International Journal of Environmental Research and Public Health

MDPI

Article

Associations between Sex and Risk Factors for Predicting Chronic Kidney Disease

Hao-Yun Kao [1], Chi-Chang Chang [2,3,*], Chin-Fang Chang [4,5,6,7,*], Ying-Chen Chen[2], Chalong Cheewakriangkrai [8] and Ya-Ling Tu [9]

[1] Department of Healthcare Administration and Medical Informatics, College of Health Sciences, Kaohsiung Medical University, Kaohsiung 80708, Taiwan; haoyun@kmu.edu.tw (H.-Y.K.)
[2] School of Medical Informatics, Chung Shan Medical University & IT Office, Chung Shan Medical University Hospital, Taichung City 40201, Taiwan; threec@csmu.edu.tw (C.-C.C.); amy0988147957@gmail.com (Y.-C.C.)
[3] Department of Information Management, Ming Chuan University, Taoyuan City 33300, Taiwan
[4] Department of Otorhinolaryngology, Head and Neck Surgery, Jen-Ai Hospital, Taichung City 41222, Taiwan; benglung@hotmail.com (C.-F.C.)
[5] Cancer Medicine Center, Jen-Ai Hospital, Taichung City 41222, Taiwan
[6] Basic Medical Education Center, Central Taiwan University of Science and Technology, Taichung City 40601, Taiwan
[7] Department of Medical Education and Research, Jen-Ai Hospital, Taichung City 41222, Taiwan
[8] Department of Obstetrics and Gynecology, Division of Gynecologic Oncology, Faculty of Medicine, Chiang Mai University, Chiang Mai 50200, Thailand; chalong.c@cmu.ac.th (C.C.)
[9] Center for General Education, National Taichung University of Science and Technology, Taichung City 40401, Taiwan; a54sandra@gmail.com (Y.-L.T.)
* Correspondence: changintw@gmail.com (C.-C.C.); benglung@hotmail.com (C.-F.C.)

Abstract: Gender is an important risk factor in predicting chronic kidney disease (CKD); however, it is under-researched. The purpose of this study was to examine whether~~Using~~ gender differences affect the ~~for~~risk factors of early CKD prediction ~~allows for earlier detection and intervention~~. This study used data from 19,270 adult health screenings, including 5101 with CKD, to screen for 11 independent variables selected as risk factors and to test for the significant effects of statistical Chi-square test variables, using seven machine-learning techniques to train the predictive models. Performance indicators included classification accuracy, sensitivity, specificity, and precision. Unbalanced category issues were addressed using three extraction methods: manual sampling, the synthetic minority oversampling technique, and SpreadSubsample. The Chi-square test revealed statistically significant results ($p < 0.001$) for gender, age, red blood cell count in urine, urine protein (PRO) content, and PRO/urinary creatinine ratio. In terms of classifier prediction performance, the manual extraction method, logistic regression, exhibited the highest average prediction accuracy rate (0.8053) for men, whereas the manual extraction method, linear discriminant analysis, demonstrated the highest average prediction accuracy rate (0.8485) for women. The clinical features of a normal or abnormal PRO/urinary creatinine ratio indicated that PRO ratio, age, and urine red blood cell count are the most important risk factors with which to predict CKD in both genders. As a result, this study proposed a prediction model with acceptable prediction accuracy. The model supports doctors in diagnosis and treatment and achieves the goal of early detection and treatment. Based on the academic foundation of evidence-based medicine, this research through practical perspectives and develops an CKD risk prediction model based on machine learning. The model has proven to support as much as possible the prediction of early clinical risk of CKD to improve the efficacy and quality of clinical decision-making.~~The national health and disease prevention strategies focus on immediate risk assessment for CKD and appropriate community-based primary care. Given the difficulty of early-stage detection of CKD, the prevention of end-stage kidney disease deterioration through early prediction, diagnosis, and treatment is necessary. This study developed a risk prediction model based on adult health-screening data, which can robustly predict clinical risk for early CKD in different genders.~~

Citation: Kao, H.-Y.; Chang, C.-C.; Chang, C.-F.; Chen, Y.-C.; Cheewakriangkrai, C.; Tu, Y.-L. Associations Between Sex and Risk Factors for Predicting Chronic Kidney Disease. *Int. J. Environ. Res. Public e next* **2021**, *17*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s):

Received: date
Accepted: date
Published: date

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Keywords: gender differences; chronic kidney disease; machine learning; risk factors

1. Introduction

Chronic kidney disease (CKD) is a global public health problem that is related to severe morbidity, mortality, and high medical resource utilization [1]. In 2017, the global estimated prevalence of CKD is 9.1% with totaled 69.75 million cases and caused 1.2 million deaths.~~In 2017, global CKD cases totaled 69.75 million with 1.2 million deaths~~ [2]. The global estimated prevalence of CKD is 9.1%. According to Taiwan's Ministry of Health and Welfare, CKD accounts for the largest proportion of health insurance claims in Taiwan, with a total of 364,000 patients accounting for costs of approximately NT$51.3 billion in 2018 [3]. Population aging and the associated increases in hypertension continue to raise the prevalence of hyperlipidemia and hyperglycemia, and thus the incidence of CKD. For the unobvious symptoms and signs at early stage,~~Early CKD has no obvious symptoms~~. Generally, two basic definitions of CKD are used: 1. Kidney damage for >3 months, including structural and functional abnormalities, which may be pathological, blood, urine, or imaging abnormalities; and 2. Glomerular filtration rate (GFR) <60 mL/min/1.73 m² for over 3 months. CKD is usually divided into five stages based on GFR [4] (see Table 1). However, the adverse consequences and clinical risks of renal function insufficiency in early cases~~CKD~~ differ between genders.

Table 1. Five stages of chronic kidney disease.

Stages	Description	GFR Value
1	CKD with normal or high GFR	≥90 mL/min/1.73 m²
2	Mild CKD	60–89.9 mL/min/1.73 m²
3	Moderate CKD	30–59.9 mL/min/1.73 m²
	3a	45–59.9 mL/min/1.73 m²
	3b	30–44.9 mL/min/1.73 m²
4	Severe CKD	15–29.9 mL/min/1.73 m²
5	End stage CKD	<15 mL/min/1.73 m²

GFR: Glomerular Filtration Rate; 3a: Stage 3a of kidney disease; 3b: Stage 3b of kidney disease.

