ICIC Express Letters
Part B: Applications
Volume **12**, Number **8**, August **2021**

ICIC International ©2021 ISSN 2185-2766
pp. **723–731**

PREDICTING STROKE, HYPERTENSION, AND DIABETES DISEASES BASED ON INDIVIDUAL CHARACTERISTICS

Ferry Vincenttius Ferdinand[1,*], Johan Sebastian[1] and Friska Natalia[2]

[1]Department of Mathematics
Universitas Pelita Harapan
Tangerang, Banten 15811, Indonesia
[*]Corresponding author: ferry.vincenttius@uph.edu

[2]Department of Information Systems
Universitas Multimedia Nusantara
Scientia Boulevard, Gading Serpong, Tangerang, Banten 15811, Indonesia
friska.natalia@umn.ac.id

Received December 2020; accepted February 2021

Abstract. *Health awareness has been significantly increasing nowadays. However, it is expensive and time-consuming to assess a person's health condition precisely. The purpose of this study is to predict specific diseases based on individual characteristics. This prediction can be used in the future as an alternative to determine one's health condition instead of using medical check-up. Several methods such as ARIMA (Autoregressive Integrated Moving Average), Principal Component Analysis (PCA), and regression are used to get to this study's purpose. The result, which includes demography, lifestyle behavior, economic condition, and geographic location, can predict the prevalence of stroke, hypertension, and diabetes.*

Keywords: Stroke, Hypertension, Diabetes, ARIMA, PCA, Regression, GLM, Lifestyle, Geographical location, GRDP

1. **Introduction.** As time goes by, the world knows the danger of disease and the importance of health. WHO declared that based on Noncommunicable Diseases Country Profiles 2018, 41% of Indonesia's death is based on cardiovascular and diabetes [1]. Currently, the only way to know the health condition is through a medical check-up, which takes much time, energy, and money [2]. Therefore, much research has been conducted to predict these diseases [3]. One possible solution is making a prediction based on the easily obtained data from one's life. It is essential to determine factors considered as the significant and/or minor risk before predicting. Hopefully, one could get a clearer picture of his/her health condition without going into a medical check-up.

Nowadays, one of the most developed methods to predict these data is by using predictive analytics. The connection between the risk factors and the condition could be obtained by predictive analytics. Furthermore, big data usage greatly aids prediction accuracy [3]. It would be easier to predict using their data since it is their health and lifestyle. It also costs nothing. While in other research, machine learning was used to predict many diseases, such as cardiovascular disease risk [4]. The results indicated that the accuracy is above 80%, which can be concluded that prediction using big data is a good alternative.

In this study, a few dimensions are inspected, such as lifestyle behavior [5], demographics such as gender [6], economic [6-8], and geography [6]. From these aspects, one could get a prediction on the health condition. The health conditions that will be predicted are hypertension, stroke, and diabetes. Hypertension is a condition where a person's blood

DOI: 10.24507/icicelb.12.08.723

vessels consistently have high blood pressure, which is hard to decrease. Diabetes is a chronic disease based on a high level of sugar in one's blood. Meanwhile, stroke is a disease caused by a lack of oxygen flowing into a brain and can partly stop one's brain. This is in accordance with the previous research, which states that obesity, behavior, gender, economic, and geographical location increases the risk of hypertension, cardiovascular disease, and diabetes [9-13].

This paper will describe how the prediction of diseases in Indonesia can be found based on individual characteristics, mainly using GLM. The remainder of the paper is organized as follows. Section 2 points to the problem statement, Section 3 presents the research methodology, Section 4 discusses the main results, and Section 5 makes to the conclusions.

2. **Problem Statement.** According to the Ministry of Health Indonesia (Kementrian Kesehatan Republik Indonesia) in 2018 shown in Figure 1, the prevalence of hypertension, stroke, and diabetes in Indonesia is 34.1% (age 18 and above), 10.9% (age 15 and above), and 10.9% (age 15 and above) respectively. While this number seems concerning, much research has also concluded that lifestyle and health conditions are related. Despite that, research for this particular field in Indonesia is still rare. Thus, this research has taken more significant dimensions to work on, such as age group, gender, economic condition, and geographic location. In effect, health conditions can be predicted with more accuracy.

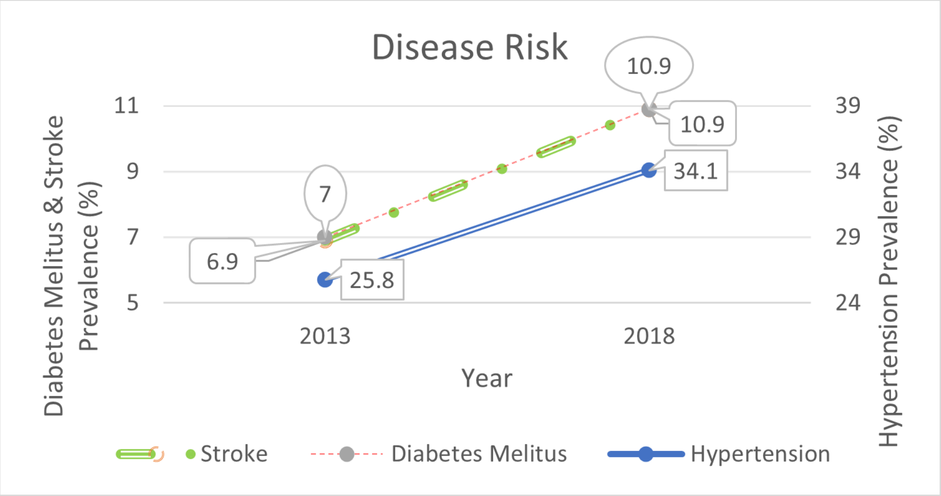


FIGURE 1. Disease risk

Thus, it will lead to some problems in this research, such as predicting lifestyle, economic situation, and geographic location for hypertension, stroke, and diabetes. It is essential to find out which linear regression model will be used to predict these health conditions.

