblatt are exploited on perceptron. The Neural Network (NN) model could be proved to converge to the correct weights that will resolve the problem. By Herb, the weight adjustment or learning algorithm used in the perceptron was established more potent than the learning rules. Parker and Lacuna discovered a learning algorithm for MLP networks called Back Propagation (BP) that could solve problems for non-linearly separable.

The associations of blood pressure with the different manifestations of incident cardiovascular disease in a contemporary population have not been compared. In this study, we aimed to analyze the

Manuscript received June 5, 2026
Manuscript revised June 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.6.12

associations of blood pressure with 12 different presentations of cardiovascular disease (Rapsomaniki, 2014). The aim of this study is to compare the performance of MLP, RBF and Cox PH model using Heart attack data.

2. Cox Proportional Hazard Model

Cox proportional hazard model (D.R. Cox), includes risk factors, which are as follows:

$$h(t,X) = h_0(t)e^{\sum_{i=1}^{p}\beta_i X_i}$$

Which $X = (X_1, X_2, \dots, X_p)$ are explanatory variables and $h_0(t)$ is basis risk and it depends on "t" not on "X"s. also $e^{\sum_{i=1}^{p}\beta_i X_i}$ is depends on "X"s and is not dependent on "t" and X are independent of time. In this model, a definition for risk at time "t" for each individual is corresponded with a set of explanatory variables which called "x". X is a set (sometimes called vector) of predictor variables which appears in model for determining the risk. The Cox Model declares that risk at time t is the result of product of two quantities. The first value is h_0 (t) which is called basis risk function and the second quantity value is e in total power of $\beta_i x_i$ which is collected on p explanatory variable. An important component of this formula is related to the assumption of proportional hazards which is a basis risk function of t function but does not include x. The formula consists of x but not including t. Here Xs are called is independent of time. Also, it is possible that Xs do not include t. in this case Xs are called independent of the time. If time-depended variables have been concerned, we still use Cox model but this detailed model which is not proper for PH is called developed Cox model. A variable is called independent of time when its value does not change over time. The gender of SMK is an example for those variables. Note that for a person's smoking status may change over time but to do SMK variable analysis we assume that after a time, a measure of it does not change so for each person a specific amount has been used. Also, note that although variables such as age and weight change over time may be assigned to treatments which this kind of variables are time-independent if do not change over time or if the effect of such factors depended on survival risk inherent to the measured value of it. One feature of Cox model is that if all values of X be equal to zero in this case the formula has reduced to basis risk function which in this case exponent of e becomes zero and the result will be one. This feature of Cox model shows why we call h_0 (t) the basis function. As a result, the cox has reduced to the baseline hazard function where there is not any X in the model. So this is possible that h_0 (t) as a pioneer or a version of baseline risk function has been considered before considering any Xs. Another important feature of this model is that h_0 (t) function of baseline risk function is an uncertain function and due to this feature the Cox model is a semi-parametric model.

A parametric model is a totally clear function except for unknown parametric values. For example, viable risk modeling is a parametric modeling and its form has been shown here which its unknown parameters include λ ، p, and β_i . Being semi-parametric is the main reason for Cox model to be common. A correct and reasonable estimate for the regression coefficients, the risk ratios, and adjusted survival curves is another reason for the popularity of this model with uncertain risk baseline which can be created for different data in different positions. Another reason is the fairly close results of Cox model to the correct parametric modeling. For example, if the results of correct parametric are viable then the results of Cox model will be equal to viable model results or if the correct model the correct model be an exponential model, the results of Cox model will be equal to exponential model results. The parametric model is better than other methods if the model is correct. Although there is a lot of method for assessing the goodness of fit in a parametric modeling but it is possible that we would not completely be able to recognize the accuracy of the parametric model. Notice that risk function of $h(t,X)$ and accordingly survival curves s (t , x) can fitted for Cox model and without Identify the baseline function. So for this model and by using minimum assumption we can obtain basic information such as survival analysis, the hazard ratio, and survival curves. This model is preferred to the logistic model when survival time information is available. The Cox modeling uses more survival time information in compare with logistic modeling. The logistic modeling uses the outcome of (0, 1) and ignores survival times.

