Part I: Exploratory data analysis

An exploratory analysis was performed by producing a scatter matrix of all covariates excluding “Systolic” (as per instructions) and “Illness” (due to its binary nature, Figure 1). According to the scatter matrix, the relationships between the following variables were investigated:

1. The correlation between heart rate and cholesterol level

Previous literature has emphasized on the relationship between serum cholesterol and elevated cardiovascular markers including heart rate (1) (2) (3). These evidence point to a physiological relationship between hypercholesterolemia and heart rate: fat deposits, especially low density lipoprotein (LDL) cholesterol and triglycerides, are prone to form atherosclerotic plaques. These plaques cause narrowing of arteries and predispose patients to a variety of arterial diseases, which are characterized by abnormal hemodynamic parameters such as increased heart rate (HR) and blood pressure (BP) (3). However, the scatter matrix (Figure 1) and scatterplot with the best line for heart rate vs. cholesterol level (Figure 2) suggests that there is a weak positive correlation between the two variables. This is, however, not surprising, as a multivariate linear regression model conducted by (4) reported a statistically insignificant association between total serum cholesterol and resting heart rates ($p > 0.05$) in a sample of 1054 urban dwellers (male and female), black South African between the age of 25-75. Furthermore, the data measures total serum cholesterol, which includes both the “bad” LDL and “good” HDL cholesterol (high density lipoprotein, which collects serum cholesterol from tissues and arterial wall for hepatic excretion, effectively lowering total serum cholesterol and improving overall cardiovascular outcomes, such as lowering BP and HR (5). Therefore, even though there is a distinction between the sub-component LDL and HDL cholesterol regarding their effect on cardiovascular outcomes, there might not be a strong positive relationship between the total serum cholesterol and heart rate) as Figure 2 suggested.

2. The correlation between hypercholesterolemia and predictions of BMI and obesity status

Hypercholesterolemia, which is sometimes classified under the umbrella term “dyslipidemia", has been implicated as one of the main predictors of BMI and obesity status (8) (9). Physiologically, a high serum cholesterol could mean that there is a significant proportion of both circulating fats and fat deposits in peripheral tissues, which results in a high body weight and hence higher BMI. Nevertheless, Figure 3 shows a scatter plot for the level of cholesterol and BMI, which demonstrates a weak positive correlation between the 2 variables. The weak association between these 2 variables can be explained by the following reasons: 1) BMI itself is not a primary variable, but relies on both height and weight; therefore, it is possible that individuals with a higher bone density, higher muscle mass, or is significantly taller for their weight will have a high BM, but a low cholesterol level (10) Likewise, individuals who are considered “skinny fat” might have a normal weight, but an alarming high serum cholesterol level (11).

3. The correlation between high cholesterol (hypercholesterolemia) and progression to heart-related diseases

Hypercholesterolemia has also been implicated as a predictor of cardiovascular mortality (3) (6). Here, a box plot was produced to compare the cholesterol level between individuals who do and do not develop heart-related conditions (Figure 4). According to Figure 4, two conclusions can be deduced: 1) the data is independent and normally distributed with equal portions around the median, and 2) there is no significant difference between the median, upper quartile and lower quartile cholesterol level between the two groups, even after outliers have been removed. Given that the data set has satisfied the independence, normality, and continuous data assumption, a two-sample Student’s t test was performed to investigate whether there is a difference in the mean cholesterol level between the two groups. According to the STATA output (Figure 5), there is statistically significant difference between the mean cholesterol level of Illness = 0 and Illness = 1 group (p-value = 0.0087), suggesting that there is a difference in the cholesterol level between those who do and do not suffer from heart-related conditions.

Part II: Simple linear regression

Discussion of the most appropriate covariate to produce a simple linear regression model with Systolic

In order to create the “best” simple linear regression between a single covariate and the variable Systolic, a second scatter matrix was produced to visualize any discernible linear pattern between Systolic and other variables (Figure 6), in addition to a correlation coefficient table which details the r-values of *all* variables against Systolic (Figure 7). According to Figure 9, the 3 variables with the highest r-values (in decreasing order) are Age, BMI and Cholesterol. Nevertheless, a full regression model of Systolic and other variables is required.

