International Journal of Applied Mathematics and Computing

E-ISSN: 3047-146X

Review Article

Identification of Risk Factors for Chronic Kidney Disease Using Binary Logistic Regression

Eva Kosasih[1], Ni Kadek Wulanda Asmara Santhi[2], Ni Wayan Atik Febriyanti[3], Eka Valencia Br Barus[4], Made Susilawati[5]

[1] Department of Mathematics, Udayana University, Bali, Indonesia;
e-mail : kosasih.2208541054@student.unud.ac.id
[2] Department of Mathematics, Udayana University, Bali, Indonesia;
e-mail : santhi.2208541017@student.unud.ac.id
[3] Department of Mathematics, Udayana University, Bali, Indonesia;
e-mail : febriyanti.2208541020@student.unud.ac.id
[4] Department of Mathematics, Udayana University, Bali, Indonesia;
e-mail : barus.2208541015@student.unud.ac.id
[5] Department of Mathematics, Udayana University, Bali, Indonesia;
e-mail : mdsusilawati@unud.ac.id

* Corresponding Author : Eva Kosasih

Abstract: Chronic Kidney Disease (CKD) is a major global health issue that can lead to serious complications and long-term medical care. This study aims to identify key clinical factors associated with CKD status using binary logistic regression analysis. The dataset, obtained from Kaggle, contains 400 patient records with various clinical and demographic attributes. The dependent variable is CKD status (positive or negative), while the independent variables include age, blood pressure, hemoglobin level, urine albumin level, and serum creatinine. Initial analysis involved descriptive statistics and multicollinearity checks, followed by model estimation and evaluation using likelihood ratio and Wald tests. The final model identified four significant predictors: blood pressure, hemoglobin, urine albumin, and serum creatinine. The model achieved a high classification accuracy of 95.50% and an Area Under the ROC Curve (AUC) of 98.78%, indicating excellent predictive performance. These results highlight the importance of these clinical indicators in early CKD detection and support their use in risk assessment models for kidney disease screening

Keywords: Chronic Kidney Disease, Binary Logistic Regression, Likelihood Ratio Test, Wald Test, Classification Accuracy

Received: April, 10th 2025
Revised: April, 28th 2025
Accepted: May, 16th 2025
Online Available: July 14th 2025
Curr. Ver.: July 14th 2025

1. Introduction

Chronic Kidney Disease (CKD) is a global health problem with a steadily increasing prevalence each year. According to data from the Centers for Disease Control and Prevention (CDC, 2022), CKD affects more than 10% of the adult population worldwide, and this number is expected to continue rising due to population aging and the high prevalence of comorbid conditions such as hypertension and diabetes mellitus. CKD is defined as kidney

damage or a reduction in glomerular filtration rate (GFR) to less than 60 mL/min/1.73 m^2 for at least three months, with implications for overall health (Levey & Coresh, 2012; KDIGO, 2013).

CKD is a progressive condition marked by a gradual decline in kidney function until reaching the terminal stage, where the kidneys can no longer perform their excretory and homeostatic functions optimally. According to the Kidney Disease: Improving Global Outcomes (KDIGO, 2013) guidelines, CKD is classified into five stages based on the estimated glomerular filtration rate (eGFR), ranging from mild kidney damage (stage 1) to end-stage renal disease (stage 5). In its advanced stages, CKD can cause various serious complications, including anemia, mineral and bone disorders, metabolic acidosis, and an increased risk of cardiovascular disease and mortality (Levey & Coresh, 2012).

Early detection of CKD plays a crucial role in slowing disease progression and preventing more severe complications. However, because early-stage CKD symptoms are often nonspecific and go unnoticed by patients, diagnosis is frequently made only after significant kidney damage has occurred (KDIGO, 2013). Therefore, there is a need for a data-driven predictive approach that can accurately and promptly identify key risk factors for CKD, enabling interventions before irreversible kidney damage occurs.