3. **Research Methodology.** Figure 2 explained how this research would be done. This research's data are lifestyle behavior, body mass index, gender, age group, stroke prevalence, hypertension prevalence, and diabetes prevalence in Indonesia for each province in the years 2007, 2013, and 2018 from RISKESDAS (Riset Kesehatan Dasar). These data were analyzed by using curve fitting to get its predicted data from 2007 to 2020. The function for curve fitting $f_i(x)$ to the M-th degree can be written as:

$$f_i(x) = c_{i,1} + c_{i,2}x + c_{i,3}x^2 + \cdots + c_{i,M+1}x^M \quad (1)$$

with $c_{i,1}, c_{i,2}, \ldots, c_{i,M+1}$ as the corresponding coefficient of the i-th variable. Since GRDP (Gross Regional Domestic Product) is being collected countless time and periodically, it is analyzed using Autoregressive Integrated Moving Average (ARIMA (p, d, q)) to get its predicted data from 2007 to 2020. A time series $\{Y_t\}$ is said to follow ARIMA (p, d, q)

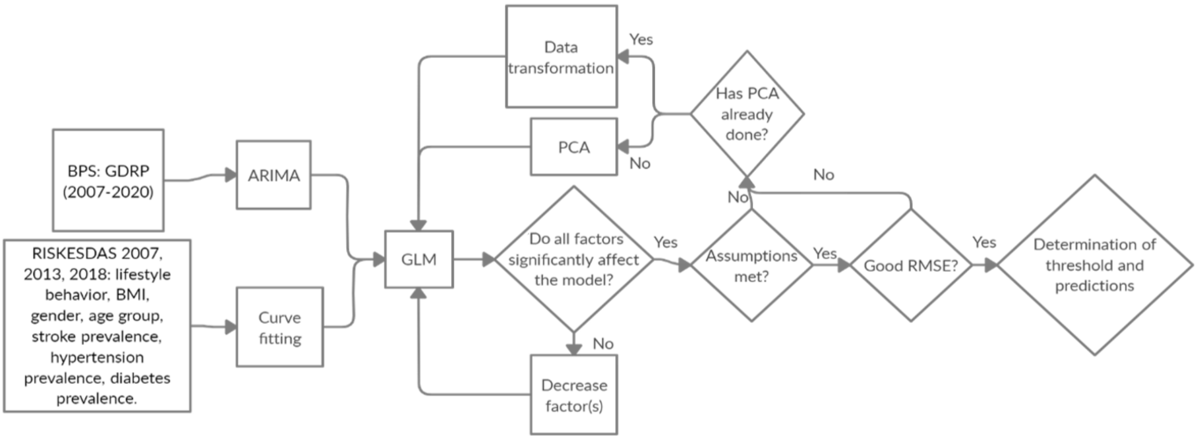

BPS: GDRP (2007-2020)
RISKESDAS 2007, 2013, 2018: lifestyle behavior, BMI, gender, age group, stroke prevalence, hypertension prevalence, diabetes prevalence.
ARIMA
Curve fitting
GLM
Data transformation
PCA
Yes
No
Has PCA already done?
Do all factors significantly affect the model?
Yes
No
Decrease factor(s)
Assumptions met?
No
Yes
Good RMSE?
No
Yes
Determination of threshold and predictions

FIGURE 2. Flowchart

model if the d-th difference $\nabla^d Y_t = W_{it}$ is a stationary Autoregressive Moving Average (ARMA (p, q)) [14] for the i-th province whereas the variable p shows the autoregressive number, d shows how many differences the time series needed for $W_{i,t}$ to be stationary, and q shows the error of delayed prediction on the prediction equation. The general prediction equation will be

$$W_{i,t} = \phi_{i,1}W_{i,t-1}+\phi_{i,2}W_{i,t-2}+\cdots+\phi_{i,p}W_{i,t-p}+e_{i,t}-\theta_{i,1}e_{i,t-1}-\theta_{i,2}e_{i,t-2}-\cdots-\theta_{i,q}e_{i,t-q} \quad (2)$$

To identify which ARIMA model is suitable, d must be selected in the first place to create a time-series, which has a stable mean and variance.

The data for the diseases are presented as prevalence. For the province data, 34 provinces in Indonesia were put into a single variable (X_1) by code for each province from 1 to 6 and were converted into 5 dummy variables $(D_1, D_2, D_3, D_4, D_5)$ so that it can fit in the regression models as binary variables. $\forall_i D_i = 0$ except $D_1 = 1$ for Jawa, $D_2 = 1$ for Bali, Nusa Tenggara, and Maluku, $D_3 = 1$ for Kalimantan, $D_4 = 1$ for Sulawesi, and $D_5 = 1$ for Papua. The technique used is binary logistic regression in the Generalized Linear Model (GLM). Binary logistic regression is used since the results are binary [15]. It estimates the probability of having a characteristic given a condition. The probability of the i-th trial is denoted by

$$\pi_{i,j} = Pr\left(Y_{i,j} = 1 | X_{i,j} = x_{i,j}\right) = \frac{\exp\left(\beta_{i,0} + \beta_{i,1}x_{i,1} + \cdots + \beta_{i,j}x_{i,j}\right)}{1 + \exp\left(\beta_{i,0} + \beta_{i,1}x_{i,1} + \cdots + \beta_{i,j}x_{i,j}\right)} \quad (3)$$

whereas the variable $Y_{i,j}$ shows the response variable to determine if any characteristic wanted by this research with $Y_{i,j} = 1$ if characteristic wanted by this research is shown and $Y_{i,j} = 0$ if it is not shown, $x_{i,j}$ shows the j-th observation data for the i-th response variables, $\beta_{i,j}$ shows the coefficient of $x_{i,j}$. Using binary logistic regression, assumptions used are the data $Y_{i,1}, Y_{i,2}, \ldots, Y_{i,n}$ which are independently distributed, i.e., cases are independent, distribution of $Y_{i,j}$ is $Bin\left(n_{i,j}, \pi_{i,j}\right)$, and the dependent variable does not need to be normally distributed. It also does not assume a linear relationship between the dependent variable and the independent variable, but linear relationship exists between the logit of the response and the explanatory variables and written as:

$$logit\left(\pi_{i,j}\right) = \beta_{i,0} + \sum_{k=1}^{j} \beta_{i,k}X_{i,k} \quad (4)$$