In order to determine the covariate which would produce the best simple linear regression model with Systolic, the following criteria needs to be satisfied:

1) Highest R-squared value + lowest root MSE: the R-squared value represents the proportion of the variation in Systolic explained by the covariates, whereas the root MSE represents the deviation or variance between the

recorded value and the theoretical value as predicted by the linear regression model. Typically, the higher the R-squared value, the higher the percentage of variations in Systolic is explained by a covariate, and the lower the root MSE value, the more accurately the model can predict the target value for Systolic;

2) p-value < 0.05: this means there is a statistically significant relationship between Systolic and the covariate;

Table 1 summarizes the adjusted R-squared, root MSE and p-value (Prob > F) of all potential covariate. According to these results, it is reasonable to deduce that **the 3 aforementioned variables (Age, BMI and Cholesterol)** are indeed potential covariates to produce the best simple linear regression model for the variable Systolic.

Next, a scatterplot of each covariate against Systolic is created to confirm criterion #3 (Figure 8-10). According to these figures, it is reasonable to deduce that all 3 covariates show a positive linear relationship with Systolic.

In addition to the 2 aforementioned criteria, the following 3 criteria need to be satisfied:

3) The relationship between y and x is linear: this assumption is validated by producing a scatterplot between each covariate against Systolic. Unless the graph shows a clear non-linear relationship (i.e quadratic, logarithmic), the assumption is fulfilled;
4) The error term must be normally distributed: $\epsilon \sim N(0, \sigma 2)$: this assumption is validated by graphing the quantile-quantile, or QQ plot, which plots where the residuals fall on a standard normal distribution against the sample quantiles of the normal distribution. This assumption is fulfilled if all the points lie perfectly on the line;
5) The error term variance is constant & independent: this assumption is validated by graphing the residuals-versus-fitted plot (or RVF plot for short), which plots the x value against their deviation from the line of best fit (the horizontal x=0 line). This assumption for a linear regression model is fulfilled if 1) all data points deviate randomly from the x=0 line with no discernible pattern, 2) the residuals cluster around the y=0 line suggesting that the variances of the error terms are constant;

Figure 11-16 summarizes the QQ and RVF for the covariate Age, BMI and Cholesterol against the outcome Systolic. Starting with the covariate Age, it is clear that the assumption for linear regression model is not violated: the residuals cluster around the y=0 line with no clear pattern, suggesting that the error terms are independent, and the spread/ deviation of the residual points is constant along the horizontal y=0 line (i.e there is no specific clusters in the graph). Moreover, the QQ plot for Systolic vs. Age shows that the majority of the data falls on the horizontal line within/ between the 95th percentile, with a slight deviation from the tail which might mean that the variable Age could be right-skewed.

Similarly, the QQ plot for Systolic vs. BMI and Systolic vs. Cholesterol demonstrates the same pattern. Previous research has concluded that BMI is not always normally distributed but positive-skewed in the higher weight range (i.e overweight and obese individuals (7); moreover, it is clear that the tail deviation for both QQ plots (Systolic vs. BMI and Systolic vs Cholesterol) is largely affected by outliers, which are identifiable (< 10 outliers) and can be easily removed by their IDs. Nevertheless, it is noteworthy that the RvF plot for the covariate BMI and Cholesterol, unlike that of Age, shows a deviation trend that is not uniform throughout the horizontal y=0 line. For instance, the residuals in the RVF plot for Systolic vs. BMI (Figure 15) clusters around a Systolic blood pressure of 120 mmHg, then diverges in both a positive (towards +50) and negative direction (towards -50) when the Systolic blood pressure is between 130-140mmHg). Even though the residuals for Systolic vs. Cholesterol appear more evenly spread out along the horizontal line, they still show a recognizable trend (they show more deviation at Systolic blood pressure ~ 135 mmHg and converges when the Systolic pressure approaches 150 mmHg.

Thus far, these observations point to the variable Age as the potential covariate to produce the best simple linear regression model for Systolic. By comparing the R-squared, root MSE, and p-value of Age, Cholesterol, and BMI, Age is indeed the “best” and most suitable covariate to explain variations in Systolic blood pressure, with the highest (adjusted) R-squared value and the lowest MSE, among the 3 covariates. A simple regression model between Systolic and Age is thus provided (Figure 19).