Binary logistic regression is one of the most commonly used statistical methods for developing predictive models with a dichotomous dependent variable (e.g., having CKD or not). This method allows the analysis of relationships between the outcome variable and one or more predictor variables, whether categorical or numerical (Hosmer, Lemeshow, & Sturdivant, 2013). In the context of CKD, logistic regression has been widely used to identify significant clinical and demographic predictors of CKD incidence.

Previous studies have identified several risk factors contributing to the incidence of CKD. Sumaili et al. (2009), in a study conducted in the Democratic Republic of Congo, found that hypertension, the use of traditional medicines, and family history were dominant factors associated with CKD. A systematic review by Bello et al. (2017) also indicated that social factors such as education level, economic status, and access to healthcare significantly influenced the occurrence of CKD, particularly in developing countries. In addition, Brück et al. (2020) reported that diabetes mellitus, hypertension, and older age were the most associated factors with CKD in the German population.

Based on this background, this study aims to identify the significant factors influencing the incidence of chronic kidney disease using a binary logistic regression approach. The data used in this study is secondary data from the Kaggle platform, specifically the Chronic Kidney Disease Dataset, which contains 400 patient observations with various clinical and demographic characteristics. The predictive model will be evaluated based on the Area Under the Receiver Operating Characteristic Curve (AUC) and goodness-of-fit tests. The results are expected to contribute to the development of an early detection system for CKD that is both statistically accurate and clinically relevant.

2. Research Methods

This study utilized secondary data from the Chronic Kidney Disease Dataset obtained through the Kaggle platform, which contains 400 patient observations with various clinical and demographic characteristics. The dependent variable in this study is the status of Chronic Kidney Disease (CKD), classified into two categories: 0 for patients without CKD and 1 for patients with CKD. The independent or predictor variables used in this study include five variables: age (X_1), blood pressure (X_2), hemoglobin level (X_3), urine albumin level (X_4), and serum creatinine level (X_5). The definitions and measurement scales of each variable are presented in Table 1 below.

Table 1. Research Variables

No	Symbol	Variable	Scale	Category/Description
1	Y	Chronic Kidney Disease (CKD) Status	Categorical	0 = No CKD; 1 = CKD Positive
2	X_1	Age	Numerical	In years
3	X_2	Blood Pressure	Numerical	In mmHg
4	X_3	Hemoglobin Level	Numerical	In g/dL
5	X_4	Urine Albumin Level	Categorical	0 = None; 1 = Trace; 2 = Mild; 3 = Moderate; 4 = Severe; 5 = Very Severe
6	X_5	Serum Creatinine Level	Numerical	In mg/dL

In this study, data processing and analysis were conducted using the Google Colab platform. The analytical steps are as follows:

1. Conduct descriptive statistics;
2. Identify and address multicollinearity among predictors using the Variance Inflation Factor (VIF), which is calculated using the formula:

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \tag{1}$$

where R_j^2 represents the coefficient of determination from the regression of variable x_j on the other predictor variables. According to Montgomery et al. (2012), a VIF value greater than 5 indicates the presence of high multicollinearity, which should be handled by removing the variable;

3. Build the initial binary logistic regression model using the Maximum Likelihood Estimation (MLE) approach, where the likelihood function is:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \left[\pi(x_i)^{y_i}\left(1 - \pi(x_i)\right)^{1-y_i}\right] \tag{2}$$

with $\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}}}$

4. Perform parameter significance tests, including:
   a) Simultaneous testing using the Likelihood Ratio Test to determine whether all predictor variables simultaneously influence the dependent variable. The test statistic used is:

$$G = -2\ln\left(\frac{L_0}{L_1}\right) \tag{3}$$

where L_0 is the likelihood of the null model and L_1 is the likelihood of the full model. The value of G is compared to the χ^2 distribution table with degrees of freedom equal to the number of predictors;

   b) Partial testing using the Wald Test to examine the individual contribution of each parameter to the model, with the formula:

$$W_j = \left(\frac{\widehat{\beta_J}}{SE(\widehat{\beta_J})}\right)^2 ; j = 1,2,\ldots,\text{k} \tag{2}$$

where $\widehat{\beta_j}$ is the estimated parameter and $SE(\widehat{\beta_j})$ is its standard error. A parameter is considered significant if $|W_j| > Z_{\alpha/2}$ (Gujarati & Porter, 2009);