The current literature mostly focuses on differences in kidney disease prognoses for men and women of average age. In men, urinary albumin excretion, plasma glucose, and systolic blood pressure are the most important predictors of severely declined renal function [5,6]. In addition, waist circumference and cholesterol/high-density lipoprotein ratio are positively associated with the maintenance of kidney function. Relative to women, plasma glucose and systolic blood pressure are risk predictors of decreased renal function, and triglycerides are positively related to maintaining renal function [5,6]. Related epidemiological studies show that the incidence of CKD in women is lower than in men [7,8]. In addition, estrogen was shown to protect kidney function in related animal models, including glomerular sclerosis reduction and ischemic damage prevention [9,10]. Ricardo et al. (2019) [11] found a significant difference in eGFR decline between men and women (−1.09 mL/min/1.73 m² for women and −1.43 mL/min/1.73 m² for men ($p < 0.001$)).

However, only few researches~~little research attention is~~ focused on gender differences among adults [12]. Current screening standards ~~for CKD~~ provide measurements but hard to~~are unable to~~ effectively predict CKD precisely [13]. At present, at least 2 million patients are diagnosed with CKD in Taiwan, but only 3.5% can secure an early diagnosis for timely treatment, lead to the loss~~and patients often lose~~ 25% of kidney function before diagnosis. Thus, early detection helps to effectively prevent deterioration. Evidence-based medicine requires gender heterogeneity to be taken into account in CKD deterioration to

格式化: 靠左
格式化: 靠左
格式化表格
格式化: 靠左
格式化: 靠左
格式化: 靠左
格式化: 靠左
格式化: 字型: (英文)Palatino Linotype, (中文) 標楷體, 10 點, 字型色彩: 自動

inform the risk assessment, monitoring, and prognosis. A recent survey showed that Taiwan has a high incidence of CKD, but public awareness of the condition is extremely low. According to the 2012 report of the United States Preventive Services Task Force and the American College of Physicians, kidney function~~CKD screening~~ is insufficient for asymptomatic individuals, and the effective tools for ~~effective~~ CKD detection~~screening~~ are lacking [14]. Regardless of the risk factors, the American Society of Nephrology strongly recommends regular CKD detection~~screenings~~ [14].

This study used preventive health-screening data from Taiwanese adults to assess the impact of gender differences as risk factors for CKD in order to achieve early prediction of reduced renal function. Since 2012~~Beginning in 2012~~, Taiwan implemented a 5-year plan for CKD prevention and quality care improvement to reduce the need for dialysis after kidney transplantation and improve the 5-year survival rate. However, in 2017, Taiwan reported 275,000 cases of CKD, and 6743 of these resulted in deaths. The uncertainty of diagnosis usually results form h ~~H~~eterogeneity screening and clinical practices~~ for CKD produce unclear results~~; thus, an accurate tool is needed for early prediction to ensure that potential patients receive and comply with preventive health check-ups. Data mining has been successfully used to build predictive models for healthcare prediction tasks [15–~~17~~21]. The present study sought to evaluate the novel hypothesis that men and women with CKD possess different risk factors. The objective of this study is to develop a risk prediction model among gender by using seven machine-learning algorithms to predict early CKD.~~Therefore, in this study, seven machine-learning algorithms were used to predict early CKD, seeking to identify CKD risk factors through the analysis of large sets of adult health-screening data; thus, allowing early and effective intervention.~~

2. Materials

2.1. Data Source

~~Previous studies revealed that gender is associated with CKD deterioration [18–21]. The urine protein (PRO)/urine creatinine ratio (UPCR) is the most important risk factor for CKD deterioration [22–24]. The UPCR is used for the next section of analyzing the impact of gender differences to predict the risk factors for CKD. The research design is as follows:~~

First, adult health-screening data were collected from 32 health-screening clinics and three specialized laboratories; this included 19,270 screening results that were recorded between 1 January 2015 and 31 December 2019. Previous studies revealed that gender is associated with CKD deterioration [22–24]. The urine protein (PRO)/urine creatinine ratio (UPCR) is the most important risk factor this consequence [25–27]. The UPCR is used for the next section of analyzing the impact of gender differences to predict the risk factors for CKD. Important risk factors of CKD were analyzed through discussions with clinical experts and a review of the relevant literature. The data were cleaned, re-encoded, ranked in importance using the gain ratio and information gain classifier, and then divided into training and test sets (at a 7:3 ratio) through 10 rounds of random sampling. Gender prediction and a clinical risk factor analysis were then performed using seven machine-learning classifiers. On the basis of expert input and the relevant literature, 11 independent predictive variables were selected (see Table 2).

Table 2. Data sources and variable codes.

Variables	Name	Normal Range
X1	Gender	1 male/2 female
X2	Age	Continuous
X3	~~Red Blood Cell,~~ RBC	0–5
X4	~~Glucose,~~ GLU	70–100
X5	~~Triglycerides,~~ TG	50–150

格式化: 字型: (英文)Palatino Linotype, (中文) 標楷體, 字型色彩: 自動

格式化: 縮排: 第一行: 0.02 公分

格式化: 靠左

格式化: 靠左

格式化: 靠左

格式化: 靠左

格式化: 靠左

格式化: 靠左

X6	~~Total Cholesterol,~~ T-CHO	50–200
X7	~~High-density Cholesterol,~~ HDL	>40
X8	~~Low-density Cholesterol,~~ LDL	<130
X9	~~Albumin,~~ ALB	3.5–5.0
X10	~~Proteinuria,~~ PRO	Random < 12 mg/dl
X11	~~Urine protein/creatinine ratio,~~ UPCR	<150
Y	~~Estimated Glomerular filtration rate,~~ eGFR	1.< 90 mL/min/1.73 2. ≥90 mL/min/1.73 m^2

RBC: Red Blood Cell; GLU: Glucose; TG: Triglycerides; T-CHO: Total Cholesterol; HDL: High-density Cholesterol; LDL: Low-density Cholesterol; ALB: Albumin; PRO: Proteinuria; UPCR: Urine protein/creatinine ratio; eGFR: Estimated Glomerular filtration rate

2.2. Method

Seven machine-learning techniques were used to predict early CKD, including support vector machine (SVM), linear discriminant analysis (LDA), logistic regression (LGR), C4.5 decision tree (C4.5), classification and regression tree (CART), random forest (RF), and C5.0 decision tree (C5. 0) [15–17].

CART is a decision tree algorithm that uses a binary process to sequentially divide the data space and generate a simple prediction rule in each partition. CART not only solves classification problems but also performs a regression analysis, which means that this method ~~CART~~ generates a classification tree when the target of the predictor variable is clear. The first step in the CART analysis is to construct all decision rules through a binary decomposition process; the second step is to prune the overgrown trees to eliminate unnecessary rule trees; and the last step is to determine the best tree rules using cross-validation.

