scientific reports

OPEN

Use of consensus clustering to identify distinct subtypes of chronic kidney disease and associated mortality risk

Yi Qin[1,6], Liping Xuan[2,6], Zhe Wu[1], Yujie Deng[4], Bin Liu[5] & Shujie Wang[3,4]✉

Background Chronic kidney disease (CKD) is a complex condition with diverse etiology and outcomes. Utilizing a data-driven clustering approach holds promise in identifying distinct CKD subgroups associated with specific risk profiles for death.

Methods Unsupervised consensus clustering was utilized to classify chronic kidney disease (CKD) into subtypes based on 45 baseline characteristics in a cohort of 6,526 participants from the US National Health and Nutrition Examination Survey (NHANES) spanning the years 1999–2000 to 2017–2018. We examined the associations between CKD subgroups and clinical endpoints related to mortality, including all-cause mortality, cardiovascular disease mortality, cancer mortality, and mortality due to other causes.

Results A total of 6,526 individuals with CKD were classified into four clusters at baseline. Cluster 1 ($n=508$) comprised patients with relatively favorable levels of cardiac and kidney function markers, lower prevalence of cancer and higher prevalence of obesity, lower medication usage, and younger age. Cluster 4 ($n=2,029$) comprised patients with the worst cardiac and kidney function markers. The characteristics of cluster 2 ($n=1,439$) and 3 ($n=2,550$) fell in between these two clusters. From cluster 1 to cluster 4, we observed a gradual increase in the hazard ratios of all-cause mortality, cardiovascular disease mortality, and mortality due to other causes. Additionally, further sensitivity analysis revealed patient heterogeneity among predefined subgroups with similar baseline kidney function and mortality risks.

Conclusions Consensus clustering integrated baseline clinical and laboratory measures, revealing distinct CKD subgroups with markedly different risks of death, suggesting that further examination of patient subgroups could advance precision medicine.

Keywords Chronic kidney disease, Clustering approach, Mortality

Background

Chronic kidney disease (CKD) involves kidney damage or a reduction in function, typically defined by an eGFR less than 60 mL/min per 1·73 m² or markers of kidney damage like albuminuria[1]. CKD poses a significant and escalating global burden, with an estimated 10% of adults worldwide affected by this condition. CKD is independently associated with an elevated risk of all-cause mortality, cardiovascular mortality, and progression to end-stage renal disease[2–4]. It leads to 1.2 million deaths and 28.0 million years of life lost annually[5,6].

CKD is a complex disorder with a wide range of causes, including systemic illnesses and conditions like hypertension, diabetes, autoimmune diseases, genetic predisposition, and congenital abnormalities[7–10]. Other factors, such as inflammation, exposure to toxins in the environment, and certain medications, can also contribute to the development of CKD[11–15]. The heterogeneity and complexity of CKD pose challenges to its control and management. The dilemma between health benefits from the intervention and economic consideration in terms

[1]Department of Thoracic Surgery, The Affiliated Hospital of Qingdao University, Qingdao, China. [2]Department of Endocrinology, Guangdong Provincial Geriatrics Institute, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China. [3]Department of Geriatric Medicine, The Affiliated Hospital of Qingdao University, Jiangsu Road No.19, Qingdao, China. [4]Department of Endocrinology and Metabolism, The Affiliated Hospital of Qingdao University, Qingdao, China. [5]Department of Rheumatology, The Affiliated Hospital of Qingdao University, Qingdao, China. [6]Yi Qin and Liping Xuan contributed equally to this work. ✉email: 18765917610@163.com

of cost-effectiveness calls for re-classification of CKD to enable precise and effective intervention in those at the greatest risk of mortality. In the present study, we hypothesize that distinct subpopulations within individuals with CKD can be identified through the use of multidimensional phenotypic data by performing the consensus clustering analysis[16,17]. It is further hypothesized that these subpopulations will have varying risks of future death. We aimed to explore the heterogeneity among CKD patients, and further investigated the differences in mortality risks between clusters.

Results

Study population

Totally, 6,526 individuals with CKD were included in the final analysis. The mean age of study population was 60.3 years, 50.5% of the subjects were women and 58.0% were non-white. The mean eGFR was 95.1 ± 24.2 ml/min per 1.73 m^2 and the median UACR was 72.5 mg/g (interquartile range, 42.0–181.0 mg/g). The total person-months of follow-up were 591,634 person-months.

