Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 3
BLSME 2024

The Research of Risk Factor Prediction of Chronic Kidney Disease

Yuxi Li [1], Yuzhi Liu [2], Guanyu Lu [3, *]

[1] School of Science, China University of Geosciences Beijing, Beijing, 100083, China

[2] Yangzhou High School of Jiangsu Province, Yangzhou, 225009, China

[3] School of mathematics and statistics, Nanjing University of Information Science and Technology, Nanjing, 211800, China

* Corresponding Author Email: 202013870062@nuist.edu.cn

Abstract. Kidney diseases, particularly chronic kidney disease (CKD), represent a significant health concern, particularly as one ages. This paper delves into an in-depth analysis of CKD, exploring its prevalence, risk factors, and the complex interplay of conventional and nontraditional determinants. Notably, socioeconomic, genetic, and lifestyle factors are pivotal in understanding CKD's multifaceted etiology. The research employs a diverse dataset from Bangladesh, applying statistical techniques including ANOVA, logistic regression, recursive feature elimination, and random forest to identify and evaluate crucial factors associated with CKD. The study aims to construct a robust predictive model for CKD risk assessment, integrating traditional and lesser-explored variables. This comprehensive approach holds promise in refining risk assessment models and guiding targeted intervention strategies for enhanced CKD prevention and management. The findings emphasize the necessity for a holistic understanding of CKD, emphasizing personalized approaches for effective disease management and prevention.

Keywords: Correlation analysis; single-factor ANOVA; binary logistic regression.

1. Introduction

Although with the development of technology, more experienced doctors, and people's attention, the life expectancy of most people worldwide has been increasing, and the mortality rate has been declining. However, under the influence of more and unhealthier diets and irregular lifestyles, more and more people have begun to suffer from chronic diseases, such as cardiovascular diseases, cancer, chronic respiratory diseases, and diabetes [1]. Diseases of the kidneys cannot be ignored, such as chronic kidney disease (CKD), which means a gradual loss of kidney function. There are people in our lives who have suffered from CKD. The older you get, the higher the chances of having developed CKD [2]. At the same time, people's understanding of kidney disease has not been very clear. Most people think that kidney pain or slight discomfort is some minor problem, but if not treated, there may be a high risk of having become a fatal cancer. A large part of this is due to malnutrition and smoking [3]. To achieve this, this paper has first learned a little about the disease, got a general understanding of the problem, and performed reasonable and statistical analysis of some of the data sets this paper has found [4]. Through the analysis and statistics of various data, this paper has carried out technical measurements and systematically collected data from different medical institutions, which have been verified by medical experts. At the same time, due to the high-intensity work pressure and the life of leaving early and returning late, overtime from time to time has invisibly caused damage to our kidneys.

The identification of risk factors has been essential for comprehending the intricate nature of CKD and crafting effective preventive strategies. Prior research has extensively examined various risk factors linked with CKD. Traditional risk factors such as diabetes, hypertension, and obesity have garnered significant attention, emerging as noteworthy contributors to CKD development [5]. Additionally, investigations have explored nontraditional risk factors encompassing nephrotoxic exposure, kidney stones, fetal and maternal influences, infections, environmental elements, and acute kidney injury, all associated with CKD progression [5]. While these studies have provided

comprehensive insights into the multifaceted etiology of CKD, they may not have offered an exhaustive exploration of less conventional risk factors. Furthermore, socioeconomic determinants and genetic predispositions have been implicated in the prevalence and advancement of CKD [6].

Lifestyle factors, including poor dietary habits, tobacco use, and poverty, have also been probed, highlighting the intricate interplay between personal behaviors and CKD development [7]. Geographical disparities in demographics and management approaches have been studied, revealing variances in CKD burden among regions [8]. Notably, specific factors, like socioeconomic status and race-ethnicity, have been underscored in distinct studies [8]. In this context, our research extends the current paradigm by meticulously examining an extensive array of variables, encompassing blood pressure, glucose levels, urine composition, and medical history [5]. By synergizing both traditional and nontraditional factors, our study aims to provide a nuanced and comprehensive comprehension of the CKD risk landscape. This holistic approach not only capitalizes on the strengths of prior research but also introduces a more exhaustive evaluation of the intricate nexus of factors impacting CKD. By dissecting the intricacies of these diverse constituents, our study strives to craft a more precise predictive model for CKD risk assessment. Within the realm of CKD research, the integration of a diverse array of factors, as proposed in our study, bestows a distinctive advantage in capturing the intricate interplay between conventional and nontraditional, genetic, lifestyle, and environmental determinants. This inclusive evaluation holds the potential to refine risk assessment models and inform targeted intervention strategies, thereby significantly contributing to the progression of CKD prevention and management.