5. Construct the final binary logistic regression model using only the significant variables based on the Wald test;
6. Evaluate the goodness-of-fit of the final model using the Hosmer-Lemeshow Test. The test statistic is formulated as:

$$G^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{5}$$

where O_{ij} and E_{ij} denote the observed and expected values for each group, respectively. The model is considered a good fit if $G^2 < \chi^2$ (Hosmer et al., 2013);

7. Calculate the Odds Ratio to measure the likelihood of CKD occurrence between two groups with different characteristics. The odds formula is:

$$\text{Odds} = \frac{\pi(x_i)}{1 - \pi(x_i)} \tag{6}$$

An odds ratio greater than 1 indicates an increased risk of CKD, while a value less than 1 indicates a reduced risk (Mariani, Santosa, & Setiawan, 2023);

8. Evaluate model performance using a classification table (confusion matrix) and the ROC curve;
9. Interpret the results.

3. Results And Discussion

The data used in this study were obtained from the Chronic Kidney Disease Dataset available on the Kaggle platform. This dataset contains 400 patient observations with various clinical and demographic characteristics. The dependent variable in this analysis is the status of chronic kidney disease (CKD), while the independent variables consist of age (X_1), blood pressure (X_2), hemoglobin level (X_3), urine albumin level (X_4), and serum creatinine level (X_5).

To understand the basic characteristics of the data, descriptive analysis was performed on numerical and categorical variables. Based on the results of descriptive statistics, patient age ranged from 2 to 90 years with an average of 51 years. Blood pressure ranged from 50 to 180 mmHg with an average of 76.47 mmHg. Hemoglobin levels ranged from 3.1 to 17.80 g/dL with an average value of 12.53 g/dL. Meanwhile, serum creatinine levels varied from 0.40 to 76 mg/dL with an average of 3.07 mg/dL. The summary of descriptive statistics of numerical variables is presented in Table 2 below.

Table 2 Descriptive Statistics of Numerical Variables

Variable	Min	Max	Mean
X_1	2	90	51
X_2	50	180	76.47
X_3	3.1	17.80	12.53
X_5	0.40	76	3.07

In addition to numerical variables, the categorical variable urine albumin was also analyzed to see its distribution toward CKD status. The results showed that all patients with

albumin levels in the categories of "trace", "mild", "moderate", "severe", to "very severe" were diagnosed with CKD (100%). Conversely, in the "none" albumin category, most patients (72.9%) did not experience CKD, while 27.1% (54 out of 199) were CKD positive. This indicates a strong relationship between urine albumin level and CKD incidence. The complete distribution of albumin level toward CKD status is presented in Table 3.

Table 3 Distribution of Albumin Level toward CKD Status

Albumin Category	Chronic Kidney Disease		Total
	Non- CKD	CKD Positive	
None	145	54	199
Trace	0	44	44
Mild	0	42	42
Moderate	0	43	43
Severe	0	24	24
Very Severe	0	1	1

Before building the binary logistic regression model, multicollinearity testing was first conducted to ensure that there was no high correlation among predictor variables. This testing was carried out by calculating the Variance Inflation Factor (VIF) using Equation (1). Based on the calculation results, all variables had VIF values below 5, as presented in Table 4. This indicates that there is no indication of high multicollinearity among the independent variables used.

Table 4 VIF Values

	X_1	X_2	X_3	X_4	X_5
VIF	1.04	1.09	1.64	1.45	1.13

Next, all independent variables were included in the construction of the initial binary logistic regression model using the Maximum Likelihood Estimation (MLE) approach. The likelihood calculation was carried out using Equation (2), and the resulting initial model is as follows:

$$logit\,[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$$
$$= 4{,}7893 + 0{,}0033X_1 + 0{,}7316X_2 - 4{,}0035X_3 + 2{,}9232X_4 + 5{,}5057X_5$$

The estimation results show that the variables blood pressure (X_2), urine albumin (X_4), and serum creatinine (X_5) have a positive effect on a person's probability of suffering from CKD. Conversely, hemoglobin (X_3) has a negative effect, meaning that an increase in hemoglobin levels tends to reduce the risk of CKD. Meanwhile, age (X_1) shows a relatively small contribution to the probability of occurrence. This initial model will be further analyzed to evaluate parameter significance and the overall feasibility of the model.