Determination of cluster number

Figure 1 shows the matrix heatmaps of the pairwise consensus for each number of cluster analysis (Fig. 1. a-f). The cumulative distribution functions (CDFs) (Fig. 2a) and the proportion increase of the area under the CDFs (Fig. 2b) indicated that to category patients into 4 clusters could best display the profiles of participants with CKD. For 4 clusters, the mean consensus scores were larger than 0.96 for all clusters, with a larger number suggesting better stability of cluster analysis (Fig. 2c).

Distributions of characteristics by clusters

Demographic, anthropometric, behavioral and clinical data for the four clusters are shown in Table 1. The distribution of the 45 baseline variables exhibited statistically significant differences between the four clusters,

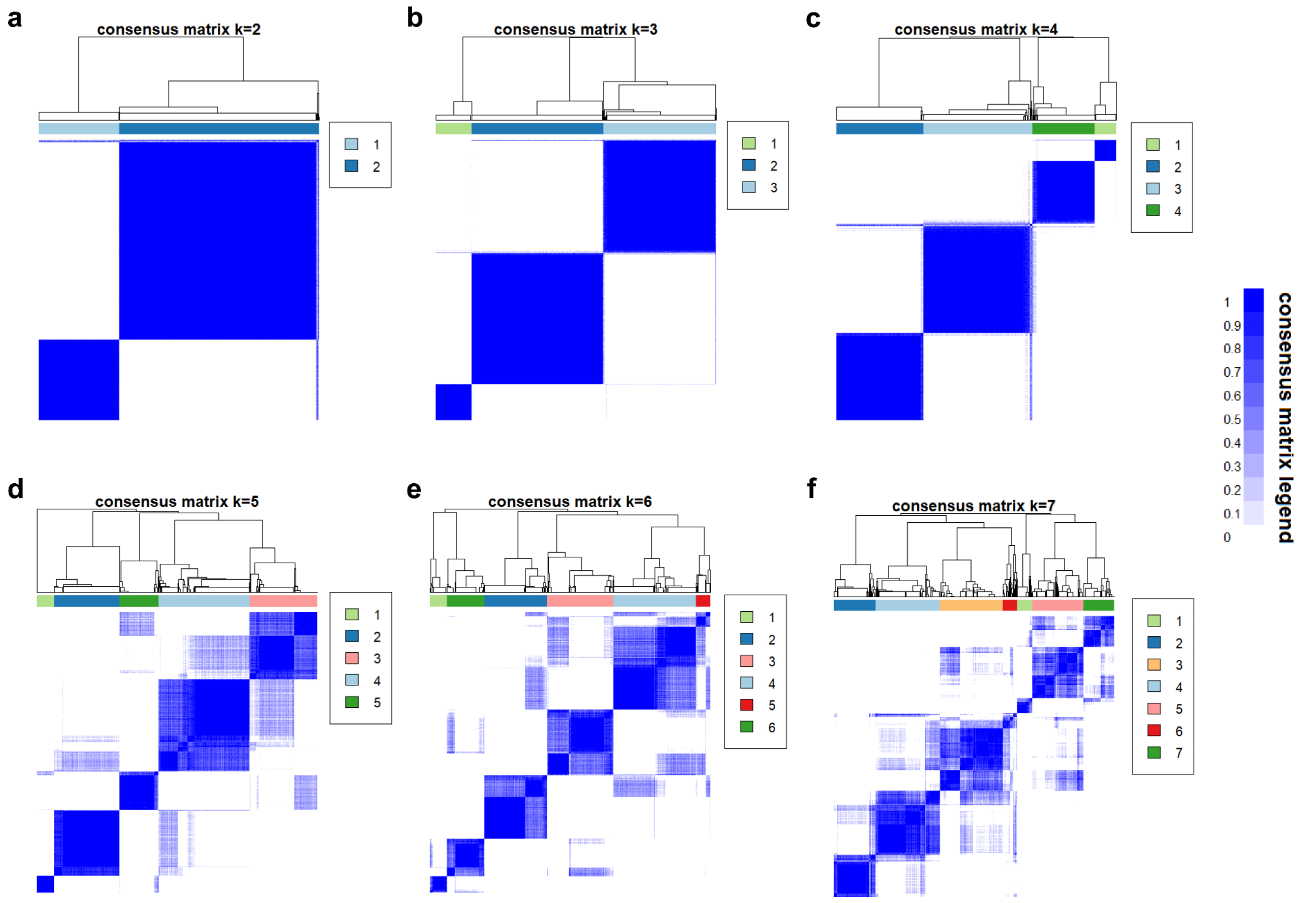

Fig. 1. Consensus matrix heatmaps using behavioral risk factors, metabolic risk factors, diseases history, and social determinants.

The consensus matrix heat maps of K = 2 to K = 7 using 45 variables. The brightest blue color in the diagonals represents perfect consensus where two individuals always group together, the white color represents perfect consensus where two individuals always group separately, and the blue color scales in between represent ambiguous consensus where two individuals are grouped together in some runs but separately in others. **(a)** K = 2; **(b)** K = 3; **(c)** K = 4; **(d)** K = 5; **(e)** K = 6; **(f)** K = 7.