C4.5 is a decision tree algorithm that uses the ID3 algorithm to produce an improved iterative binary tree. It selects the attributes of the decision tree at each node according to the information acquisition rate ~~[20]~~. According to information theory and probability statistics, the smallest is the expected information value, the greatest is the information gain, and the highest stands the information entropy (i.e., purity). C4.5 reflects the training set and gradually develops a tree structure for attribute variables according to the maximum information gain ratio.

LDA is a widely used method for dimensionality reduction and classification ~~[21]~~. For the data to be analyzed, LDA looks for different categories with their own dimensional space in which the distance between samples of different categories gradually increases, and, contrarily, the distance between samples of the same category decreases. In the learning process, LDA can obtain the function to project different types of samples into low-dimensional space and then performs feature decomposition to calculate the best projection.

The C5.0 algorithm can generate decision trees or rule sets. The C5.0 model splits the samples based on the maximum information gain. The sample subset determined by the first split is then split again, usually based on another field, and this process is repeated to cause the sample subset to split until it cannot be split any further. Finally, the lowest-level split is re-captured, and the sample subsets that do not significantly contribute to the model value are pruned.

RF is an overall classification method based on statistical learning theory that combines several separate classification trees ~~[25]~~. RF is a supervised machine-learning algorithm that considers the unweighted majority of classified votes. It first uses bootstrapping to select various random variables as the training dataset. This widely used algorithm uses random sampling and replacement to simultaneously reduce variance and avoid overfitting. The classification tree of the selected sample is then used to construct the training process, using a large number of classification trees to form the RF from selected samples. CART is a classification method that is widely used for RF modeling. Finally, all

格式化: 靠左
格式化: 靠左
格式化: 靠左
格式化: 靠左
格式化: 靠左
格式化: 靠左
格式化: 靠左
格式化: 字型: (英文)Palatino Linotype, (中文) 標楷體, 字型色彩: 自動
格式化: 縮排: 第一行: 0.02 公分

classification trees are merged, with voting for each category, and then the winning category is selected according to the number of votes to obtain the final classification result.

SVM is a machine-learning algorithm that is based on the principle of structural risk minimization and is used to estimate the function by minimizing the upper limit of the generalization error ~~[26]~~. SVM modeling can initially use a linear or nonlinear kernel function to map the input vector to a feature space. Then, in the feature space, SVM tries to find the best linear division to construct a hyperplane that separates various types.

LGR is a classification algorithm rather than a regression algorithm. Usually, the known independent variable is used to predict the value of a discrete dependent variable (for example, a binary value of 0/1, yes/no, and true/false). It predicts the probability of an event by fitting a logit function, thus producing a probability value between 0 and 1.

Regarding to the imbalance issue, we adopted the kind of sampling techniques: manual extraction technique, synthetic minority over-sampling technique (SMOTE) and SpreadSubsample. Manual extraction is an easy extraction technique, and it under-sampled the majority category data at random. The synthetic minority over-sampling technique (SMOTE) is an over-sampled minority category data by creating synthetic samples, and it makes extra data by actual examples. SpreadSubsample is an under-sampling and de-creasing majority category data technique, and it can be balanced with the minority class. Three types of extraction techniques were used to solve majority and minority data imbalanced problems.

Using the above classification methods, we have used the Raprt package (version R4.1.15) to build the CART prediction model. The OptimClassifier suite (version R0.1.5) was used to determine the tree depth, the number of observations in the terminal node, and the pruning tree parameters to search for the best parameter set to generate the CART model [~~19~~22]. The RWeka suite (version R0.4-42) was used to construct the C4.5 model. The caret suite (version R 6.0-84) was used to identify the best parameter set to effectively build the C4.5 model ~~[27]~~. LDA used the MASS suite (version R7.3-51.5). The ELM model was constructed by running the elmNN package in the R1.0 version [~~28~~23]. The caret suite (version R6.0-84) was used to adjust important hyperparameters to find the best number of hidden layer neurons to generate the best ELM model [~~27~~]. Classification accuracy was assessed based on accuracy, sensitivity, specificity, and the receiver operating characteristic curve by estimating the area under the curve (AUC).

格式化: 字型: 非粗體, 字型色彩: 自動, (中文) 中文 (台灣)

格式化: 字型: (中文) 標楷體

3. Results

In this study, we collected 19,270 valid health-screening records from 32 health-screening clinics and three specialized laboratories. Of the 8073 male and 11197 female records, ~~2465~~ 5101 had CKD patients and 14169 healthy samples~~, whereas of the 11,197 female records, 2636 had CKD~~. The average age of CKD patients was 69.19, range 58.45-79.93 years, shows in table 3. This study used the adult health-screening data to predict risk factors for CKD. The results of the Chi-square statistical test are summarized in Table 3. Gender, age, urinary red blood cells (RBC), fasting blood glucose (GLU), neutral lipid (triglyceride) (TG), high-density lipoprotein (HDL), PRO, and urinary protein/creatinine ratio (UPCR) were all found to be statistically significant. In terms of gender, the majority of male participants were found to have CKD (48.3% CKD vs. 39.6% healthy). The proportion of participants with abnormal RBC in the CKD group was higher than that in the healthy participants group (23.2% vs. 19.1%). The GLU in the CKD group was higher than in the healthy participants group (20.7% vs. 18.8%). Compared with the healthy group, the CKD group had a higher proportion of TG abnormalities (60.6% vs. 58.5%). In CKD group, the proportion of PRO abnormalities was higher than in healthy participants group (82.1% vs. 35.0%), as was the UPCR (67.9% vs. 12.7%). No significant differences were found for serum total cholesterol (p = 0.491), low-density lipoprotein (p = 0.782), or albumin (p = 0.457).

Table 3. Descriptive analysis of variables.