Since independent variables can be some other nonlinear transformations of the original independent variables, the homogeneity of variance does not need to be satisfied; errors need to be independent but not necessary to be normally distributed; parameters being estimated using MLE (Maximum Likelihood Estimation); not more than 20% of the expected cells counts are less than 5 in a goodness-of-fit measures; multicollinearities must not exist through VIF (Variance Inflation Factor). Suppose that a dataset contains m independent variables $(x_{i,1}, x_{i,2}, \ldots, x_{i,m})$, $\mathit{VIF}_{i,j}$ (predictor of $x_{i,j}$) can be calculated using the linear relationship between $x_{i,j}$ and other independent variables. $\mathit{VIF}_{i,j}$ can be written as:

$$\mathit{VIF}_{i,j} = \frac{1}{\left(1 - R_{i,j}^2\right)} \tag{5}$$

whereas $R_{i,j}^2$ is the coefficient of determination of regression of $x_{i,j}$ on the other independent variables. For examining the multicollinearity problem, VIF value of 5 is used so that for every VIF value more than or equal to 5, there exists multicollinearity [16].

If all these assumptions are met, GLM will be applied to determining its regression coefficient. Otherwise, PCA (Principal Component Analysis) will be used by creating new variables (which are the linear combinations of all the variables). The number of new variables can be determined by the last eigenvalue that has a value more than or equal to 1. However, assumptions that need to be satisfied are all the basis vectors $\{p_1, p_2, \ldots, p_n\}$ which are orthonormal and the directions with the largest variances are the most principal [17].

4. **Main Results.** GLM binary logistic regression is conducted for each disease separately. The general equation for predicting these disease(s):

$$Y_i = \frac{e^{(\beta_{i,0}+\alpha_{i,1}D_{i,1}+\alpha_{i,2}D_{i,2}+\alpha_{i,3}D_{i,3}+\alpha_{i,4}D_{i,4}+\alpha_{i,5}D_{i,5}+\beta_{i,2}X_{i,2}+\beta_{i,3}X_{i,3}+\cdots+\beta_{i,34}X_{i,34}+\beta_{i,35}X_{i,35})}}{1+e^{(\beta_{i,0}+\alpha_{i,1}D_{i,1}+\alpha_{i,2}D_{i,2}+\alpha_{i,3}D_{i,3}+\alpha_{i,4}D_{i,4}+\alpha_{i,5}D_{i,5}+\beta_{i,2}X_{i,2}+\beta_{i,3}X_{i,3}+\cdots+\beta_{i,34}X_{i,34}+\beta_{i,35}X_{i,35})}} \tag{6}$$

Variable Y_i represents prevalence of the ith disease with $i = \{1, 2, 3\}$ and Y_1 represents stroke prevalence, Y_2 represents hypertension prevalence, and Y_3 represents diabetes prevalence. The variables α_1 to α_5 and β_0 to β_{35} are used as the coefficient. Other variables in equation are defined in Table 1.

Before using Equation (6), multicollinearity test is needed. Since the result shows GVIF (Generalized Variance Inflation Factor) and GVIF$^{1/(2\cdot df)}$, there will be several conditions

Table 1. Variables and its description

Variables	Description	Variables	Description	Variables	Description
Factor (X_1) or D_1, D_2, D_3, D_4, D_5	Province code	X_{13}	Proportion of tobacco consumed	X_{25}	Age 25-29
X_2	Proportion of not consuming fruits/vegetables each day in a week.	X_{14}	Average sum of tobacco consumed (per day)	X_{26}	Age 30-34
X_3	Average of fruit consumption (portion per day in a week)	X_{15}	Proportion of inactive physical activity (no physical activity done $\geq$ 30 minutes per day)	X_{27}	Age 35-39
X_4	Average of vegetable consumption (portion per day in a week)	X_{16}	Gross Regional Domestic Product (GRDP)	X_{28}	Age 40-44
X_5	Proportion of sweet food consumption ($\geq$ 1 time per day)	X_{17}	Body mass classified as obesity	X_{29}	Age 45-49
X_6	Proportion of salty food consumption ($\geq$ 1 time per day)	X_{18}	Body mass classified as overweight	X_{30}	Age 50-54
X_7	Proportion of fatty food consumption ($\geq$ 1 time per day)	X_{19}	Body mass classified as normal	X_{31}	Age 55-59
X_8	Proportion of grilled food consumption ($\geq$ 1 time per day)	X_{20}	Body mass classified as underweight	X_{32}	Age 60-64
X_9	Proportion of food that contains animal protein consumption ($\geq$ 1 time per day)	X_{21}	Proportion of male	X_{33}	Age 65-69
X_{10}	Proportion of food that contains preservatives ($\geq$ 1 time per day)	X_{22}	Proportion of female	X_{34}	Age 70-74
X_{11}	Proportion of caffeinated drinks consumption ($\geq$ 1 time per day)	X_{23}	Age 15-19	X_{35}	Age 75+
X_{12}	Proportion of instant noodle consumption ($\geq$ 1 time per day)	X_{24}	Age 20-24		