2.1 Estimating of Maximum Likelihood of Cox Proportional Hazard Model

β_i are the parameters in general formula of Cox model and the obtained outcomes of this parameters are called Maximum Likelihood which is shown by $\widehat{\beta}_i$ and those are the correct maximum likelihood (Fox 1997). With logistic regression the estimates of ML are derived for Cox modeling parameters with maximum of a likelihood function which usually is shown by "L". Likelihood function is a mathematic declaration which shows the possibility

of simultaneous observations based on unknown parameters of the model (β_i). The formula of likelihood function for Cox model is called partial likelihood function in proportion to the likelihood function. The term partial likelihood has been used because we only considered the observations that have occurred and we do not directly considered the possibility of censored observations. So the likelihood function in the Cox model might not address all observations and because of this, called partial likelihood. Partial likelihood can be written as follow:

$$L_j = L_1 \times L_2 \times \ldots \times L_K = \sum_{j=1}^{K} L_j$$

K shows the number of time repetition. So in the danger commands in time L_J is called likelihood of risk at this time that survival is lost at this time. The individuals' hazard ratio is called as risk set at a time $R(t_j)$. Although in partial likelihood we focus on occurred observations, former survival time information also has been used for censored observations. A person who censored after the j-th risk at the time has been used as a part of risk ratio until L_J has been calculated. After the formation of the likelihood function for a given model, the next step is calculating the function, Max. This work has been done by maximizing the L logarithm. The maximizing process is done by driveling from L log to any unknown parameters in the model and then, equations that are shown here are solved. This equation is performed by using repetition and resume. Usually, there is interest to Statistical interpretations with estimation of obtained ML. The estimated hazard ratios HR of power by the coefficient in the range of

$$h(t, X(t)) = h_0(t) exp\left[\sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \delta_j X_j(t)\right]$$

Here X(t) includes $X_1,\ldots, X_{P1}$ independent-time and $X_1,\ldots, X_{P2}$ dependent-time. Just like PH Cox Model, the extended model includes a primary risk function h_0 (t) which is multiplied in the exponential function. Therefore, in the extended model the exponential includes both independent-time predictors which have been shown by X_i and the variable of dependent-time shown as X_i (t). All predictor sets in T time has been shown by X(t). A simple example of extended Cox model is a model with one independent-time variable and one dependent-time variable.

$h(t, X(t)) = h_0(t)\exp\left[\beta E + \delta(E \times t)\right]$

(1 and 0) were calculated $e^{\hat{\beta}}$. Note that the model is not in front of any Interaction effects.

2.2 Adjusted survival curves by using COX PH model

The two desired amount in terms of survival analysis estimates the hazard ratio and estimate survival curves (Collett, 2015). If no model has been fit to survival data, a survival curves using Kaplan-Meier method can be estimated. Like KM curves that are a step function. When Cox model used for survival data analysis adjusted survival curves can be obtained for the explanatory variables as predictors (. These are called adjusted survival curves and are drawn like KM curves that are step functions. The formula for the Cox hazard function can turn to survival formula. This survival function formula is the basis for the determination of adjusted survival curves. This formula shows that survival function in the time of t reached to exp ($\sum_{i=1}^{p} \beta_i X_I$) as a predictor of the baseline survival function namely $S_0(t)$. The estimation of $\hat{S}_0(t)$ and $\hat{\beta}_i$ is achieved by a computer program.

2.3 Extended Cox model for dependent variables – time

For a survival analyzes revealed including two predictor variables (independent-time and dependent-time) the extended Cox model can be set down as follow:

$$X(t) = (X_1 = E, X_1(t) = (E \times t)$$

$$p_1 = 1\,, p_2 = 1$$

The independent-time variable shows how E gets (0,1) and $E \times t$ shows dependent-time variable. Just like PH Cox model the regression coefficients in extended Cox model has been achieved by maximum likelihood method. The estimation of ML has been determined by Max of likelihood function but estimation of extended Cox method is more complicated than PH Cox method because risk categories that are used in the form of the likelihood function became very complex with dependent-time variables.

Computer programs which can fit the syndetic cox model include stata, SAS, and SPSS. The methods of Vald, likelihood ratio test and large sample confidence interval are also can be used. An important hypothesis of syndetic cox model is that the effect of a time-dependent variable $X_j(t)$ on the probability of survival in the time t is dependent on the amount of this variable at the time t not sooner or later time. It should be considered that each value of the variable may change during the time, risk model $X_j(t)$ creates only one coefficient for each time-dependent variable in the model.Therefore, at the time t, there is only one value of the variable $X_j(t)$ that has an effect on the risk that is the same measured amount at the t. However, with the term time-dependent, it can be also defined with the effect lag-time. For showing the meaning of lag-time effect, for example, assume that the circumstance of employment is assessed weekly and shown with EMP(t) that the time-dependent variable is under the analysis. then, syndetic cox model that does not consider lag-time hypotheses, the effect of employment circumstance on the probability of survival at the time t is dependent on the amount of observation of this variable at the same time t for example, one week sooner. Lag-time one week, the variable of the circumstance of employment may be modified. Therefore, the risk model is predicted by the circumstance of employment in the week t-1 at the time t. hence, the variable EMP(t) placed by the variable EMP(t-1) in the model (Ross, 2014).