Referring back to the scatterplot for Systolic vs. Age (Figure 8), several outliers can be identified. The regression model and RVF plot is repeated after excluding individual outliers (Figure 17 and 18, respectively), and changes to the adjusted R-squared is shown in Table 2. Even though removing ID=270 shows the greatest change to the adjusted R-squared value and root MSE, excluding *all* the stated IDs shows a more evenly spread out distribution around the horizontal y=0 line in the new RVF plot, while the new scatterplot for Systolic vs Age shows a greater “fit” with the data being “compressed” in the vertical direction. Therefore, the α and β coefficient for this regression model (Figure 21) will be used for the final “best” simple linear regression with the value Systolic.

Determining the “best” simple linear regression model: y = α + βx

From the regression model of Systolic vs Age *excluding* outliers (Figure 19), the *α* (constant) is calculated as 90.89493 and the *β* (coefficient) is calculated as 0.812071. The full linear equation for this regression model is thus as followed:

$$y = 0.812071x + 90.89493$$

According to this model, when Aco(x-value), the Systolic blood pressure is predicted to increase by 0.812071 units, and it is safe to conclude with 95% confidence that the increase in Systolic blood pressure will take values between 0.6813549 and 0.9427871 units. Likewise, when the age is equal to 0 (x = 0), the systolic blood pressure value would be 90.89493, and it is safe to conclude with 95% confidence that the value for Systolic blood pressure will be between 83.60576 and 98.1841 mmHg (since Age cannot take the value of 0, this interpretation is irrelevant in practice).

Prediction of Systolic when the covariate takes its lower quartile, median, and upper quartile values

A full summary of the lower quartile, median, and upper quartile value of Age is obtained in Table 3. From here, the respective value of Age at the 25th, 50th, and 75th percentile is plugged into the x-value of the above equation to predict the value of Systolic when this covariate takes it lower quartile, median, and upper quartile values:

Regression model equation: $y_{Systolic} = 0.812071x_{Age} + 90.89493$			
	Lower quartile	**Median**	**Upper quartile**
Age	48	55	62
Systolic blood pressure (mmHg)	129.874338	135.558835	141.243332

Table 3: summary of the predicted value of Systolic when Age takes its 25th, 50th, and 75th percentile

Part III: Multiple linear regression

Discussion of the most appropriate covariates to produce a multiple linear regression model with Systolic

Firstly, a multivariate regression model of all covariates was conducted to determine their individual correlation coefficient with the outcome Systolic (Figure 7 and 20). Among these, the covariates with the highest r value are Age, BMI, and Cholesterol (similar to the previous discussion), with all r-values being statistically significant (p-value << 0.05). From these calculations, 3 base models were initially produced (starting with 2 covariates for base model 1, and adding one additional covariate for the subsequent model 2 and 3, and their RVF + QQ plots are obtained respectively (for each base model) to ensure that they satisfy the necessary assumptions for a linear model.

Age and BMI were chosen for base model 1 because they both have the highest correlation coefficient (0.365 and 0.301) as well as a statistically significant p-value of << 0.05 with the outcome Systolic, consistent with the previous investigation. Moreover, their RVF and QQ plot (Figure 22 - 23) demonstrated that they have satisfied the assumptions of linear regression regarding the normality of the error term and the constant error term variance. Nevertheless, this regression model returns a lower adjusted R-squared value compared to the original regression model with all covariates, which means that after adjusting for the number of covariates and observations of each covariates, a smaller proportion of the variance in Systolic value is explained by Age & BMI (compare to the adjusted R-squared value for the multivariate regression model in Figure 21), which requires further modification & adjustments.

Cholesterol was the next covariate to be added to the initial base model 1 (termed “base model 2”) because it shows the third highest correlation coefficient with Systolic. According to the regression analysis (Figure 24) shows a higher R-squared value and a p-value of < 0.05, suggesting that there is a statistically significant linear regression between these 3 covariates with Systolic (after confirming that they have satisfied the assumptions for linear model by means of graphing their RVF and QQ plot, Figure 25 - 26). Since there are no clear recognizable patterns in these plots, it is reasonable to conclude that the error terms of this multivariate regression model is normally distributed and that the variance is constant & independent - satisfying the assumptions for linear regression. A summary of the adjusted R-squared value for Base model 1 and 2 can be found in Table 4 (Appendix).