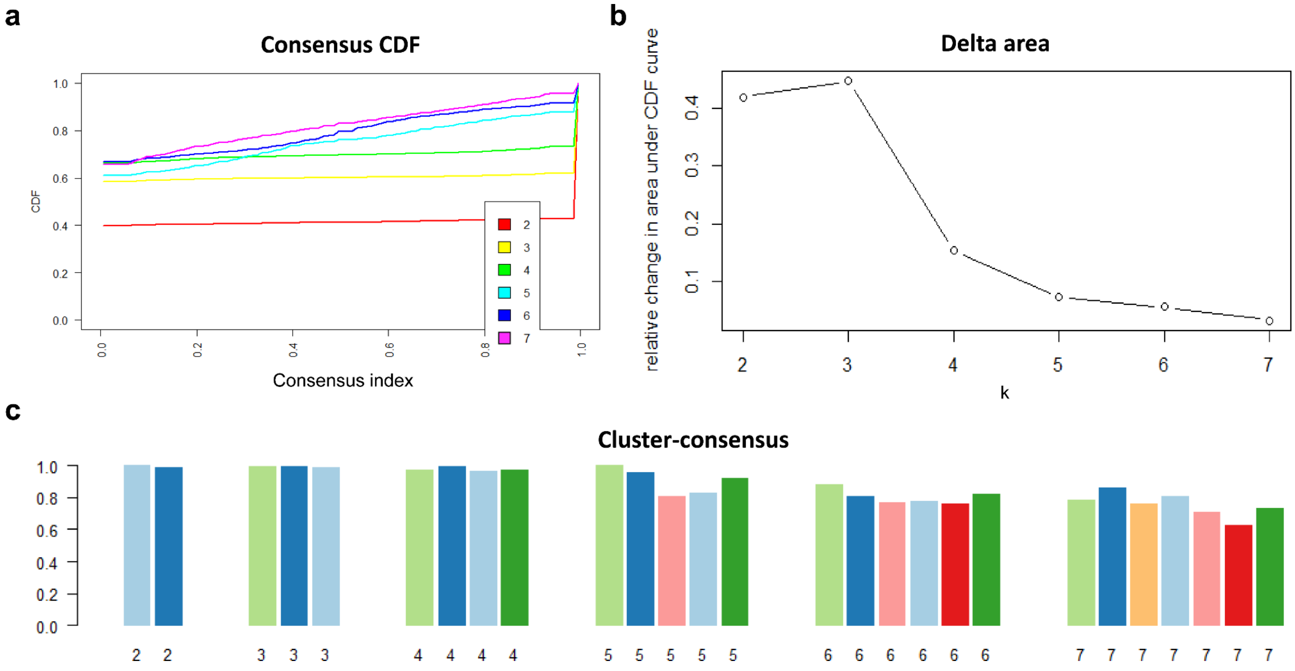

Fig. 2. Consensus cumulative distribution function and cluster consensus score to determine at what number of clusters.
(a) The lines by colors indicating the cumulative distribution functions (CDF) of the consensus matrix for each number of clusters. The CDF reaches an approximate maximum, consensus and cluster confidence is at a maximum at this K. **(b)** The changes in area under the CDF curves comparing K and K − 1. For K = 2, there is no K − 1, so the total area under the curve rather than the relative increase is plotted. The relative increases in consensus are used to determine K at which there is appreciable increase. **(c)** The mean consensus score for different numbers of clusters (K ranges from 2 to 7). Cluster is indicated by color following the same color scheme as the cluster matrices and tracking plots. The bars are grouped by K which is marked on the horizontal axis. High values indicate a cluster has high stability and low values indicate a cluster has low stability. For K = 4, the mean consensus score was 0.97 for cluster 1, 0.99 for cluster 2, 0.96 for cluster 3, 0.97 for cluster 4.

with the exception of the following variables: not being a citizen of the US, levels of fasting plasma glucose, aspartate transaminase, γ-glutamyl transferase, triglyceride, and uric acid; history of diagnosed diabetes, hypercholesterolemia, congestive heart failure, and treated diabetes (Table 1). The four clusters showed distinctive patterns displayed by standardized means of cluster variables (Fig. 3). Supplementary Table 2 showed the pairwise comparisons of the clustering variables between clusters, and most differences achieved Bonferroni adjusted statistical significance ($p < 0.0083$). Supplementary Table 3 shows the standard mean differences in age and metabolic risk factors between each two clusters. The distributions of the metabolic-related factors that with large differences (standard mean differences > 0.2) are shown in Fig. 4, and the values of the features were centered to a mean value of 0 and a standard deviation of 1. Cluster 1, including 508 (7.8%) participants, was marked by relatively higher levels of BMI, WC, eGFR, ALT, and DBP, but lower levels of 2 h PG and SBP, and younger age. Cluster 2 comprised 1,439 (22.1%) participants and cluster 3 constituted 2,550 (39.1%) participants. The characteristics of cluster 2 and 3 fell somewhere in between cluster 1 and 4. Cluster 4, included 2,029 (31.1%) participants, was marked by the worst metabolic traits (higher levels of SBP, 2 h PG) and kidney function markers (lowest levels of eGFR).

Associations of CKD clusters with mortality

Totally, 2,327 participants with CKD at baseline died during the follow-up, including 850 CVD death, 368 cancer death, and 1109 death due to other causes. The Kaplan-Meier survival curves showed that there were significant different risks of all-cause mortality, CVD mortality, and other cause mortality by cluster (log-rank test all $P < 0.001$) (Fig. 5). However, there was no significant difference in cancer mortality between clusters ($P = 0.310$) (Fig. 5). The risks of all-cause mortality were significantly higher in cluster 2 (HR, 1.25; 95%CI, 1.02–1.54) and cluster 3 (HR, 1.25; 95%CI, 1.03–1.52), and in cluster 4 (HR, 1.48; 95%CI, 1.22–1.81) compared with cluster 1 (Table 2). Only cluster 4 presented significantly higher risk of CVD mortality (HR, 1.45; 95%CI, 1.05–2.02) and mortality due to other causes (HR, 1.60; 95%CI, 1.20–2.14) compared with cluster 1 (Table 2).

Sensitivity analysis

We further performed the sensitivity analysis among the 1,207 individuals with impaired baseline kidney function (eGFR < 45 ml/min per 1.73 m^2 or UACR ≥ 300 mg/g), we also identified four clusters (Supplemental Fig. 2) with similar cluster consensus scores and proportion of clustered values (Supplemental Fig. 3) as the main analyses.