Through the relevant literature reviewed, this paper has found that the previous people have used it when researching the relevant content. Analysis of variance (ANOVA) is a statistical method used to test whether the mean of three or more groups differs significantly. It is a common hypothesis testing tool that is widely used in a variety of fields, including finance, biology, engineering, and the social sciences. etc. have been used to select features related to important features in CKD diagnosis. Logistic regression has been used to determine the prediction accuracy of CKD and decision stump. Risk factors for CKD have been calculated using a linear regression model [10]. RFE has mainly involved repeatedly building a model and then selecting the best features, selecting the selected features, and then repeating the process on the remaining features until all features have been traversed [9]. It has assumed that the appearance of all features is independent of each other, that each feature is equally important, and that it is simple and very interpretable. Random forest is a simpler forecasting method that has had an effective way to estimate missing data and maintain accuracy when most data has been lost [11]. Logistic regression has been used for classification, which is a generalized linear regression analysis model that has been beneficial to solving the problem of binary classification [11].

This paper has preliminarily found the relevant features related to important features of CKD diagnosis. Next, the dataset was split into training and test sets. ANOVA and logistic regression were applied to calculate accuracy and other data. Subsequently, the main factors causing kidney disease were identified using a linear regression model.

2. Methods

2.1. Data source

This article uses a dataset from the UC Irvine Machine Learning Repository. The dataset consists of real patient data from Bangladesh. It was collected at the Enam Medical College in Dhaka, Bangladesh. The authors of the dataset are Md. Ashiqul Islam and Shamima Akter.

2.2. Variable selection

The original dataset consists of 27 variables and a predictor variable. Due to the excessive number of variables, a variable selection process is being considered. The objective is to identify variables that

exhibit a substantial degree of correlation with the predictor variable, namely the 'affected' variable, for further investigation. Given that the majority of the variables are discrete rather than continuous, the correlation analysis employs the Kendall's Tau-b method. The correlation tests based on Kendall's Tau-b method were conducted using the SPSS software, yielding a correlation test table. Variables with significance values below 0.001 can be considered as factors significantly correlated with the 'affected' variable, thus holding substantial value for further research endeavors.

Table 1. Correlation Test Table based on Kendall's Tau-b Method

Variable	affected	Variable	affected
affected	0	sc	0.007
age	0.704	pot	0.454
bp_Diastolic	0.197	hemo	<0.001
bp_limit	0.063	pcv	<0.001
sg	<0.001	rbcc	<0.001
al	<0.001	wbcc	0.257
rbc	0.064	htn	<0.001
su	0.004	dm	<0.001
pc	0.004	cad	0.01
pcc	0.013	appet	0.006
ba	0.184	pe	0.008
bgr	<0.001	ane	0.041
bu	<0.001	grf	<0.001
sod	<0.001	stage	<0.001

Based on the aforementioned analysis results, the variables that exhibit notable correlation with chronic kidney disease are tentatively identified as Table 2 shows.

Table 2. Selected variables

CKD dataset	Attributes meaning	Category	Scale
sg	Specific gravity	Nominal	1.005 to 1.025
al	Albumin	Nominal	0 to 5
bgr	Blood glucose random	Numerical	mgs/dl
bu	Blood urea	Numerical	mgs/dl
sod	Sodium	Numerical	mEq/L
hemo	Hemoglobin	Numerical	gms
pcv	Packed cell volume	Numerical	P cv
rbcc	Red blood cell count	Nominal	Millions/cmm
htn	Hypertension	Nominal	No (0), Yes (1)
dm	Diabetes mellitus	Nominal	No (0), Yes (1)
grf	Glomerular FiltrationRate	Numerical	ml/min
stage	Stage	Nominal	1 to 5
affected	Affected	Nominal	No (0), Yes (1)

2.3. Research protocol

In this paper, CKD was the dependent variable (Y), and the 28 factors including dichotomous and multinomial variables were independent variables (X). In the dichotomous variables, 0 means No or the variable data is normal, and 1 means Yes or the variable data is abnormal. In multinomial variables, higher numbers represent higher numeric values for the variable data. Through correlation analysis, the correlation coefficient of each variable and CKD variable was calculated, and 12 variables with high correlation with CKD were screened out by Pearson correlation coefficient, then, factor analysis

(ANOVA) were used respectively to analyze the variables with strong correlation through the method of factor analysis, so as to better understand the potential structure of the variables and better understand the factors affecting CKD. After finding the independent influencing factors and significant factors, this paper calculates the AUC of each independent influencing factor, and compares them with the joint prediction, and plots the evaluation indicators of the model such as the sensitivity and specificity of ROC calculation.