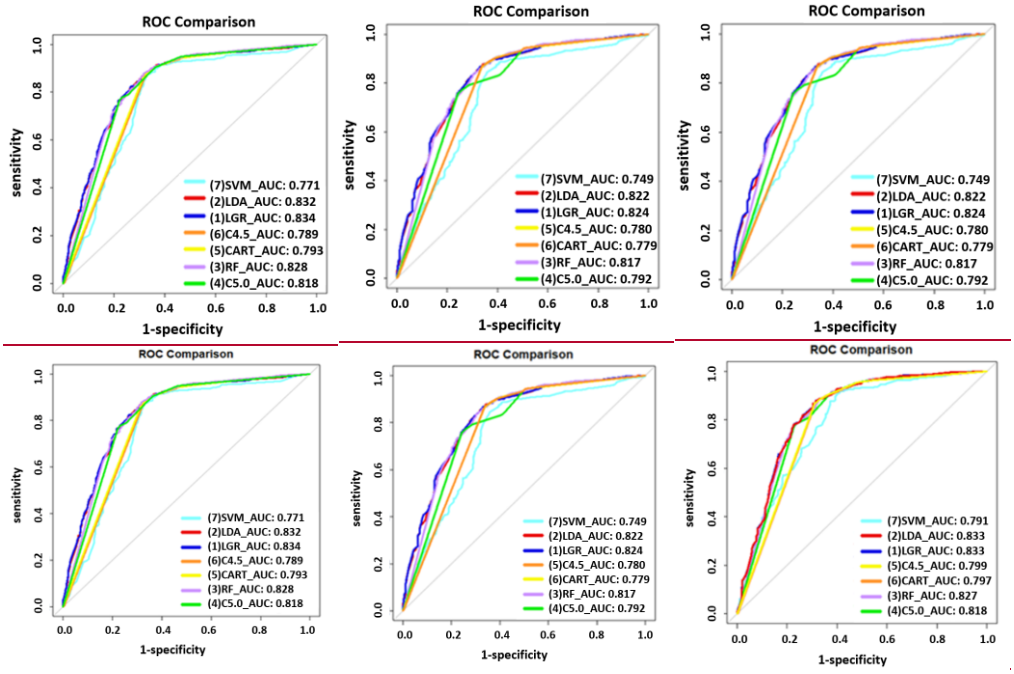
Items	Healthy	CKD	p-value	χ^2
n (%)	14,169(73.5%)	5,101(26.5%)		
Gender				
Male	5,608(39.6%)	2,465(48.3%)	<0.001 **	117.817
Female	8,561(60.4%)	2,636(51.7%)		
Age				
Mean (±SD)	63.37 ± 11.56	69.19 ± 10.74	<0.001 *	699.271
RBC				
Normal	11,460(80.9%)	3,917(76.8%)	<0.001 **	38.956
Abnormal	2,709(19.1%)	1,184(23.2%)		
GLU				
Normal	2667(18.8%)	1,055(20.7%)	0.004 **	8.321
Abnormal	11,502(81.2%)	4,046(79.3%)		
TG				
Normal	5,878(41.5%)	2,012(39.4%)	0.011 *	6.466
Abnormal	8,291(58.5%)	3,089(60.6%)		
T-CHO				
Normal	9,198(64.9%)	3,284(64.4%)	0.491	0.474
Abnormal	4,971(35.1%)	1,817(35.6%)		
HDL				
Normal	11,954(84.4%)	4,369(85.6%)	0.029 *	4.763
Abnormal	2,215(15.6%)	732(14.4%)		
HDL				
Normal	11,400(80.5%)	4,095(80.3%)	0.782	0.076
Abnormal	2,769(19.5%)	1,006(19.7%)		
ALB				
Normal	14,162(100.0%)	5,097(99.9%)	0.457	0.553
Abnormal	7(0.0%)	4(0.1%)		
PRO				
Normal	9,203(65.0%)	915(17.9%)	<0.001 *	3,324.451
Abnormal	4,966(35.0%)	4,186(82.1%)		
UPCR				
Normal	12,364(87.3%)	1,639(32.1%)	<0.001 *	5,739.411

Abnormal	1,805(12.7%)	3,462(67.9%)

** *p*-value < 0.01; * *p*-value < 0.05.

RBC: Red Blood Cell; GLU: Glucose; TG: Triglycerides; T-CHO: Total Cholesterol; HDL: High-density Cholesterol; LDL: Low-density Cholesterol; ALB: Albumin; PRO: Proteinuria; UPCR: Urine protein/creatinine ratio; eGFR: Estimated Glomerular filtration rate

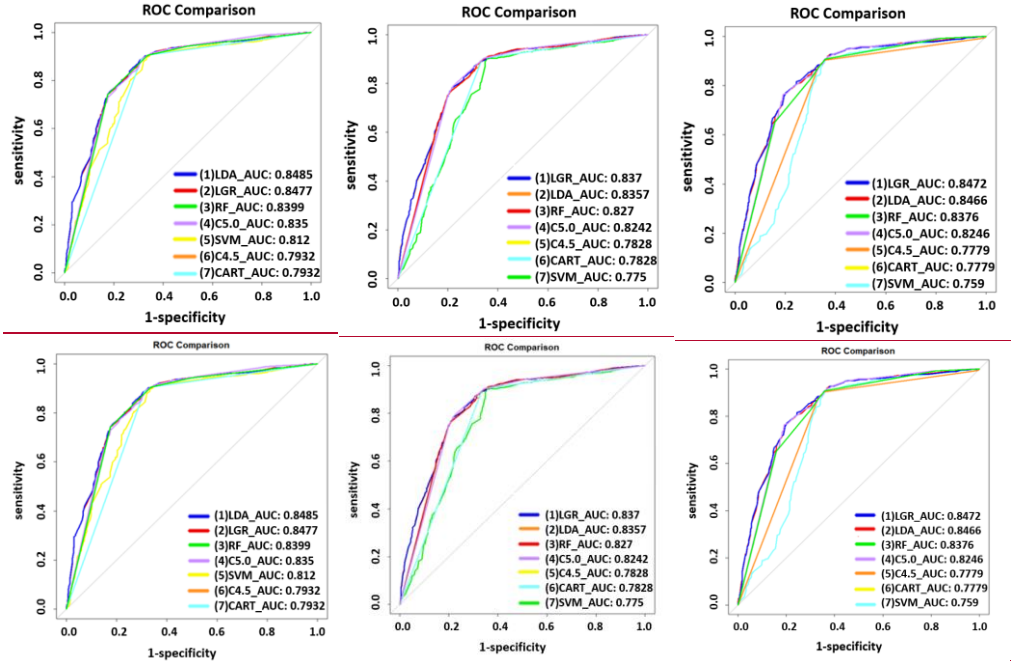
On the basis of manual extraction techniques, the statistically significant risk factors in the male participants group were age, RBC, GLU, HDL, PRO, and UPCR, with LGR exhibiting the highest AUC (0.834) (Figure 1a); whereas in the female participants group, they were age, RBC, T-CHO, PRO, and UPCR, with LDA exhibiting the highest AUC (0.8485) (Figure 2a). Common risk factors for both genders were age, RBC, PRO, and UPCR. Using the synthetic minority oversampling technique (SMOTE) extraction analysis, the statistically significant risk factors in the male participants group were age, RBC, GLU, HDL, PRO, and UPCR, with LGR exhibiting the highest AUC (0.824) (Figure 1b); whereas in the female participants group, they were age, RBC, T-CHO, HDL, PRO, and UPCR, with LGR exhibiting the highest AUC (0.837) (Figure 2b). The common risk factors for both genders were age, RBC, HDL, PRO, and UPCR. Using SpreadSubsample extraction analysis techniques, the statistically significant risk factors in the male participants group were age, RBC, GLU, HDL, PRO, and UPCR, with LGR exhibiting the highest AUC (0.833) (Figure 1c); whereas in the female participants group, they were age, RBC, T-CHO, PRO, and UPCR, with LGR exhibiting the highest AUC (0.8472) (Figure 2c). The common statistically significant risk factors for both genders were age, RBC, PRO, and UPCR. Regardless of the extraction method, the top three risk factors for deteriorating CKD in male participants were UPCR, PRO, and age, whereas in females, they were UPCR, PRO, and age.