Generally, the syndetic cox model can frequently write by a lag-time modification for each considered time-dependent variable. If we define L_j as the certain lag-time for the time-dependent variable j, the syndetic lag-time model can be written as what has brought here. Notice that the variable $X_j(t)$ is placed in the variable $X_j(t-L_j)$ in the most simple type of syndetic model.

2.4 Risk set for the syndetic model of cox

The formula of the obtained risk set from the syndetic cox model will be explained (Sprent and Smeeton, 2016).

$$\widehat{HR}(t) = \frac{\hat{h}(t,X^*(t))}{\hat{h}(t,X(t))} = \exp\left(\sum_{i=1}^{p_1} \hat{\beta}_i (X_i^* - X_i)\right) + \sum_{j=1}^{p_2} \delta_j \left(X_j^*(t) - X_j(t)\right) \quad (5)$$

Two predictor collections are as follow.

$$X^*(t) = (X_1^*, \dots, X_{p_1}^*, X_1^*(t), \dots, X_{p_2}^*(t))$$

$$\mathrm{X(t)} = (\mathrm{X}_1, \dots, X_{p_1}, X_1(t), \dots, X_{p_2}(t)) \quad (6)$$

The most important part of this formula is the hypotheses of proportional hazards, they did not occur when the syndetic model of cox is used.

This formula explains the risk ratio at a certain time t and the determination of the characteristics of two categories of predictions at the time t. these two categories are determined wth $X^*(t)$ and $X(t)$. Two collections of predictors $X^*(t)$ and $X(t)$ determine two characteristics at the moment t for the combination of the category of predictions include both variables time-dependent and time-independent. The single components for each category for predictors are shown here.

As a simple example, assume that the model only includes a time-independent prediction, means the circumstance of E, a variable (0 and 1) and a time-dependent prediction, means $E \times t$. Then, we compare the people who are at the risk of experiment E=1, with the people who are not at the risk of experimnet E=0 at the moment t. $X^*(t)$ is a collection of predictions that have two components E=1 and $E \times t$ =t; and $X(t)$ has the two combinations E=0 and E×t=0. If we now evaluate hazard ratio, that compare people who are at the risk of experiment with the who are not at this risk, the formula that have written here will be obtained. In other words, $\hat{HR}$ is equal with e to the power $\hat{\beta}$ plus $\hat{\delta}$ multiplied by t. This formula indicates that the hazard ratio of a function is from the time specially, if $\hat{\delta}$ positive. The ratio of hazard increases with the increase of time. so the ratio of hazard is this example is defienetly not stable. Therefore, hypotheses of PH are not applied for this model. Generally, because this formula of public hazard ratio includes different values of time-dependent variables, this hazard ratio is a function of time. Hence generally, the syndetic cox model does not occur PH hypotheses if δ_i did not equal with zero. Note that the formula of hazard ratio, coeffinceit of δ_i that gains differen values for time-dependent JM variable are not time-dependent themselves. These

coefficents explain the effect of the similar time-dependent variable according to the time that each variable assessed in the study.

2.5 Model optimization

First step: unsupervised

The hidden item parameters like c_j and σ specified with those methods which used the only input vector called Z.

Second step: supervised

The linear optimization of output parameters of β_{ϕ_j} was done. It is possible to compute the amount of hidden function after choosing the amount of c_j and σ parameters , then linear regression model with the regression coefficients (β_{ϕ_j}) will use by new variables.

Many techniques are suggested in order to optimizing the **RBF** function. The whole operation is as the following:

1) All centers of C_j specified to **Z** vectors from the input space of **S** and only the amount of σ computed for the mean of distances between C_j centers. We can use this process when the dimension of $S_j$is not big.
2) Otherwise when the dimension of S is big, the centers of c_j specified to a big vector of z from **S.** the consistent factor is σ for high value parameters because the space between centers depends on sampling operation.

The variables in Z vector can measure with some codes or by different scales. Those variables

Scale between 0 and 1. (0 - 1)

It is possible for both functions to be correlated and PCA is used in order to reject the problem. So, all the values of RBF replace with the values of main ratio ($\emptyset^s$). The approximation of regression parameters in the second layer of $\beta_{\phi_j^S}$ will obtain with iterative reweighted least square. All calculated models begin with some increasing ratio of $\emptyset^S$. Because of the acceptable rule about avoid of measuring in regression models in order to keep the maximum parameters limited in $\frac{1}{10}$ domain, bigger models should not be considered. Choosing of ϕ^S took place in order to accept the model by using the AIC scale.