Illness was the fourth covariate added (term “Base model 3”) because it was one of the variables with a low correlation coefficient but yielded a statistically significant p-value (< 0.05). The regression analysis (Figure 27) shows that the adjusted R-squared values of this base model increases while the root MSE decreases, suggesting that variations in Age, BMI, Cholesterol *and* Illness explain the largest proportion of the variance in Systolic value, compared to Base model 1 and 2. The QQ and RVF plots for Base model 3 are summarized in Figure 28 - 29.

After a Base model has been established, the remaining variables are added either individually or in combinations to the regression analysis of Base model 3 until the highest adjusted R-squared and the lowest root MSE are obtained (in other words, the final model with the greatest positive difference in the adjusted R-squared and negative difference in the root MSE from the original multivariate linear regression containing *all* given variables). The possible combinations and the adjusted R-squared value + root MSE is detailed in Table 5:

Potential final models	Adjusted R-squared	Deviations from the base model R-squared value (= 0.2344)	Root MSE	Deviations from the base model root MSE (= 19.6)
Base model 3 + Heart rate	0.2396	0.0052	19.533	-0.067
Base model 3 + Heart rate + Glucose	0.2395	0.0051	19.534	-0.066
Base model 3 + Heart rate + Cigarette	0.2389	0.0045	19.542	-0.058
Base model 3 + Glucose	0.2348	0.0004	19.595	-0.005
Base model 3 + Cigarettes	0.2336	-0.0008	19.610	0.010
Base model 3 + Glucose + Cigarettes	0.2340	-0.0004	19.605	0.005

Table 5: summary of the regression analysis parameters of different potential final multiple linear regression models

According to Table, it is evident that the regression model of option 1: base model 3 + Heart rate yields the highest adjusted R-squared, lowest root MSE as well as the greatest positive difference and negative difference from the original multivariate regression model, respectively (a full multivariate regression analysis of the final model is shown in Figure 30). The QQ and RVF plot of the final multivariate regression model confirms that the assumptions for linear regression are fulfilled (Figure 31 - 32).

Determining the "best" multiple linear regression model: y = ax + b

A multivariate regression analysis is conducted for the variable Age + BMI + Cholesterol + Illness + Heart rate against the outcome Systolic (Figure 33). The coefficient for different covariates are used to determine for the multiple regression model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_m x_m$ as seen below:

$$y = 0.7434254x_1 + 1.395386x_2 + 0.0487971x_3 + 5.27386x_4 + 0.1430449x_5 + 34.87331$$

where x_1 = Age, x_2 = BMI, x_3 = Cholesterol, x_4 = Illness, and x_5 = Heart rate. According to this model, for every 1 unit increase in Systolic blood pressure, Age (x_1) is predicted to increase by 0.7434254 units, BMI (x_2) is predicted to increase by 1.395386 units, Cholesterol (x_3) is predicted to increase by 0.0487971 units, Illness (x_4) is predicted to increase by 5.27386 units, and Heart rate (x_5) is predicted to increase by 0.1430449 units, on average. Likewise,when all covariates take the value of x ($x_1 = x_2 = x_3 = x_4 = x_5 = 0$), Systolic blood pressure would be 34.87331 mmHg

Prediction of the value of Systolic when the covariates take their lower quartile, median, and upper quartile

A full summary of the lower quartile, median, and upper quartile values of Age, Cholesterol, Illness, and Heart rate is obtained in Table 6. From here, the respective value of these variables at the 25th, 50th, and 75th percentile is plugged into the x-value of the above equation to predict the value of Systolic when this covariate takes it lower quartile, median, and upper quartile values (an exemplar calculation performed in Excel is provided in Figure 33).

Regression model: $y_{Systolic} = 0.7434254x_{Age} + 1.395386x_{BMI} + 0.0487971x_{Cholesterol} + 5.27386x_{Illness} + 0.1430449x_{Heart\ rate} + 34.87331$			
Covariates	**Lower quartile (25th percentile)**	**Median (50th percentile)**	**Upper quartile (75th percentile)**
Age	48	55	62
BMI	23.09	25.46	28.065
Cholesterol	210	240	270
Illness	0	0	0.5
Heartrate	68	75	84
Systolic	122.7516361	133.7279061	147.9551115

Table 6: summary of the predicted value of Systolic when all covariates takes their 25th, 50th, and 75th percentile

Part IV: Logistic regression

Application of a hypothesis test regarding the variable Systolic & Illness

A two-sample Student's t-test was conducted to determine if there is a difference in the mean Systolic blood pressure in people who do and do not develop heart-related conditions (Illness = 1 and Illness = 0, respectively). The null (H_0) and alternative hypothesis are as followed:

- H_0: $\mu_x = \mu_y$ or $\mu_x - \mu_y = 0$; there is no difference in the mean Systolic blood in group 1 and group 0,
- $H_{1'}$: $\mu_x \neq \mu_y$ or $\mu_x - \mu_y \neq 0$; there is a difference in the mean Systolic blood in group 1 and group 0,

A box plot (Figure 34) and histogram (Figure 35) comparing the systolic blood pressure of the 2 groups is obtained to ensure the data follows a normal distribution. Several outliers are identified and removed before constructing the box plot

(not shown) and the results show that data in both groups follows a normal distribution with equal tails in the box plot (although there is a marginal degree of right-skewness). Figure 36 summarizes the mean, standard deviation (SD), and min/ max value for Systolic blood pressure in group 0 and 1. Since the difference in SD(Illness = 0) and SD(Illness = 1) is modest, it is safe to conclude that the 2 groups have equal variances. Lastly, since each participant (by ID) can only take a value of Illness = 0 or Illness =1 (i.e they do or do not have heart-related problems), the independence assumption is satisfied.

Figure 37 summarizes the results of a 2-sample t-test assuming equal variance. With a t_0 of -6.4089 under the $t_{(750\ -1)\ +\ (250\ -1)}$ distribution with a degree of freedom = 998, we obtain a p-value of 0.000 or << 0.05. Therefore, we reject the null hypothesis and conclude that there is a difference in the mean Systolic blood pressure.

Binary logistic regression model for Illness with the covariate Systolic

For an outcome such that P(Illness = 1) or the probability that a random ID has a heart-related problem, the odds of this event occurring is: $\frac{P\ (Illness = 1)}{1-\ P(Illness\ =1)}$. Since the odds can only take values between 0 → ∞, in order to "transform" this data set so that it can take on the same "scale" as the continuous variable Systolic (-∞ → +∞), this ratio needs to take the natural log transformation, or $\ln(\frac{P\ (Illness = 1)}{1-\ P(Illness\ =1)})$. Hence, we can rewrite this equation as logitP(Illness = 1) = $\beta_0 + \beta_1 x$. In order to calculate P(Illness), we can rearrange this equation to: P(Illness = 1) = $\frac{e^{\beta 0\ +\ \beta 1 x}}{1\ +\ e^{\beta 0\ +\ \beta 1 x}}$.

Using the "logit" command in STATA, the corresponding constant β_0 and coefficient β_1 can be obtained from Figure 38. Since the P>|z| of the coefficient is less than 0.05, there is a statistically significant correlation between the log of the probability of a participant having heart-related conditions and changes in Systolic blood pressure. The final logistic regression model for Illness using the covariate Systolic is summarized by the equation:

$$\text{logit}(P = \text{Illness}) = -3.832896 + 0.0197971x, \text{ for the log of the odds of P(Illness = 1)}$$

$$P(\text{Illness} = 1) = \frac{e^{-3.832896\ +\ 0.0197971x}}{1\ +\ e^{-3.832896\ +\ 0.0197971x}}, \text{ for the probability that a random participant has heart-related problems}$$

The coefficient β = 0.0197971 indicates the impact of systolic blood pressure on the *log of the odds* of having heart-related conditions: for every 1 unit increase in Systolic (x), the log-odds of having heart-related conditions increase by 0.0197971 units. Likewise, when the Systolic blood pressure is 0, the *log of the probability* of having heart-related conditions is 0.0197971 (although this interpretation is irrelevant in practice).

Prediction of P(Illness = 1) when Systolic takes its lower quartile, median, and upper quartile

A full summary of the lower quartile, median, and upper quartile values of Systolic is obtained in Table 7. From here, the respective value of these variables at the 25th, 50th, and 75th percentile is plugged into the x-value of the above equation to predict probability of an individual having heart-related conditions when the value of Systolic when this covariate takes it lower quartile, median, and upper quartile values (detailed calculations is shown in Figure 39).