In the etiology of the disease, it is often necessary to analyze the quantitative relationship between the occurrence of the disease and the risk factors, and also need to analyze the confounding effect of multiple factors, and use the logistic regression model to solve the problem better when the dependent variable Y is a dichotomous variable.

The dependent variable y is a dichotomous variable, its value is y=1 (have CKD) or y=0 (without CKD), and the m independent variables that affect the value of y are variables $x_1, x_2, x_3, \cdots, x_m$, and the conditional probability of CKD results occurring under the action of m independent variables p: $p = p(y = 1|x_1, x_2, x_3, \ldots, x_m)$, then the logistic regression model can be represented as:

$$P = \frac{exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots\beta_m x_m)}{1+exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots\beta_m x_m)} \quad (1)$$

β_0 is constant term, $\beta_1, \beta_2, \cdots, \beta_m$ are partial regression coefficients. Perform a logit transformation on $f(x) = \frac{1}{1+e^{-x}}$, then $L(p) = ln\frac{p}{1-p}$. So, the logistic regression model can be expressed in the following linear form:

$$ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1, \beta_2 x_2 \ldots, \beta_m x_m \quad (2)$$

3. Results and discussion

3.1. Descriptive statistical analysis

This paper classified two groups with or without CKD diseases and plot pie or bar charts for each variable to explore the distribution of each variable, and the following is a display of part of the data variables (Figure 1, 2 and 3).

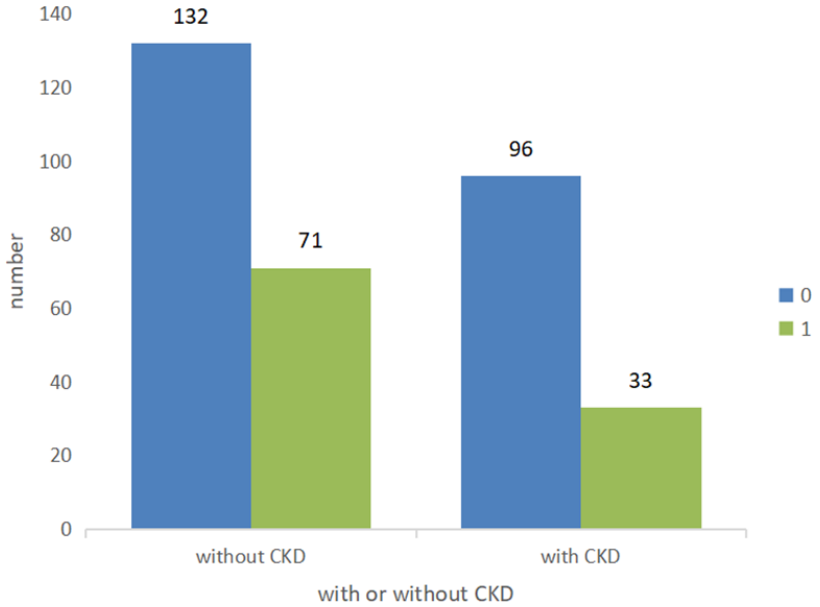

Figure 1. The number of DM with or without CKD.