to convert it into VIF. GVIF can be directly used for the variables that have one degree of freedom, while others need to take the square root of GVIF$^{1/(2\cdot df)}$ to use it as its VIF. If VIF ≥ 5, we conclude there exists multicollinearity. Based on the result, the variable $X_{i,17}$ to $X_{i,20}$ and $X_{i,23}$ to $X_{i,35}$ give scores of VIF more than 5, which indicates multicollinearity. Thus, we use PCA to overcome this multicollinearity. By performing PCA, 7 components are being used $(FAC_1, FAC_2, \ldots, FAC_7)$ that are linear combinations of the 35 variables $(X_1, X_2, \ldots, X_{35})$ since the eigenvalue of the 8th component and above is less than one and the total variance described by all those 7 factors already reached 83.58%. Thus, each new variable can be written as:

$$\begin{aligned} FAC_1 = {} & 0.007X_1 - 0.017X_2 - 0.029X_3 - 0.009X_4 - 0.021X_5 + 0.034X_6 - 0.001X_7 \\ & + 0.043X_8 + 0.017X_9 - 0.047X_{10} - 0.003X_{11} + 0.011X_{12} - 0.021X_{13} \\ & - 0.001X_{14} - 0.004X_{15} + 0.048X_{16} - 0.023X_{17} + 0.018X_{18} + 0.002X_{19} \\ & - 0.007X_{20} + 0.021X_{21} - 0.024X_{22} + 0.079X_{23} + 0.079X_{24} + 0.079X_{25} \\ & + 0.077X_{26} + 0.077X_{27} + 0.078X_{28} + 0.078X_{29} + 0.077X_{30} + 0.076X_{31} \\ & + 0.075X_{32} + 0.076X_{33} + 0.075X_{34} + 0.073X_{35} \end{aligned} \tag{7}$$

$$\begin{aligned} FAC_2 = {} & 0.033X_1 + 0.166X_2 + 0.163X_3 - 0.007X_4 + 0.027X_5 - 0.033X_6 + 0.126X_7 \\ & - 0.008X_8 - 0.0002X_9 + 0.036X_{10} - 0.077X_{11} - 0.053X_{12} + 0.01X_{13} \\ & + 0.007X_{14} + 0.022X_{15} + 0.067X_{16} - 0.082X_{17} - 0.207X_{18} + 0.166X_{19} \\ & + 0.189X_{20} - 0.044X_{21} + 0.052X_{22} - 0.03X_{23} - 0.026X_{24} - 0.024X_{25} \\ & - 0.016X_{26} - 0.01X_{27} - 0.09X_{28} - 0.09X_{29} - 0.04X_{30} + 0.001X_{31} \\ & + 0.004X_{32} - 0.003X_{33} - 0.012X_{34} - 0.005X_{35} \end{aligned} \tag{8}$$

$$\begin{aligned} FAC_3 = {} & 0.051X_1 + 0.106X_2 + 0.032X_3 - 0.076X_4 - 0.083X_5 - 0.146X_6 + 0.006X_7 \\ & - 0.06X_8 + 0.011X_9 - 0.017X_{10} + 0.058X_{11} - 0.017X_{12} - 0.025X_{13} \\ & - 0.103X_{14} + 0.125X_{15} + 0.027X_{16} + 0.173X_{17} - 0.125X_{18} - 0.022X_{19} \\ & + 0.042X_{20} - 0.427X_{21} + 0.435X_{22} - 0.027X_{23} - 0.021X_{24} - 0.023X_{25} \\ & - 0.02X_{26} - 0.019X_{27} - 0.017X_{28} - 0.015X_{29} - 0.011X_{30} - 0.005X_{31} \\ & + 0.001X_{32} + 0.004X_{33} + 0.001X_{34} + 0.003X_{35} \end{aligned} \tag{9}$$

$$\begin{aligned} FAC_4 = {} & -0.048X_1 + 0.038X_2 + 0.044X_3 - 0.088X_4 + 0.386X_5 + 0.178X_6 - 0.068X_7 \\ & + 0.009X_8 + 0.317X_9 + 0.017X_{10} - 0.140X_{11} + 0.154X_{12} - 0.101X_{13} \\ & - 0.012X_{14} + 0.343X_{15} + 0.043X_{16} + 0.04X_{17} + 0.008X_{18} - 0.055X_{19} \\ & - 0.008X_{20} - 0.010X_{21} + 0.004X_{22} + 0.011X_{23} + 0.007X_{24} + 0.009X_{25} \\ & + 0.011X_{26} + 0.006X_{27} - 0.003X_{28} - 0.010X_{29} - 0.013X_{30} - 0.015X_{31} \\ & - 0.016X_{32} - 0.020X_{33} - 0.022X_{34} - 0.024X_{35} \end{aligned} \tag{10}$$