格式化: 字型: 10 點
格式化: 字型: 11 點
格式化: 字型色彩: 黑色
格式化: MDPI_3.1_text, 靠左, 縮排: 左: 0 公分, 右: 0 公分

(**a**) The ROC curve of manual extraction for males. (**b**) The ROC curve of SMOTE extraction for males. (**c**) The ROC curve of SpreadSubsample extraction for males.

Figure 1. The Receiver operating characteristic (ROC) ~~ROC~~ curves for comparing extraction methods for male.



(**a**) The ROC curve of manual extraction for females. (**b**) The ROC curve of SMOTE extraction for females. (**c**) The ROC curve of SpreadSubsample extraction for females.

Figure 2. The Receiver operating characteristic (ROC)~~ROC~~ curves for comparing extraction methods for female.

UPCR was further used to divide the following stage, which was used to analyze gender differences and predict CKD risk factors based on UPCR status. Different sampling method results are presented below.

Through manual extraction analysis techniques, the statistically significant risk factors for normal UPCR in male participants were age, RBC, TG, HDL, and PRO, with LGR exhibiting the highest AUC (0.715); whereas in female participants, they were age, RBC, GLU, TG, HDL, and PRO, with LGR exhibiting the highest AUC (0.7039). Common risk factors for both genders were age, RBC, TG, HDL, and PRO. Using SMOTE extraction analysis, the statistically significant risk factors for normal UPCR in male participants were age, RBC, TG, T-CHO, HDL, and PRO, with LDA exhibiting the highest AUC (0.693); whereas in females, they were age, RBC, GLU, TG, T-CHO, HDL, LDL, and PRO, with LDA exhibiting the highest AUC (0.7061). Common risk factors for UPCR in both genders were age, RBC, TG, T-CHO, HDL, and PRO. Using SpreadSubsample extraction analysis techniques, the statistically significant risk factors of normal UPCR in male participants were age, RBC, TG, and PRO, with LDA exhibiting the highest AUC (0.657); whereas in females, they were age, RBC, GLU, TG, and PRO, with LDA exhibiting the highest AUC (0.7142). Common statistically significant risk factors for UPCR in both genders were age, RBC, TG, and PRO. Regardless of the extraction method, the top three risk factors for CKD

deterioration under normal UPCR in male participants were PRO, RBC, and age, whereas for females, they were PRO, age, and RBC.

The statistically significant risk factors for abnormal UPCR in male participants included age, RBC, GLU, HDL, and PRO, with LDA exhibiting the highest AUC (0.741); whereas in females, they were age, RBC, GLU, GLU, T-CHO, and PRO, with LGR exhibiting the highest AUC (0.7318). Common risk factors for UPCR abnormalities in both genders were age, RBC, GLU, and PRO. Our findings were consistent with some previous reports, including the National Health Research Institutes Annual Report on Kidney Disease in UPCR [24] and the RBC [25]. The findings of the ALB and GLU are consistent with previous our studies [18]. Similarly, Kshirsagar et al. [26] reported that age, and gender are important risk factors for prediction of chronic kidney disease.

格式化: 非醒目提示

格式化: 非醒目提示

格式化: 非醒目提示

格式化: 非醒目提示

4. Discussion

The key analyzed risk factor in this study was gender. ~~The key risk factor analyzed in this study was gender.~~ The results are consistent with previous findings in that it was demonstrated that gender is very important to predict CKD [9–11,~~29~~].

~~Other risk factors included UPC, age, RBC, GLU, TG, T-CHO, HDL-C, LDL-C, ALB, and PRO.~~

The data analysis results showed that the most important risk factor for CKD prediction in both genders was UPCR, followed by PRO and age. Regardless of normal or abnormal UPCR in male participants, the top three risk factors for CKD were PRO, age, and RBC, and the same results were observed in the females analyzed with UPCR.~~obtained for the females with normal and abnormal UPCR~~.

Other risk factors included UPC, age, RBC, GLU, TG, T-CHO, HDL-C, LDL-C, ALB, and PRO. From the above, UPCR, PRO, age, and RBC are important risk factors in this gender difference analysis [28]. This result is consistent with previous reports, including research reports on the role of UPCR and RBC in CKD from the National Institutes of Health [24]. The results for PRO are consistent with [29-31], whereas the findings for ALB and GLU are consistent with previous studies [32-34].

Other studies found that the prevalence of CKD is often positively correlated with diabetes, hypertension, and obesity [35-37], which is also consistent with the results of the present study.

~~From the above, UPCR, PRO, age, and RBC are important risk factors in this gender difference analysis. This result is consistent with previous reports, including research reports on the role of UPCR and RBC in CKD from the National Institutes of Health [30].~~

~~The results for PRO are consistent with Matsushita et al. [31] and Wu [32], whereas the findings for ALB and GLU are consistent with previous studies [29,33]. Other studies found that the prevalence of CKD is often positively correlated with diabetes, hypertension, and obesity [30], which is also consistent with the results of the present study.~~

This study sought to set-up a precise prediction ~~produce predictions~~ from routine health screenings from asymptomatic adult patients. A comprehensive clinical prevention method that considers all these factors is needed to successfully solve the problem of high-risk exposure in the adult population and to produce a successful prevention~~comprehensive prevention method~~ specifically for these populations. As shown in Figure 3, through decision tree analyses, all samples can pass through 12 subsets of different branches, from the root node to the leaf node, by conditional screening. As mentioned earlier, gender has a strong influence on CKD interpretation; , and~~thus, it~~ was selected as the root node of the classification decision tree. The second-order decision tree node is UPCR. The third-order decision tree is generated by age, RBC, TG, PRO, T-CHO, and LDL, with classification prediction accuracy ranging from 55.2% to 84.9%. Table 4 presents the results of the 12 combinations of conditions.