Simultaneous significance testing was carried out to test the combined influence of all independent variables on the dependent variable in the logistic regression model. According to Hosmer and Lemeshow (2000), overall model significance testing can be conducted using the Likelihood Ratio Test (LRT), which produces the test statistic G. The hypotheses tested are formulated as follows:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_5 = 0$ (no influence of independent variables on the dependent variable)

H_1: at least one $\beta_p \neq 0; \quad p = 1,2,3,4,5$ (at least one independent variable influences the dependent variable)

The result of the Likelihood Ratio Test (LRT) calculation using Equation (3) with the help of Google Colab software shows that the test statistic G = 348.852 is greater than the critical value $\chi^2_{(0,05;5)} = 11.070$, thus H_0 is rejected. This means that simultaneously there is a significant influence of the independent variables on chronic kidney disease status.

Next, a partial test was conducted to evaluate the individual effect of each predictor variable on the dependent variable. This test was conducted using the Wald test statistic, with the following hypotheses:

$H_0: \beta_p = 0; \quad p = 1,2,3,4,5$ (variable X_p has no influence on Y)

$H_1: \beta_p \neq 0; \quad p = 1,2,3,4,5$ (variable X_p influences Y)

The results of the Wald test were obtained using Google Colab software and are shown in Table 5 below:

Table 5 Partial Test Results (Wald Test)

Variable	β	S.E.	Wald	Decision
X_1	0.1464	0.3191	0.4589	Not Significant
X_2	1.0199	0.4170	2.4459	Significant
X_3	-4.2396	0.7808	-5.4295	Significant
X_4	3.1952	1.0357	3.0849	Significant
X_5	7.3664	2.6196	2.8121	Significant

Based on the results in Table 5, the decision is made by comparing the Wald value with the critical value of the standard normal distribution at a 5% significance level, which is $Z_{(0,025)} = 1.96$. It can be seen that variable X_1 (age) has a Wald value less than 1.96, so the decision is to fail to reject H_0. This indicates that the age variable does not have a significant effect on the diagnosis of chronic kidney disease. Conversely, the variables X_2 (blood pressure), X_3 (hemoglobin), X_4 (urine albumin), and X_5 (serum creatinine) have Wald values greater than 1.96, thus H_0 is rejected, which means these four variables have a significant effect on CKD diagnosis status.

Based on the Wald test results, only four variables were found to have a significant effect on CKD status, namely blood pressure (X_2), hemoglobin (X_3), urine albumin (X_4), and serum creatinine (X_5). Therefore, the final binary logistic regression model was constructed using these significant variables, and the resulting final model is as follows:

$$logit\,[\pi(x)] = 5{,}5678 + 1{,}0199X_2 - 4{,}2396X_3 + 3{,}1952X_4 + 7{,}3664X_5$$

This model was then evaluated to measure its fit with the observed data. The test was conducted using the Hosmer–Lemeshow test, where the test statistic value was calculated using Equation (5). The hypotheses tested are:

H_0: The resulting binary logistic regression model fits the data.

H_1: The resulting binary logistic regression model does not fit the data.

Based on the results calculated using Google Colab software, the test statistic value G^2=6.14 is smaller than the critical value $\chi^2_{(0,05;6)}$ =12.59. Therefore, the decision is to fail to reject H_0, which means that the binary logistic regression model obtained fits the data.

Furthermore, to measure the magnitude of the influence of each variable in the model on the probability of CKD occurrence, the Odds Ratio values were calculated using Equation (6). The Odds Ratio results of the four significant variables are shown in Table 6 below:

Table 6 Odds Ratio

Variable	Odds Ratio
X_2	2.8224
X_3	0.0145
X_4	25.5327
X_5	1741.8592

Based on Table 6, the variable blood pressure (X_2) has an Odds Ratio value of 2.8224, meaning that every increase of 1 mmHg in blood pressure will increase the probability of a person having chronic kidney disease by 2.82 times compared to individuals with lower blood pressure. This result supports the findings of Wang et al. (2017) who state that hypertension is one of the main risk factors for CKD.