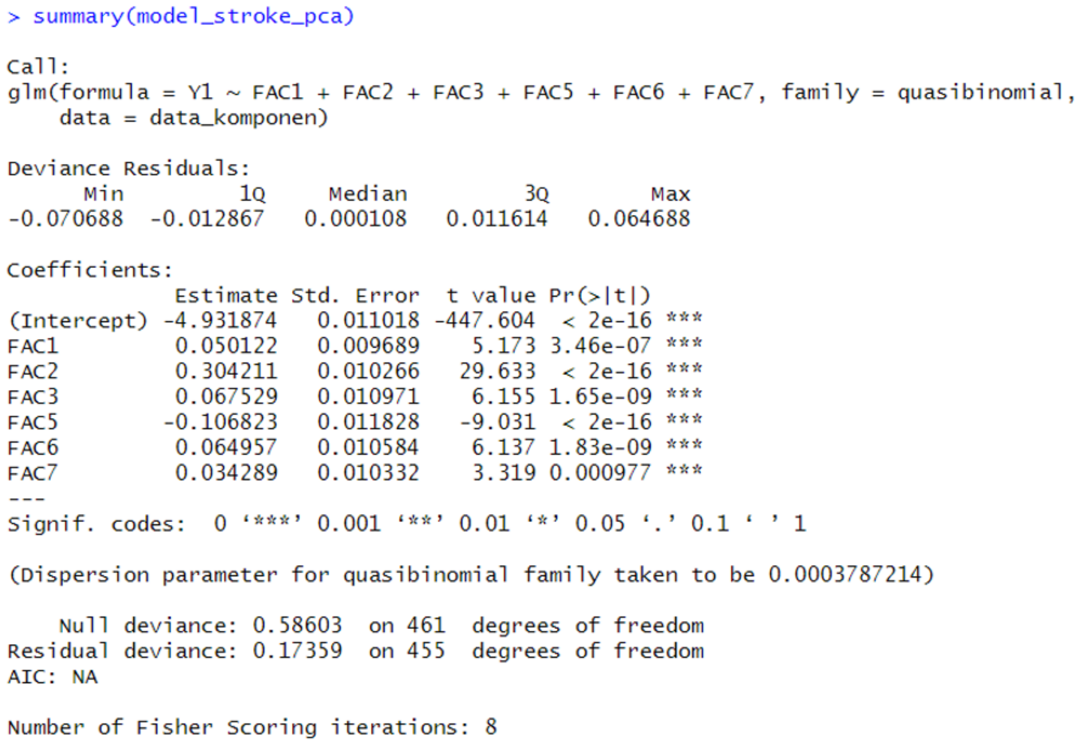
$$\begin{aligned} FAC_5 = {} & 0.404X_1 + 0.017X_2 - 0.105X_3 - 0.027X_4 - 0.246X_5 - 0.134X_6 - 0.001X_7 \\ & + 0.364X_8 + 0.128X_9 - 0.025X_{10} + 0.134X_{11} + 0.204X_{12} - 0.196X_{13} \\ & - 0.086X_{14} + 0.036X_{15} + 0.027X_{16} - 0.045X_{17} - 0.067X_{18} + 0.088X_{19} \\ & + 0.058X_{20} - 0.019X_{21} + 0.036X_{22} - 0.001X_{23} + 0.004X_{24} + 0.002X_{25} \\ & + 0.008X_{26} + 0.019X_{27} + 0.026X_{28} + 0.025X_{29} + 0.023X_{30} + 0.025X_{31} \\ & + 0.028X_{32} + 0.028X_{33} + 0.020X_{34} + 0.015X_{35} \end{aligned} \tag{11}$$

$$\begin{aligned} FAC_6 = {} & 0.150X_1 + 0.026X_2 + 0.082X_3 + 0.011X_4 - 0.023X_5 + 0.118X_6 + 0.264X_7 \\ & - 0.108X_8 + 0.060X_9 + 0.598X_{10} + 0.238X_{11} + 0.161X_{12} + 0.182X_{13} \\ & - 0.237X_{14} - 0.107X_{15} + 0.024X_{16} + 0.101X_{17} - 0.088X_{18} + 0.022X_{19} \\ & + 0.026X_{20} + 0.014X_{21} - 0.025X_{22} - 0.050X_{23} - 0.052X_{24} - 0.046X_{25} \end{aligned}$$

$$
\begin{aligned}
&- 0.027X_{26} - 0.010X_{27} - 0.011X_{28} - 0.021X_{29} - 0.028X_{30} - 0.030X_{31} \\
&- 0.029X_{32} - 0.033X_{33} - 0.044X_{34} - 0.042X_{35} \qquad (12)
\end{aligned}
$$

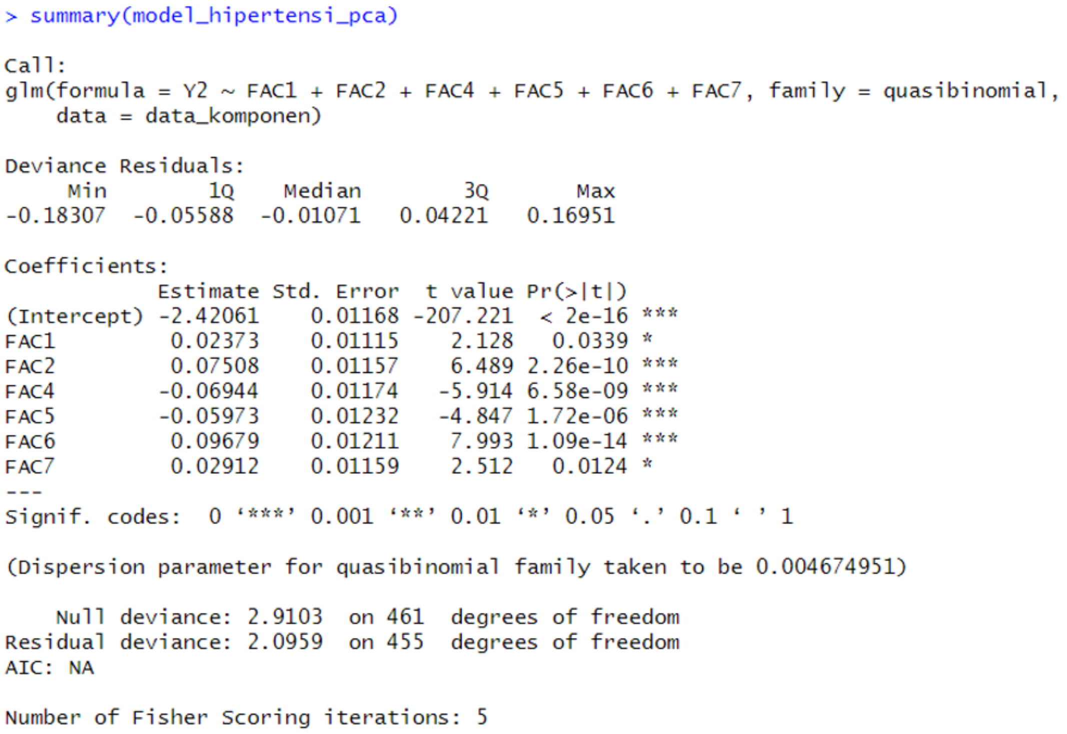
$$
\begin{aligned}
FAC_7 = {} & 0.036X_1 + 0.131X_2 - 0.170X_3 - 0.585X_4 - 0.281X_5 - 0.017X_6 + 0.002X_7 \\
&+ 0.026X_8 + 0.035X_9 - 0.017X_{10} + 0.235X_{11} + 0.200X_{12} + 0.271X_{13} \\
&+ 0.231X_{14} + 0.085X_{15} + 0.068X_{16} - 0.072X_{17} - 0.038X_{18} + 0.059X_{19} \\
&+ 0.054X_{20} - 0.095X_{21} + 0.064X_{22} + 0.029X_{23} + 0.011X_{24} + 0.008X_{25} \\
&+ 0.018X_{26} + 0.034X_{27} + 0.033X_{28} + 0.017X_{29} + 0.003X_{30} - 0.002X_{31} \\
&- 0.004X_{32} - 0.010X_{33} - 0.029X_{34} - 0.050X_{35} \qquad (13)
\end{aligned}
$$