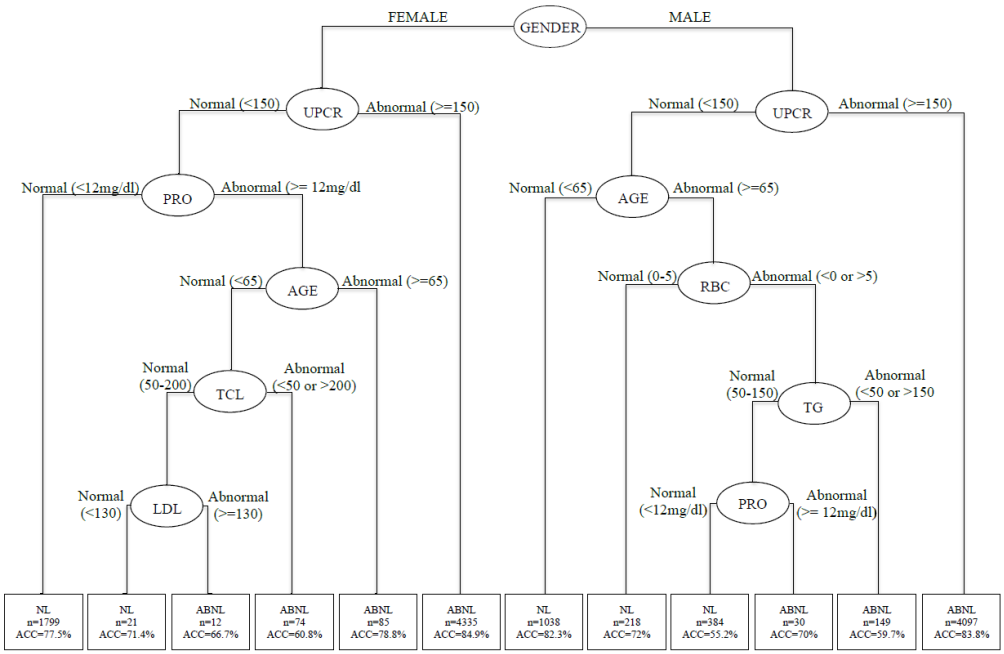


Figure 3. Decision tree for predicting variables by gender.

RBC: Red Blood Cell; GLU: Glucose; TG: Triglycerides; T-CHO: Total Cholesterol; HDL: High-density Cholesterol; LDL: Low-density Cholesterol; ALB: Albumin; PRO: Proteinuria; UPCR: Urine protein/creatinine ratio; eGFR: Estimated Glomerular filtration rate

Table 4. Classification and regression tree (CART)~~CART~~ decision rule for predicting variables by gender.

Rule No.	The Composition of Risk Factors	No.	Status	Accuracy
1	Gender (Female) + UPCR(<150) + PRO(<12)	1799	Non-CKD	77.5%
2	Gender (Female) + UPCR(<150) + PRO(≥ 12) + Age(<65) + T-CHO(50–200) + LDL(<130)	21	Non-CKD	71.4%
3	Gender (Female) + UPCR(<150) + PRO(≥ 12)+Age(<65) + T-CHO(50–200)+LDL(≥ 130)	12	CKD	66.7%
4	Gender (Female) + UPCR(<150)+PRO≥ 12) + Age(<65) + T-CHO(<50 or >200)	74	CKD	60.8%
5	Gender (Female) + UPCR(<150) + PRO(≥ 12) + Age(≥ 65)	85	CKD	78.8%
6	Gender (Female) + UPCR(≥ 150)	4335	CKD	84.9%
7	Gender (Male) + UPCR(<150) + Age(<65)	1038	Non-CKD	82.3%
8	Gender (Male) + UPCR(<150) + Age(≥ 65) + RBC(0–5)	218	Non-CKD	72%
9	Gender (Male) + UPCR(<150) + Age(≥ 65)+RBC(<0 or >5) + TG(50–150) + PRO(<12)	384	Non-CKD	55.2%
10	Gender (Male) + UPCR(<150) + Age(≥ 65) + RBC(<0 or >5) + TG(50–150) + PRO(≥ 12)	30	CKD	70%

格式化: 靠左
格式化: 靠右
格式化: 靠左
格式化: 靠右
格式化: 靠左
格式化: 靠右
格式化: 靠左
格式化: 靠右
格式化: 靠左
格式化: 靠右
格式化: 靠左
格式化: 靠右
格式化: 靠左
格式化: 靠右
格式化: 靠左
格式化: 靠右
格式化: 靠左
格式化: 靠右
格式化: 靠左
格式化: 靠右

11	Gender (Male) + UPCR(<150) + Age(≥ 65) + RBC(<0 or >5) + TG(<50 or >200)	149	CKD	59.7%
12	Gender (Male) + UPCR(≥ 150)	4097	CKD	83.8%

RBC: Red Blood Cell; GLU: Glucose; TG: Triglycerides; T-CHO: Total Cholesterol; HDL: High-density Cholesterol; LDL: Low-density Cholesterol; ALB: Albumin; PRO: Proteinuria; UPCR: Urine protein/creatinine ratio; eGFR: Estimated Glomerular filtration rate

格式化: 靠左
格式化: 靠右
格式化: 靠左
格式化: 靠右

5. Conclusions

CKD is a major global public health issue which~~, and~~ significant differences are reported in terms of research, health care, and public policy around the world. According to a global report from the US Kidney Data Logging System, ~~Taiwan has~~ the world's highest prevalence of dialysis is in Taiwan. In addition, according to an annual report from Taiwan's Ministry of Health and Welfare, CKD accounts for the largest proportion of health insurance claims and the largest single cost, with a total 364,000 patients requiring treatment in 2018. At present, only 10% of patients with CKD in Taiwan are aware of their condition, indicating low public awareness, possibly due to a lack of screening evidence in asymptomatic cases, and a lack of effective screening tools.