Akaike Information criterion (AIC) is a criterion that is proposed by akaike in 1974 and its purpose is to measure the goodness of the estimated model. AIC is a criterion that measures the amount of complexity of the model and the proper goodness and the amount is better if it is less.

AIC= -2Log (likelihood) + 2 ρ

the model that has a little akaile criterion, has better goodness relating to the data and it has more application.The measure of approximate parameters indicted ρ .AIC is equal with leave one out cross-validation. Having the reliability for hazard functions is not standard.Two factors should be considered: the process of choosing model and the approximation of an appropriate interval should be computational and those approximations must gain directly from the likelihood results. Making model is dependent on choosing the center of RBF and σ the amount of approximation uncertainty should be considered.

Artificial Neural Network (ANN): ANN at first developed to imitate basic biological neural systems, the human brain particularly, are collected of a numeral of interlinked simple processing elements called nodes or neurons. Every node gets an input sign which is the total information from other nodes, processes it locally though an activation or transfer function and produces a transformed output sign to other nodes. Every single neuron enforces its function rather slowly and badly, jointly a network can execute an amazing number of jobs efficiently Reilly and Cooper. This information processing characteristic create ANN a powerful mathematical device and able to find out from examples. First studies of NN were carried out in 1942 by McCullough and Pitts. After sometime, Rosenblatt conceived in 1959 the first learning algorithm, creating a model known as the perceptron, which was then only a solution to simple linear problems. Werbos reported the first non-linear processing capabilities of ANNs in 1974.

Multilayer preceptor (MLP): In MLP, the weighted sum of the inputs and bias terms are passed to activation level through a transfer function to produce the output, and the units are arranged in a layered feed –forward topology called Feed Forward Neural Network (FFNN). The diagrammatic representation of

FFNN is given in Fig.1. The three layers of ANN are input layer, hidden layer and output layer. The learining power of the MLP increases by the hidden layer. The activation function of the network changes the input to provide a desired output. The activation function is choses by the algorithm require a function with a continuous, single-valued with first derivative existence. Choosing the hidden layers' number, hidden nodes, and the type of activation function play an essential position in model building, Hecht-Nielsen, and White.

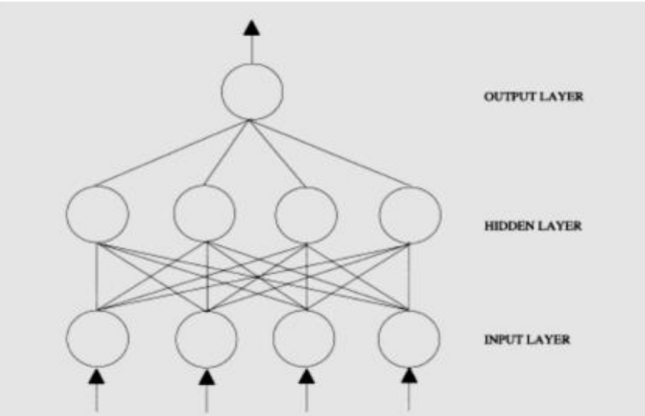

Figure 1. Feed forward neural network

3. Database

The data obtained from the website, ftp//ftp.wiley.com/public/scitech_med/survival and http://www.umass.edu/statdata/statdata. The data from the Worcester Heart Attack Study (WHAS) have been provided by Goldberg (1989) of the Department of Cardiology at the University of Massachusetts Medical School. Data collected from 1975 to 1988, on all myocardial infarction (MI) patients allowed to hospital in the Worcester, Massachusetts Standard Metropolitan Statistical Area. Event is encrypted as 1 and censoring is encrypted as 0. The subsets of covariates used. For MLP network architecture, one hidden layer with activation function of sigmoid, which is optimal for the outcome, is chosen. A BP algorithm grounded on conjugate gradient optimization technique was used to model MLP for the above data. A Cox PH model was fitted using the same input vectors as in the neural networks and heart attack status as the binary dependent variable. Constructed models efficiency was evaluated by likening the sensitivity, specificity and overall accurate predictions for datasets. Cox PH, MLP and RBFNN were constructed using SPSS and MATLAB Soft wares.

Table (1): Statistical equivalent phrases in neural networks

Neural network	**analysis**
network	model
learning	estimation
Regression	regression
Generalization	Interpolation
Learning collections	observation
Synapse	parameters
Inputs	Independent variables

outputs	Dependent variables

4. Discussion

MLP and RBFNN are the comprehensively used FFNNs. Both differ basically in the way how the hidden units combine values approaching from the inputs. The MLP use inner products and the RBF use Euclidian distance. We used both RBF and MLP algorithms for the forecast of heart attack data.