For the next step, GLM is used to find regression equation for each disease model using new components that we got from PCA before. As the assumptions, multicollinearity and autocorrelation are checked while we omit the normality, linearity, and heteroskedasticity assumptions due to GLM's properties [18]. Based on the results and analysis, regression equation for stroke, hypertension, and diabetes model can be written respectively in Figures 3-8. As the first step, the component FAC_4 in stroke model having p-value of 0.5646 which is larger than 0.05 (using significance level of 95%). Thus, FAC_4 does not significantly affect the regression model for stroke and needs to be removed. Then GLM is conducted again and all components are significantly affecting the regression model in Figure 3 by having p-value less than 0.05. Multicollinearity test is then conducted in Figure 6, and since all components have scores of VIF less than 5, it can be concluded that there are no multicollinearity and thus the regression model for stroke is valid. For the accuracy of the model, we calculate the Root Mean Square Error (RMSE) for the predicted value of the model with the initial data. The RMSE for the stroke regression model is 0.0067. The same procedure is also applied for the hypertension and diabetes model. For hypertension model, FAC_3 will be removed since its p-value is 0.3404 ($p > 0.05$). GLM is then conducted and all components are significantly affecting the regression model in Figure 4. Multicollinearity test is then conducted and shown in Figure 7. From



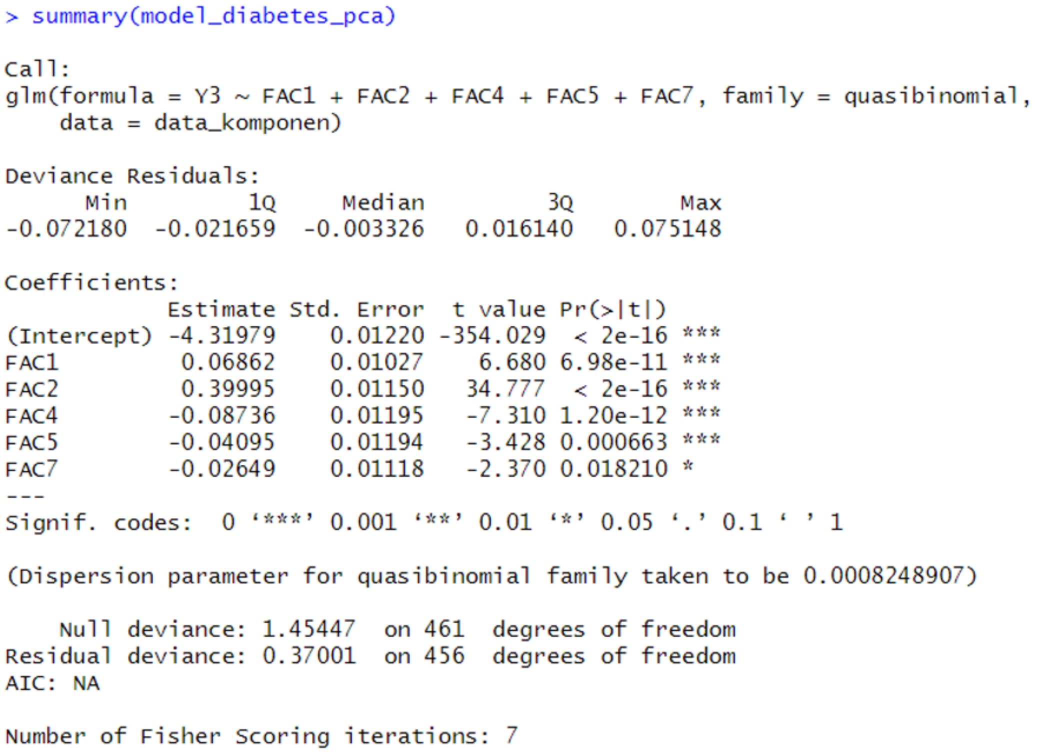
```
> summary(model_stroke_pca)

Call:
glm(formula = Y1 ~ FAC1 + FAC2 + FAC3 + FAC5 + FAC6 + FAC7, family = quasibinomial,
    data = data_komponen)

Deviance Residuals:
      Min          1Q      Median          3Q         Max
-0.070688   -0.012867    0.000108    0.011614    0.064688

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -4.931874   0.011018 -447.604  < 2e-16 ***
FAC1         0.050122   0.009689    5.173 3.46e-07 ***
FAC2         0.304211   0.010266   29.633  < 2e-16 ***
FAC3         0.067529   0.010971    6.155 1.65e-09 ***
FAC5        -0.106823   0.011828   -9.031  < 2e-16 ***
FAC6         0.064957   0.010584    6.137 1.83e-09 ***
FAC7         0.034289   0.010332    3.319 0.000977 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 0.0003787214)

    Null deviance: 0.58603  on 461  degrees of freedom
Residual deviance: 0.17359  on 455  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 8
```