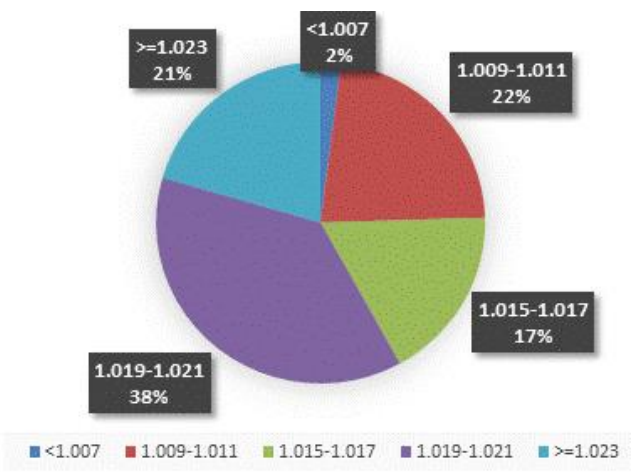

Figure 2. SG values of samples without CKD

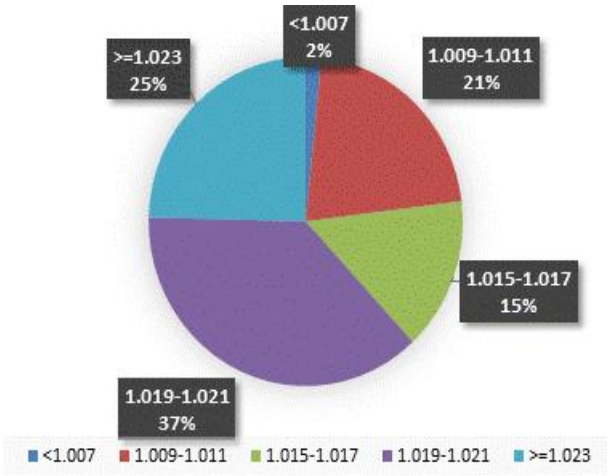

Figure 3. SG values of samples with CKD

3.2. One way ANOVA test results

This paper employs ANOVA for data preprocessing. The rationale behind this is the excessive number of variables, and ANOVA serves the purpose of dimensionality reduction, making it easier for subsequent regression predictions.

First, the variables are standardized. Here is a list of the standardized data. Due to space limitations, only a portion of the data is provided:

Table 3. ANOVA test results

Item	F	χ^2	p
sg	45.365	96.404	0.000**
al	33.505	81.466	0.162
bgr	5.168	39.06	0.000**
bu	10.765	53.638	0.094
sod	12.128	67.373	0.633
hemo	59.067	147.339	0.000**
pcv	42.909	134.049	0.000**
rbcc	26.770	105.717	0.000**
htn	111.197	71.926	0.149
dm	86.028	60.577	0.000**
grf	29.525	116.625	0.000**
stage	69.294	117.404	0.000**

** $p<0.01$

As can be seen from the above Table 3, al, bu, sod, htn all have p greater than 0.05, therefore, this paper excludes these variables, after which this paper performs regression analysis on the remaining variables.

3.3. Binary logistic regression analysis

After the preceding univariate analysis, this article will employ a binary logistic regression model to further investigate factors influencing chronic kidney disease and provide the final conclusion. Accurate regression analysis results can be obtained using SPSS Statistics software. First, the Hosmer-Lemeshow test will be conducted, and if the result is greater than 0.05, it indicates the reliability of the subsequent binary logistic regression results. In such a case, this paper can proceed with further analysis. The results of the Hosmer-Lemeshow test are displayed in the following Table 4.

Table 4. Hosmer-Lemeshow test

Chi-Square	Degrees of Freedom	Significance
6.30E-11	3	0.999

The results of the binary Logistic Regression analysis using SPSS Statistics software are presented in the table below, demonstrating the impact strength of various variables in the regression model. It is important to note that all the data under the variable names in this study are continuous in nature. Therefore, the Exp(B) values (i.e., Odds Ratios) displayed in the table represent the ratio of the outcome variable (where 0 indicates non-afflicted and 1 indicates afflicted) when a given variable increases by one unit compared to its initial value. These values reflect the extent to which a particular variable influences the outcome variable.

From the table 5, it can be confidently inferred that the variables "bgr," "dm," and "stage" have particularly significant effects on the outcome variable.

Table 5. Results of the Binary Logistic Regression

Variations	B	Standard Error	Wald	Significance	Exp(B)
sg	-115.391	1087.85	0.011	0.916	0
bgr	226.907	2472.34	0.008	0.927	3.50E+98
hemo	-33.472	902.132	0.001	0.97	0
pcv	-126.028	1197.61	0.011	0.916	0
rbcc	-96.473	1148.276	0.007	0.933	0
dm	32.002	2243.533	0	0.989	7.91E+13
grf	-26.167	260.498	0.01	0.92	0
stage	57.677	770.203	0.006	0.94	1.12E+25

ROC curves generated using SPSS software revealed that the curves for DM, GRF, and stage all lie above the reference line (the diagonal line), indicating a notably high sensitivity (Figure 4). This observation aligns with the results of the binary logistic regression, providing confirmation of the accuracy of the binary logistic regression results.