This study used a large dataset of adult health-screening results to calculate the statistical significance of variables and their Chi-square tests. Using seven classification techniques written in R language software, we identified important risk factors for gender differences in predicting CKD. The results indicate that, regardless of gender, health screenings should emphasize the UPCR, PRO, and age. For male UPCR, attention should be paid to PRO, age, and urine RBC. For female UPCR, attention should be paid to PRO, age, and urine RBC. This research sought to empirically establish the risk factors of CKD based on health-screening data from Taiwanese adults and use machine-learning methods to identify the model with the highest prediction accuracy; thus, allowing clinicians to select the appropriate classifier for early-stage detection.

Further investigation of prediction need to be conducted after CKD diagnosis, and follow-up is necessary to track patients.Effective prediction of patient risk of developing CKD allows for early detection and treatment. Machine learning-based risk prediction models can provide supportive evidence for the robustness of early clinical risk assessments and CKD prediction. Previous studies often showed imbalances in data categories; we recommend that these experiments be repeated using advanced methods and discussion are required.

Author Contributions: Data curation, C.-F.C. and C.-C.C.; Investigation, C.-C.C. and H.-Y.K.; Methodology, C.-C.C. and Y.-C.C.; Validation, C.-C.C. and Y.-C.C.; Writing—original draft, H.-Y.K., Y.-L.T. and C. C.; Writing—review and editing, C.-C.C., H.-Y.K. and C.-F.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nordqvist, C. Chronic Kidney Disease: Causes, Symptoms and Treatments. Last updated: Thu, 14. 2016.
2. Smith, D. Chronic Kidney Disease: A Global Crisis. 2018. Available online: https://www.siemens-healthineers.com/en-be/news/chronic-kidney-disease.html (March 9, 2018).
3. National Health Research Institutes Annual Report on Kidney Disease in Taiwan. Available online: http://w3.nhri.org.tw/nhri_org/rl/lib/NewWeb/nhri/ebook/39000000448683.pdf (accessed on 27 September 2021).
4. Official Journal of the International Society of Nephrology KDIGO Clinical Practice Guideline for the Management of Blood Pressure in Chronic Kidney Disease. Available online: https://kdigo.org/wp-content/uploads/2016/10/KDIGO-2012-Blood-Pressure-Guideline-English.pdf (accessed on 27 May 2020).
5. Hippisley-Cox, J.; Coupland, C. Predicting the risk of chronic Kidney Disease in men and women in England and Wales: Prospective derivation and external validation of the QKidney Scores. *BMC Fam. Pract.* **2010**, *11*, 49.

6. Halbesma, N.; Brantsma, A.H.; Bakker, S.J.; Jansen, D.F.; Stolk, R.P.; De Zeeuw, D.; De Jong, P.E.; Gansevoort, R.T.; for the PREVEND study group Gender differences in predictors of the decline of renal function in the general population. *Kidney Int.* **2008**, *74*, 505–512.
7. Go, A.S.; Chertow, G.M.; Fan, D.; McCulloch, C.E.; Hsu, C.-Y. Chronic Kidney Disease and the Risks of Death, Cardiovascular Events, and Hospitalization. *N. Engl. J. Med.* **2004**, *351*, 1296–1305.
8. Albertus, P.; Morgenstern, H.; Robinson, B.; Saran, R. Risk of ESRD in the United States. *Am. J. Kidney Dis.* **2016**, *68*, 862–872.
9. Hutchens, M.P.; Fujiyoshi, T.; Komers, R.; Herson, P.S.; Anderson, S. Estrogen protects renal endothelial barrier function from ischemia-reperfusion in vitro and in vivo. *Am. J. Physiol. Physiol.* **2012**, *303*, F377–F385.
10. Elliot, S.J.; Karl, M.; Berho, M.; Potier, M.; Zheng, F.; Leclercq, B.; Striker, G.E.; Striker, L.J. Estrogen Deficiency Accelerates Progression of Glomerulosclerosis in Susceptible Mice. *Am. J. Pathol.* **2003**, *162*, 1441–1448.
11. Ricardo, A.C.; Yang, W.; Sha, D.; Appel, L.J.; Chen, J.; Krousel-Wood, M.; Manoharan, A.; Steigerwalt, S.; Wright, J.; Rahman, M.; et al. Sex-Related Disparities in CKD Progression. *J. Am. Soc. Nephrol.* **2019**, *30*, 137–146.
12. Kattah, A.G.; Garovic, V.D. Understanding sex differences in progression and prognosis of chronic kidney disease. *Ann. Transl. Med.* **2020**, *8*, 897–897.
13. Roberti J, Cummings A, Myall M, Harvey J, Lippiett K, Hunt K, Cicora F, Alonso JP, May CR. Work of being an adult patient with chronic kidney disease: a systematic review of qualitative studies. *BMJ open.* ***2018***, 8(9), e023507
14. Moyer, V.A. U.S. Preventive Services Task Force. Screening for chronic kidney disease: U.S. Preventive Services Task Force recommendation statement. *Ann. Intern. Med.* **2012**, *157*, 567–570.
15. Tseng, C.-J.; Lu, C.-J.; Chang, C.-C.; Chen, G.-D.; Cheewakriangkrai, C. Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. *Artif. Intell. Med.* **2017**, *78*, 47–54.
16. Ting, W.C.; Lu YC, A.; Lu, C.J.; Cheewakriangkrai, C.; Chang, C.C. Recurrence impact of primary site and pathologic stage in patients diagnosed with colorectal cancer. *Journal of Quality.* **2018**, *25*, 166–184.
17. Chang, C.C.; Chen, S.H. Developing a novel machine learning-based classification scheme for predicting SPCs in breast cancer survivors. *Front. Genet.* **2019**, *10*, 848.
18. Shih, C-C; Chen, S-H; Chen, G-D; Chang, C-C; Shih, Y-L. Development of a Longitudinal Diagnosis and Prognosis in Patients with Chronic Kidney Disease: Intelligent Clinical Decision-making Scheme. *Int J Environ Res Public Health.* **2021**, 18, 12807.
19. Chang, C-C; Huang, T-H; Shueng, P-W; Chen, S-H; Chen, C-C; Lu, C-J; Tseng, Y-J. Developing a Stacked Ensemble-based Classification Scheme to Predict Second Primary Cancers in Head and Neck Cancer Survivors. *Int J Environ Res Public Health*. **2021**, 18, 12499.
20. Chang, C-C; Chen, C-C; Cheewakriangkrai, C; Chen, Y-C; Yang, S-F. Risk Prediction of Second Primary Endometrial Cancer in Obese Women: A Hospital-Based Cancer Registry Study. *Int J Environ Res Public Health.* **2021**, 18, 8997.
21. Chan, C-L; Chang, C-C. Big Data, Decision Models, and Public Health. *Int J Environ Res Public Health.* **2020**, 17, 6723.
22. Grubinger, T.; Zeileis, A.; Pfeiffer, K.P. evtree: Evolutionary learning of globally optimal classification and regression trees in R.J. Stat. Softw. **2014**, 61, 1–29.
23. Hornik, K.; Buchta, C.; Zeileis, A. Open-source machine learning: R meets Weka. Comput. Stat. **2009**, 24, 225–232.
24. National Health Research Institutes Annual Report on Kidney Disease in Taiwan. Available online: http://w3.nhri.org.tw/nhri_org/rl/lib/NewWeb/nhri/ebook/39000000448683.pdf (accessed on 27 Dec 2021).
25. The National Health Insurance Statistics. 2017. Available online: https://www.nhi.gov.tw/english/Content_List.aspx?n=0D39BCF70F478274&topn=616B97F8DF2C3614 (accessed on 27 Dec 2021).
26. Korbut, A.I.; Klimontov, V.V.; Vinogradov, I.V.; Romanov, V.V. Risk factors and urinary biomarkers of non-albuminuric and albuminuric chronic kidney disease in patients with type 2 diabetes. *World J. Diabetes* **2019**, 10, 517–533.
~~17.~~ Haroun, M.K.; Jaar, B.G.; Hoffman, S.C.; Comstock, G.W.; Klag, M.J.; Coresh, J. Risk factors for chronic kidney disease: A prospective study of 23,534 men and women in Washington County, Maryland. *J. Am. Soc. Nephrol.* **2003**, *14*, 2934–2941.
~~18. Iseki, K.; Iseki, C.; Ikemiya, Y.; Fukiyama, K. Risk of developing end-stage renal disease in a cohort of mass screening. *Kidney Int.* **1996**, *49*, 800–805.~~
~~19. Weller, J.M.; Wu, S.-C.H.; Ferguson, W.; Hawthorne, V.M. End-Stage Renal Disease in Michigan. *Am. J. Nephrol.* **1985**, *5*, 84–95.~~
~~20.~~ Neugarten, J.; Golestaneh, L. Influence of Sex on the Progression of Chronic Kidney Disease. *Mayo Clin. Proc.* **2019**, *94*, 1339–1356.
~~21.~~ Price, C.P.; Newall, R.G.; Boyd, J.C. Use of Protein: Creatinine Ratio Measurements on Random Urine Samples for Prediction of Significant Proteinuria: A Systematic Review. *Clin. Chem.* **2005**, *51*, 1577–1586.
~~22.~~ Hogg, R.J.; Furth, S.; Lemley, K.V.; Portman, R.; Schwartz, G.J.; Coresh, J.; Levey, A.S. National Kidney Foundation's Kidney Disease Outcomes Quality Initiative clinical practice guidelines for chronic kidney disease in children and adolescents: Evaluation, classification, and stratification. *Pediatrics* **2003**, *111*, 1416–1421.
~~23.~~ Abebe, J.; Eigbefoh, J.; Isabu, P.; Okogbenin, S.; Eifediyi, R.; Okusanya, B. Accuracy of urine dipsticks, 2-h and 12-h urine collections for protein measurement as compared with the 24-h collection. *J. Obstet. Gynaecol.* **2008**, *28*, 496–500.
~~24. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees;* Wadsworth, Champman &Hall/CRC: New York, NY, USA, 1984.~~
~~25. Vapnik, V.N. *The Nature of Statistical Learning Theory;* Springer: Berlin/Heidelberg, Germany, 2000.~~