Figure 3. Stroke PCA model



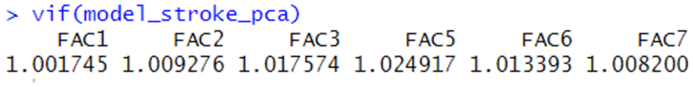
```
> summary(model_hipertensi_pca)

Call:
glm(formula = Y2 ~ FAC1 + FAC2 + FAC4 + FAC5 + FAC6 + FAC7, family = quasibinomial,
    data = data_komponen)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.18307  -0.05588  -0.01071   0.04221   0.16951

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) -2.42061    0.01168 -207.221  < 2e-16 ***
FAC1         0.02373    0.01115    2.128   0.0339 *
FAC2         0.07508    0.01157    6.489 2.26e-10 ***
FAC4        -0.06944    0.01174   -5.914 6.58e-09 ***
FAC5        -0.05973    0.01232   -4.847 1.72e-06 ***
FAC6         0.09679    0.01211    7.993 1.09e-14 ***
FAC7         0.02912    0.01159    2.512   0.0124 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 0.004674951)

    Null deviance: 2.9103  on 461  degrees of freedom
Residual deviance: 2.0959  on 455  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure 4. Hypertension PCA model



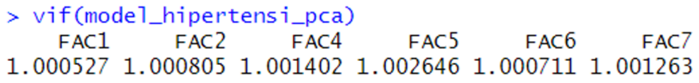
```
> summary(model_diabetes_pca)

Call:
glm(formula = Y3 ~ FAC1 + FAC2 + FAC4 + FAC5 + FAC7, family = quasibinomial,
    data = data_komponen)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.072180  -0.021659  -0.003326   0.016140   0.075148

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) -4.31979    0.01220 -354.029  < 2e-16 ***
FAC1         0.06862    0.01027    6.680 6.98e-11 ***
FAC2         0.39995    0.01150   34.777  < 2e-16 ***
FAC4        -0.08736    0.01195   -7.310 1.20e-12 ***
FAC5        -0.04095    0.01194   -3.428 0.000663 ***
FAC7        -0.02649    0.01118   -2.370 0.018210 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 0.0008248907)

    Null deviance: 1.45447  on 461  degrees of freedom
Residual deviance: 0.37001  on 456  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

Figure 5. Diabetes PCA model



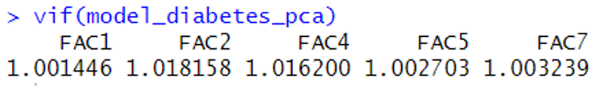
```
> vif(model_stroke_pca)
    FAC1     FAC2     FAC3     FAC5     FAC6     FAC7
1.001745 1.009276 1.017574 1.024917 1.013393 1.008200
```

Figure 6. Multicollinearity test for stroke PCA model



```
> vif(model_hipertensi_pca)
    FAC1     FAC2     FAC4     FAC5     FAC6     FAC7 
1.000527 1.000805 1.001402 1.002646 1.000711 1.001263 
```

Figure 7. Multicollinearity test for hypertension PCA model



```
> vif(model_diabetes_pca)
    FAC1     FAC2     FAC4     FAC5     FAC7 
1.001446 1.018158 1.016200 1.002703 1.003239 
```

Figure 8. Multicollinearity test for diabetes PCA model

Figure 7, it can be concluded that there are no multicollinearity and thus the regression model for hypertension is valid. The RMSE of the hypertension regression model is 0.02205. For diabetes model, the components FAC_3 and FAC_6 will be removed since its p-value(s) are 0.2089 and 0.4062 which are larger than 0.05 and GLM is conducted again. Figure 5 shows that the remaining components significantly affect the regression model. Multicollinearity test is then conducted, shown in Figure 8 and it can be concluded that there is no multicollinearity and thus the regression model for diabetes is valid. The RMSE of diabetes regression model is 0.0076. From Equations (14)-(16), we can say that FAC_2 affects the stroke and diabetes model the most with coefficient of 0.3042 and 0.4000, respectively. While for hypertension model, FAC_6 gives the most effect with a coefficient of 0.0968. For the stroke model, FAC_4 does not significantly affect the regression equation and thus it was discarded. The variables X_5, X_9, X_{15} took the dominant role in FAC_4 with coefficients of 0.386, 0.317, and 0.343, respectively. This means that the proportion of sweet food consumption, food that contains animal protein consumption, and inactive physical activity does not significantly affect stroke. This result is a little bit different from the article that was written by Johnson et al., which states that physical inactivity and unhealthy diet are the risk factors of stroke [5]. There are two factors that may be the cause of this difference. First, in this research, the variable of food consumption (both sweet food and food that contains animal protein) only considers the frequency but not the amount. Second, this research is specifically based on the data of Indonesian lifestyle as it may differ from the global perspective. For the hypertension model, FAC_3 does not significantly affect the regression equation and thus it was discarded. The variables X_{21} and X_{22} have taken the dominant role in FAC_3 with coefficients of -0.427, and 0.435, respectively. This means that there is no significant difference about the hypertension prevalence by gender. This result is in accordance with the report from WHO which states that behavior and social determinants of health (income) are the risk factors of hypertension [19]. For the diabetes model, FAC_3 and FAC_6 do not significantly affect the regression equation and thus they were discarded. The variables X_{21} and X_{22} have taken the dominant role in FAC_3 with coefficients of -0.427, and 0.435 respectively and the variable X_{10} has taken the dominant role in FAC_6 with coefficient of -0.598. This means that gender and proportion of consumption of food that contains preservatives do not significantly affect diabetes. This result agrees with the global reports from WHO that states healthy diet and regular physical activities are the associated risk factors of diabetes [20]. Income level also plays a major part as the associated risk factor of diabetes.