The variable hemoglobin (X_3) has an Odds Ratio of 0.0145, indicating that every 1 g/dL increase in hemoglobin level actually reduces the probability of having CKD to 0.0145 times. Since the OR < 1, hemoglobin has a negative effect on CKD. This finding is consistent with the study by Stauffer & Fan (2014), which mentions that anemia (low hemoglobin level) is often found in CKD patients and contributes to accelerated kidney damage.

Meanwhile, urine albumin (X_4) has an Odds Ratio of 25.5327, indicating that each increase in one level of urine albumin increases the risk of CKD by 25.53 times. This value shows a very strong predictive power of the albumin variable toward CKD. This result aligns with the findings of Levey et al. (2003) which affirm that albuminuria is a clinically recognized indicator of kidney damage. The serum creatinine variable (X_5) shows the strongest influence with an Odds Ratio of 1741.8592. This means that every 1 mg/dL increase in serum creatinine increases the likelihood of a person developing CKD more than 1700 times. This indicates that serum creatinine is a primary indicator of declining kidney function (KDIGO, 2012).

To evaluate the performance of the model in classifying patients who have CKD and those who do not, testing was carried out using two approaches: accuracy calculation from the confusion matrix and the ROC curve illustration. The confusion matrix for the final model is shown in Figure 1 below:

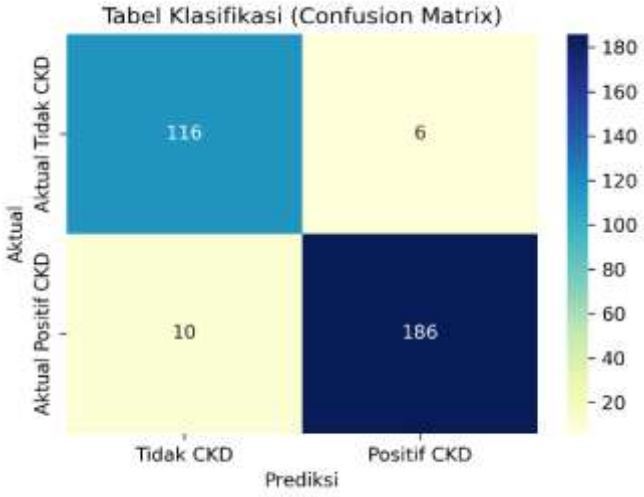

Figure 1 Confusion Matrix of Final Model

Based on the confusion matrix above, the model's accuracy value is calculated using the following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{186 + 116}{186 + 116 + 6 + 10} = \frac{302}{318} = 0.9550$$

Thus, an accuracy value of 95.50% was obtained, indicating that the model has a high level of classification accuracy in predicting CKD status. In addition, the model was also evaluated based on the Receiver Operating Characteristic (ROC) curve, which illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) at various threshold values. The ROC curve for the final model is presented in Figure 2.

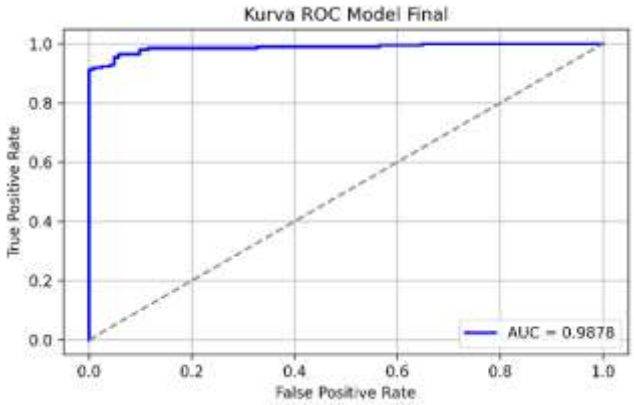

Figure 2 ROC Curve of Final Model

From the ROC curve, an Area Under the Curve (AUC) value of 0.9878 or 98.78% was obtained. This very high AUC value indicates that the model has an excellent discriminative ability in distinguishing between patients who do and do not have CKD. Overall, the high accuracy and AUC values indicate that the constructed binary logistic regression model is not only statistically significant but also highly reliable in clinical practice as a tool for early prediction of CKD.