格式化: 字型: (英文)Palatino Linotype, (中文) +本文中文字型 (新細明體)

格式化: 字型: (英文)Palatino Linotype, (中文) +本文中文字型 (新細明體)

格式化: 字型: (英文)Palatino Linotype, (中文) +本文中文字型 (新細明體)

格式化: 字型: (中文) +本文中文字型 (新細明體), 字型色彩: 黑色

格式化: 字型: (中文) +本文中文字型 (新細明體), 字型色彩: 黑色

格式化: 字型: (中文) +本文中文字型 (新細明體), 字型色彩: 黑色

格式化: 縮排: 左: 0.75 公分, 取消項目符號與編號

27.32. Kim, M.; Kim, C.S.; Bae, E.H.; Ma, S.K.; Kim, S.W. Risk factors for peptic ulcer disease in patients with end-stage renal disease receiving dialysis. *Kidney Res. Clin. Pr.* **2019**, *38*, 81–89.
28. Muntner, P.; Coresh, J.; Smith, J.C.; Eckfeldt, J.; Klag, M.J. Plasma lipids and risk of developing renal dysfunction: The Atherosclerosis Risk in Communities Study. *Kidney Int.* **2000**, *58*, 293–301.
29.33. Tangri, N.; Stevens, L.A.; Griffith, J.; Tighiouart, H.; Djurdjev, O.; Naimark, D.; Levin, A.; Levey, A.S. A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure. *JAMA* **2011**, *305*, 1553–1559.
30.34. Brück, K.; Stel, V.S.; Gambaro, G.; Hallan, S.; Völzke, H.; Ärnlöv, J.; Kastarinen, M.; Guessous, I.; Vinhas, J.; Stengel, B.; et al. CKD Prevalence Varies across the European General Population. *J. Am. Soc. Nephrol.* **2016**, *27*, 2135–2147.
31.35. Matsushita, K.; Van der Velde, M.; Astor, B.C.; Woodward, M.; Levey, A.S.; De Jong, P.E.; Gansevoort, R.T. Chronic Kidney Disease Prognosis Consortium: Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: A collaborative meta-analysis. *Lancet* **2010**, *375*, 2073–2081.
32.36. Wu, M.T.; Lam, K.K.; Lee, W.C.; Hsu, K.T.; Wu, C.H.; Cheng, B.C.; Lee, C.T. Albuminuria, proteinuria, and urinary albumin to protein ratio in chronic kidney disease. *J. Clin. Lab. Anal.* **2012**, *26*, 82–92.
33.37. Van Der Velde, M.; Matsushita, K.; Coresh, J.; Astor, B.C.; Woodward, M.; Levey, A.S.; de Jong, P.E.; Coresh, J.; Chronic Kidney Disease Prognosis Consortium. Lower estimated glomerular filtration rate and higher albuminuria are associated with all-cause and cardiovascular mortality. A collaborative meta-analysis of high-risk population cohorts. *Kidney Int.* **2011**, *79*, 1341–1352.