$$Y_1 = \frac{e^{(-4.9318+0.0501FAC_1+0.3042FAC_2+0.0675FAC_3-0.1068FAC_5+0.0650FAC_6+0.0343FAC_7)}}{1+e^{(-4.9318+0.0501FAC_1+0.3042FAC_2+0.0675FAC_3-0.1068FAC_5+0.0650FAC_6+0.0343FAC_7)}} \tag{14}$$

$$Y_2 = \frac{e^{(-2.4207+0.0240FAC_1+0.0751FAC_2-0.0694FAC_4-0.0597FAC_5+0.0968FAC_6+0.0291FAC_7)}}{1+e^{(-2.4207+0.0240FAC_1+0.0751FAC_2-0.0694FAC_4-0.0597FAC_5+0.0968FAC_6+0.0291FAC_7)}} \tag{15}$$

$$Y_3 = \frac{e^{(-4.3198+0.0686FAC_1+0.4000FAC_2-0.0874FAC_4-0.0410FAC_5-0.0265FAC_7)}}{1+e^{(-4.3198+0.0686FAC_1+0.4000FAC_2-0.0874FAC_4-0.0410FAC_5-0.0265FAC_7)}} \tag{16}$$

5. **Conclusions.** In general, it can be concluded that lifestyle behavior, economic situation, and geographical location can predict stroke, hypertension, and diabetes. Specifically, through GLM from the dimension reduction to these diseases, it is implied that the proportion of sweet food consumption, food that contains animal protein consumption, and inactive physical activity does not significantly affect stroke; there is no significant difference about the hypertension prevalence by gender; the proportion of consumption of food that contains preservatives does not significantly affect diabetes. Furthermore, methods and factors for predicting health conditions can be modified to find an alternative and hopefully better solution in future research.

Acknowledgement. The authors are grateful to the READI (Risk Management, Economic Sustainability and Actuarial Science Development in Indonesia) Project and Universitas Pelita Harapan for fully funding this research.

REFERENCES

[1] World Health Organization, *Noncommunicable Diseases Country Profiles 2018*, Geneva, Licence: CC BY-NC-SA 3.0 IGO, http://www.who.int/nmh/publications/ncd-profiles-2018/en/, accessed on 19 November 2020.

[2] Y. J. Kim and H. Park, Improving prediction of high-cost health care users with medical check-up data, *Big Data*, vol.7, no.3, DOI: 10.1089/big.2018.0096, 2019.

[3] G. Gan and E. A. Valdez, *Actiarial Statistics with R: Theory and Case Studies*, ACTEX Learning, New Hartford, 2018.

[4] I. B. K. Manuaba, I. Sutedja and R. Bahana, The evaluation of supervised classifier models to develop a machine learning API for predicting cardiovascular disease risk, *ICIC Express Letters*, vol.14, no.3, pp.219-226, 2020.

[5] W. Johnson, O. Onuma, M. Owolabi and S. Sachdev, Stroke: A global response is needed, *Bulletin of the World Health Organization*, vol.94, DOI: http://dx.doi.org/10.2471/BLT.16.181636, 2016.

[6] Z. Brokesova, T. Ondruska and E. Pastorakova, Economic and demographic determinants of life insurance industry development, *European Financial System*, pp.61-65, 2015.

[7] D. Minos, I. Butzlaff, K. M. Demmler and R. Rischke, Economic growth, climate change, and obesity, *Curr. Obes. Rep.*, 2016.

[8] M. Masood and D. D. Reidpath, Effect of national wealth on BMI: An analysis of 206266 individuals in 70 low-, middle-, and high-income countries, *PloS One*, vol.12, no.6, 2017.

[9] H. Margaretha, M. Susanto, E. O. Lionel and F. V. Ferdinand, An actuarial model of stroke long term care insurance with obesity as a risk factor, *AIP Conference Proceedings*, 2019.

[10] W. Fan, Epidemiology in diabetes mellitus and cardiovascular disease, *Cardiovascular Endocrinology*, vol.6, no.1, pp.8-16, 2017.

[11] D. Glovaci, W. Fan and N. D. Wong, Epidemiology of diabetes mellitus and cardiovascular disease, *Current Cardiology Reports*, vol.21, no.4, 2019.

[12] X. Su and D. Peng, Emerging functions of adipokines in linking the development of obesity and cardiovascular diseases, *Molecular Biology Reports*, vol.47, pp.7991-8006, 2020.

[13] A. Hannah and S. Johan, Obesity, income and gender: The changing global relationship, *Global Food Security*, vol.23, pp.267-281, 2019.

[14] C. Brooks, *Introductory Econometrics for Finance*, Cambridge University Press, 2019.

[15] R. H. Myers, *Generalized Linear Models: With Applications in Engineering and the Sciences*, Wiley-Blackwell, Oxford, 2010.

[16] A. H. Studenmund and B. K. Johnson, *Using Econometrics a Practical Guide*, Pearson, Boston, 2016.

[17] S. Kolenikov and G. Angeles, The use of discrete data in PCA: Theory, simulations, and applications to socioeconomic indices, *Carolina Population Center MEASURE Evaluation*, University of North Carolina at Chapel Hill, Chapel Hill, NC, 2004.

[18] P. McCullagh and J. Nelder, *Generalized Linear Models*, Springer US, Chicago, 1983.

[19] WHO, *A Global Brief on Hypertension*, MEO Design – Communication – Web, Switzerland, 2013.

[20] WHO, *Global Report on Diabetes*, MEO Design & Communication, France, 2016.