4. Conclusion And Suggestions

The results of this study indicate that four out of the five independent variables analyzed blood pressure, hemoglobin, urine albumin, and serum creatinine, have a significant influence

on the probability of chronic kidney disease (CKD) occurrence. In contrast, age did not contribute significantly to the model. The final binary logistic regression model is expressed as follows:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = 5{,}5678 + 1{,}0199X_2 - 4{,}2396X_3 + 3{,}1952X_4 + 7{,}3664X_5$$

This model demonstrates a strong ability to distinguish between patients with and without CKD, as reflected by an Area Under the Curve (AUC) value of 98.78% and a classification accuracy of 95.50%. Additionally, all variables in the final model satisfied the assumption of no multicollinearity, and the results of the Hosmer-Lemeshow test indicated a good fit between the model and the observed data. Based on these findings, it is recommended that clinicians consider blood pressure, hemoglobin levels, urine albumin, and serum creatinine as primary indicators in the early screening process for patients at high risk of developing chronic kidney disease (CKD).

References

[1] A. K. Bello *et al.*, "Assessment of global kidney health care status," *JAMA*, vol. 317, no. 18, pp. 1864–1881, 2017, doi: 10.1001/jama.2017.4046.

[2] K. Brück *et al.*, "CKD prevalence varies across the European general population," *J. Am. Soc. Nephrol.*, vol. 31, no. 3, pp. 614–623, 2020, doi: 10.1681/ASN.2019090930.

[3] Centers for Disease Control and Prevention (CDC), *Chronic Kidney Disease in the United States, 2021*. U.S. Department of Health and Human Services, 2022. [Online]. Available: https://www.cdc.gov/kidneydisease/publications-resources/ckd-national-facts.html

[4] D. N. Gujarati and D. C. Porter, *Basic Econometrics*, 5th ed. New York: McGraw-Hill Education, 2009.

[5] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ: Wiley, 2013, doi: 10.1002/9781118548387.

[6] KDIGO, "KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease," *Kidney Int. Suppl.*, vol. 3, no. 1, pp. 1–150, 2012, doi: 10.1038/kisup.2012.73.

[7] A. S. Levey and J. Coresh, "Chronic kidney disease," *Lancet*, vol. 379, no. 9811, pp. 165–180, 2012, doi: 10.1016/S0140-6736(11)60178-5.

[8] A. S. Levey *et al.*, "Definition and classification of chronic kidney disease: A position statement from Kidney Disease: Improving Global Outcomes (KDIGO)," *Kidney Int.*, vol. 67, no. 6, pp. 2089–2100, 2003, doi: 10.1046/j.1523-1755.2003.00409.x.

[9] R. Mariani, B. Santosa, and B. Setiawan, *Statistika untuk Penelitian Kesehatan: Teori dan Aplikasi*. Jakarta: Salemba Medika, 2023.

[10] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Hoboken, NJ: Wiley, 2012, doi: 10.1002/9781118136829.

[11] M. E. Stauffer and T. Fan, "Prevalence of anemia in chronic kidney disease in the United States," *PLoS ONE*, vol. 9, no. 1, e84943, 2014, doi: 10.1371/journal.pone.0084943.

[12] E. K. Sumaili *et al.*, "Prevalence of chronic kidney disease in Kinshasa: Results of a pilot study from the Democratic Republic of Congo," *Nephrol. Dial. Transplant.*, vol. 24, no. 1, pp. 117–122, 2009, doi: 10.1093/ndt/gfn444.

[13] H. Wang *et al.*, "Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: A systematic analysis for the Global Burden of Disease Study 2015," *Lancet*, vol. 388, no. 10053, pp. 1459–1544, 2017, doi: 10.1016/S0140-6736(16)31012-1.