Regression model: P(Illness = 1) = $\frac{e^{-3.832896\ +\ 0.0197971x}}{1+\ e^{-3.832896\ +\ 0.0197971x}}$			
	Lower quartile	**Median**	**Upper quartile**
Systolic	120	133	148
P(Illness = 1)	0.465754401	0.602459691	0.8107642

Table 7: summary of the predicted value of P(Illness = 1) when Systolic takes its 25th, 50th, and 75th percentile

Discussion of the benefits of logistic regression when determining the correlation between Illness & Systolic

In the given dataset, performing a logistic regression to investigate the correlation between Illness and Systolic is suitable for the binary nature of the covariate Illness, since a nonlinear equation will more accurately estimate and/ or the resulting response (probability of Illness = 1) from the covariate affecting it (Systolic). This is because the outcome for Illness is probabilistic and hence can only take values from 0 to 1, while that for Systolic can take any integer from -∞ to +∞. It is also less sensitive to non-normally distributed data and unequal variance in different groups (12). Moreover, in a linear regression model, since the regression model will attempt to fit all data points on the line of best fit by minimizing the error terms, it will be less sensitive if one group has a larger sample size compared to the other (which is the case in our model, considering that Illness = 1 has a sample size n = 250 and Illness = 0 has a sample size n = 750. If a scatter plot of Systolic vs Illness is produced (Figure 40), it is evident that the line might be able to predict whether an individual will or will not develop heart-related conditions at a certain range (or values) of Systolic; nevertheless, outside of those range/ values, this prediction is flawed and not statistically accurate.

Bibliography

1. Wannamethee G, Shaper AG. The Association between Heart Rate and Blood Pressure, Blood Lipids and Other Cardiovascular Risk Factors. European Journal of Cardiovascular Prevention & Rehabilitation. 1994 Oct 1;1(3):223–30.
2. Bønaa KH, Arnesen E. Association between heart rate and atherogenic blood lipid fractions in a population. The Tromsø Study. Circulation. 1992 Aug;86(2):394–405.
3. Tardif JC . Heart rate and atherosclerosis. European Heart Journal Supplements. 2009 Aug 1;11(Suppl D):D8–12.
4. Peer N, Lombard C, Steyn K, Levitt N. Elevated resting heart rate is associated with several cardiovascular disease risk factors in urban-dwelling black South Africans. Scientific Reports. 2020 Mar 12;10(1).
5. Wang C, Li Y, Li L, Wang L, Zhao J, You A, et al. Relationship between Resting Pulse Rate and Lipid Metabolic Dysfunctions in Chinese Adults Living in Rural Areas. Makishima M, editor. PLoS ONE [Internet]. 2012 Nov 7 [cited 2021 Feb 1];7(11):e49347. Available from: https://dx.doi.org/10.1371%2Fjournal.pone.0049347
6. Soran H, Adam S, Mohammad JB, Ho JH, Schofield JD, Kwok S, et al. Hypercholesterolaemia – practical information for non-specialists. Archives of Medical Science : AMS [Internet]. 2018 Jan 1;14(1):1–21. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778427/
7. Tsang S, Duncan GE, Dinescu D, Turkheimer E. Differential models of twin correlations in skew for body-mass index (BMI). Caruso C, editor. PLOS ONE. 2018 Mar 28;13(3):e0194968.
8. Kawada T. Body mass index is a good predictor of hypertension and hyperlipidemia in a rural Japanese population. International Journal of Obesity. 2002 May;26(5):725–9.
9. .Nwaiwu O, Ibe B. Relationship between Serum Cholesterol and body mass index in Nigeria schoolchildren aged 2–15 years [Internet]. Journal of Tropical Pediatrics . Oxford Academics; 2015. Available from: https://academic.oup.com/tropej/article/61/2/126/1728479
10. Humphreys S. The Unethical Use of BMI in Contemporary General Practice. British Journal of General Practice [Internet]. 2010 Sep 1;60(578):696–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2930234/
11. Kapoor N. Thin Fat Obesity: The Tropical Phenotype of Obesity [Internet]. Feingold KR, Anawalt B, Boyce A, Chrousos G, de Herder WW, Dhatariya K, et al., editors. PubMed. South Dartmouth (MA): MDText.com, Inc.; 2000. Available from: https://www.ncbi.nlm.nih.gov/books/NBK568563/
12. Jing H. Why Linear Regression is not suitable for Binary Classification [Internet]. Medium. 2020. Available from: https://towardsdatascience.com/why-linear-regression-is-not-suitable-for-binary-classification-c64457be8e28