Venkatesan P and Suresh M. L study gives show that the accuracy of ANN for the breast cancer survival forecast was better than regression models, Burke H. B et al. study concluded that neural networks are more accurate in predicting the breast cancer than LR and CART models for fifth year survival. Hacib T in his paper states that RBFNN identifies the parameters of electromagnetic, faster than MLP neural network. Dan Ardelean et al. in their paper reported that quality of the RBF model is better than the quality of the MLP. Padmavathis paper of breast cancer forecast used RBF and MLP, suggested that RBF have good predictive capabilities and time taken was less when compared to MLP. Venkatesan and Anitha study gives that performance of the RBFNN has a better performance than other models like MLP and classical logistic regression. Sereno F et al. paper entitled, comparative study of RBF and MLP neural nets in the Estimation of the Foetal weight and length, concluded with the slight confusion regarding prediction performance while comparing the RBF and MLP networks to resolve the problem of foetal weight prediction. Many researchers have compared the efficiency of RBF and MLP and majority have recommended that RBF network was better than MLP, and some of them doubt the prediction efficiency.

5. Results

WHAS data sets with 481 records were used for the study. The Cox PH models were fitted using SPSS. The covariates AGE, SHO and CHF are significantly connected with the time to event. The Cox PH regression fitted to the data gave a sensitivity of 85%, specificity 82% and overall accurate prediction of 83%. The MLP architecture had six input variable and one hidden layer with three hidden nodes and one output node. The best MLP was obtained at lowest Root Mean Square (RMS) of 0.2125. MLP sensitivity was 92%, specificity was 91% and percentage accurate prediction was 91%. RBFNN executed best at ten centres and maximum number of centers tried was 18. Root Mean Square Error (RMSE) using the best centres was 0.3212. Sensitivity of the RBFNN model was 97%, specificity was 97% and the percentage accurate prediction was 97%. Execution time of RBFNN is lesser than MLP and when compared with Cox PH model.

Table (2): models and their characteristics

Model	Sensitivity (%)	Specificity (%)	Accuracy (%)
Cox	85	82	83
MLP	92	91	91

RBF	98.5	98.5	98.5

6. Conclusion

The sensitivity and specificity of both NN models had a better predictive power compared to Cox PH regression. Also the time taken by RBF is less than that of MLP in our findings. The limitation of the RBFNN is that it is more sensitive to dimensionality and has larger difficulties if the numeral of units is huge. The forecasting capabilities of RBFNN has showed better outcomes and more applications would bring out the efficiency of this model over other models. The radical bases function (RBF) was used to estimate **PH** model, and concluded that this function has a vast role of estimating different parameters

References

[1] Pérez-Godoy, M. D., Rivera, A. J., Carmona, C. J., & del Jesús, M. J. (2014). Training algorithms for radial basis function networks to tackle learning processes with imbalanced data-sets. Applied Soft Computing, 25, 26-39.

[2] Collett, D. (2015). Modeling survival data in medical research. CRC press.

[3] Rusthoven, C. G., Jones, B. L., Flaig, T. W., Crawford, E. D., Koshy, M., Sher, D. J., ... & Pugh, T. J. (2016). Improved survival with prostate radiation in addition to androgen deprivation therapy for men with newly diagnosed metastatic prostate cancer. Journal of Clinical Oncology, 34(24), 2835-2842.

[4] Bernhart, G., Fernández, L., Mai, J. F., Schenk, S., & Scherer, M. (2015). A Survey of Dynamic Representations and Generalizations of the Marshall–Olkin Distribution. In Marshall− Olkin Distributions-Advances in Theory and Applications (pp. 1-13). Springer International Publishing.

[5] Sprent, P., & Smeeton, N. C. (2016). Applied nonparametric statistical methods. CRC Press.

[6] Ross, S. M. (2014). Introduction to probability models. Academic press.

[7] Dickman, P. W., Coviello, E., & Hills, M. (2015). Estimating and modeling relative survival. *Stata J*, *15*(1), 186-215.

[8] Kohl, M., Plischke, M., Leffondré, K., & Heinze, G. (2015). PSHREG: A SAS macro for proportional and nonproportional subdistribution hazards regression. Computer methods and programs in biomedicine, 118(2), 218-233.

[9] Rapsomaniki, E., Timmis, A., George, J., Pujades-Rodriguez, M., Shah, A. D., Denaxas, S., ... & Williams, B. (2014). Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1· 25 million people. The Lancet, 383(9932), 1899-1911.