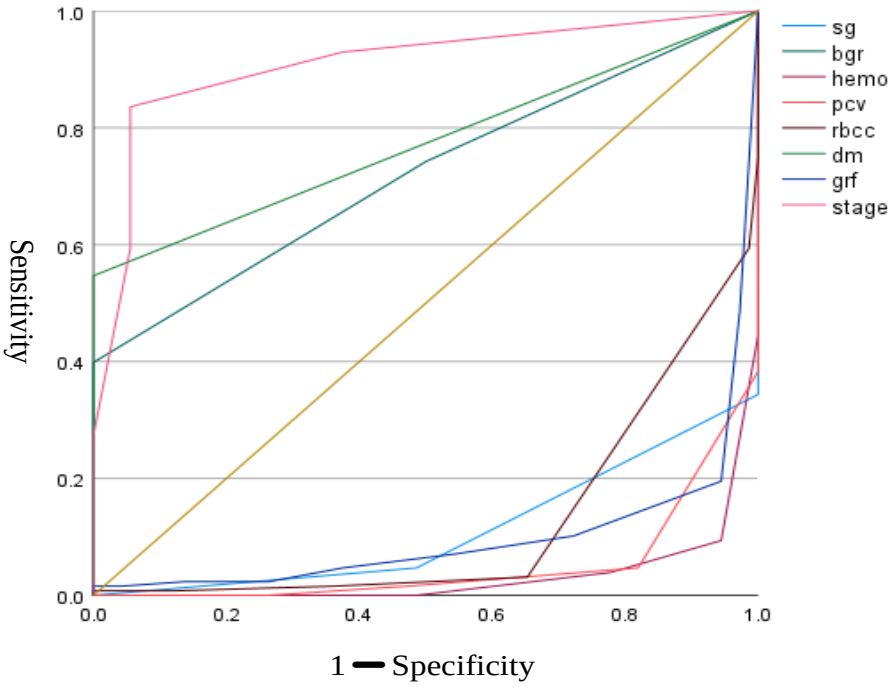

Figure 4. ROC curves

4. Conclusion

In this experiment, ANOVA was used for data preprocessing. The rationale behind this is that there are too many variables, and the purpose of ANOVA is to reduce dimensionality and make subsequent

regression predictions easier. Al, Bu, SOD, HTN were excluded since the p-value was too large, and then this paper performed regression analysis on the remaining variables. Different data emerged from the current study, focusing on factors that may be associated with the development of chronic kidney disease (CKD). The variables "bgr" stands for Blood Glucose Random, "dm" stands for Diabetes Mellitus, and "stage" stands for disease stage, these have a particularly significant effect on the outcome variable. At the same time, it was compared with binary logistic regression and found that the results were consistent. It follows that the values of "bgr", "dm", and "stage" have a significant impact on the prevalence of CKD.

Due to the limited amount of data, the model may have errors in addition to factors, and the sample does not cover all categories, which may cause discrepancies, which may also affect the accuracy of the results. The method used in this study was the SPSS method to visualize the differences in various factors between people with and without CKD. This allows conclusions to be visualized, making the results clearer and more intuitive. It has a certain positive effect on the prevention of CKD. There are many factors associated with CKD that may be worth noting and considering, such as "bgr", "dm", and "stage" factors. Truly preventing CKD requires further medical research, and the research may mean that it will point the way to further related research in the future. Once the predisposing factors are identified, this can help people detect CKD earlier and suppress CKD as soon as possible, improving survival and the patient's quality of life.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Sze S M 1969 Physics of Semiconductor Devices. *New York: Wiley–Inter science.*

[2] Dorman L I 1975 Variations of Galactic Cosmic Rays. *Moscow: Moscow State University Press*, 103.

[3] Szytula A and Leciejewicz J 1989 Handbook on the Physics and Chemistry of Rare Earths. *Amsterdam: Elsevier*, 133.

[4] Kuhn T 1998 Density matrix theory of coherent ultrafast dynamics Theory of Transport Properties of Semiconductor Nanostructures. *Electronic Materials, London: Chapman and Hall*, 173 – 214.

[5] Liu G, et al 2017 Current situation and coping strategies of four major chronic diseases. *Chinese Journal of Social Medicine*, 53 - 56.

[6] Yang L 2022 Risk factors analysis of malnutrition in patients with chronic kidney disease. *Journal of Mathematical Medicine*, 1174 - 1176.

[7] Yan S 2023 Smoking, obesity, high risk of chronic kidney disease. *Health Times*.

[8] Islam M A and Akter S 2023 Risk Factor prediction of chronic kidney disease. *UCI Machine Learning Repository*.

[9] Islam M A, et al. 2020 Risk Factor Prediction of Chronic Kidney Disease based on Machine Learning Algorithms. *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), India,* 952 - 957.

[10] Valerie A L, et al. 2017 Reducing major risk factors for chronic kidney disease. *Kidney International Supplements,* 71 - 87.

[11] Romagnani P, et al 2017 chronic kidney disease. *Nat Rev Dis Primers*, 17088.