	Cluster 1 (*n* = 508)	Cluster 2 (*n* = 1439)	Cluster 3 (*n* = 2550)	Cluster 4 (*n* = 2029)	*P* value
Social demographic factors					
Age (year)	52.7 ± 16.8	57.9 ± 17.1	60.5 ± 17.5	63.7 ± 16.7	< 0.001
Female	159 (31.3%)	558 (38.8%)	1288 (50.5%)	1293 (63.7%)	< 0.001
Race in non-white	319 (62.8%)	819 (56.9%)	1436 (56.3%)	1209 (59.6%)	0.014
Low educational attainment	132 (26.0%)	430 (29.9%)	922 (36.2%)	874 (43.1%)	< 0.001
Low family income-to-poverty ratio	112 (22.0%)	286 (19.9%)	601 (23.6%)	571 (28.1%)	< 0.001
Not married	249 (49.0%)	673 (46.8%)	1274 (50.0%)	1155 (56.9%)	< 0.001
Not a citizen of the US	59 (11.6%)	151 (10.5%)	311 (12.2%)	271 (13.4%)	0.085
Not homeowner	190 (37.4%)	485 (33.7%)	910 (35.7%)	794 (39.1%)	0.008
Unemployed	110 (21.7%)	354 (24.6%)	606 (23.8%)	585 (28.8%)	< 0.001
Without health insurance	109 (21.5%)	228 (15.8%)	391 (15.3%)	310 (15.3%)	0.005
Without regular health care access	86 (16.9%)	139 (9.7%)	251 (9.8%)	167 (8.2%)	< 0.001
Behavioral risk factors					
Currently smoking	280 (55.1%)	764 (53.1%)	1311 (51.4%)	966 (47.6%)	0.001
Currently drinking	399 (78.5%)	1053 (73.2%)	1712 (67.1%)	1181 (58.2%)	< 0.001
Higher level of physical activity	203 (40.0%)	716 (49.8%)	1352 (53.0%)	1231 (60.7%)	< 0.001
Sleep duration ≥ 8 h	182 (35.8%)	592 (41.1%)	1039 (40.7%)	883 (43.5%)	0.013
Mean sodium intake (mg)	6640 ± 858	4340 ± 526	2820 ± 403	1520 ± 456	< 0.001
Metabolic risk factors					
Body mass index (kg/m^2)	31.0 ± 7.91	30.6 ± 7.32	29.7 ± 6.82	29.2 ± 6.43	< 0.001
Waist circumference (cm)	106.3 ± 20.3	105.3 ± 18.0	102.9 ± 16.7	101.0 ± 15.6	< 0.001
Systolic blood pressure (mm Hg)	133.9 ± 21.8	136.0 ± 23.7	137.9 ± 25.2	141.1 ± 26.5	< 0.001
Diastolic blood pressure (mm Hg)	74.3 ± 15.4	71.5 ± 16.4	70.7 ± 16.7	69.2 ± 17.5	< 0.001
HbA1c (%)	6.50 ± 1.89	6.52 ± 1.82	6.39 ± 1.71	6.38 ± 1.71	0.049
Fasting plasma glucose (mmol/L)	7.16 ± 3.25	7.32 ± 3.40	7.10 ± 3.14	7.20 ± 3.35	0.212
2 h postprandial glucose (mmol/L)	9.73 ± 5.98	10.6 ± 6.11	10.6 ± 5.73	11.1 ± 6.01	< 0.001
Alanine transaminase (U/L)	21 (17–31)	21 (16–29)	19 (15–27)	18 (14–25)	< 0.001
Aspartate transaminase (U/L)	23 (19–29)	23 (19–28)	22 (19–28)	23 (19–28)	0.142
γ-glutamyl transferase (U/L)	25 (16–40)	23 (16–38)	22 (16–36)	22 (16–37)	0.025
Triglyceride (mmol/L)	1.58 (1.16–1.82)	1.58 (1.10–1.76)	1.58 (1.16–1.76)	1.58 (1.16–1.80)	0.638
HDL-cholesterol (mmol/L)	1.27 ± 0.47	1.31 ± 0.43	1.33 ± 0.43	1.36 ± 0.46	< 0.001
Total cholesterol (mmol/L)	5.03 ± 1.16	4.96 ± 1.27	4.99 ± 1.21	5.10 ± 1.23	0.002
LDL-cholesterol (mmol/L)	2.91 ± 1.04	2.83 ± 1.06	2.87 ± 1.06	2.94 ± 1.09	0.024
C-reactive protein (mg/dL)	0.60 (0.19–1.72)	0.57 (0.19–1.71)	0.49 (0.18–1.43)	0.47 (0.18–1.49)	0.048
Serum albumin (g/dL)	4.18 ± 0.35	4.13 ± 0.38	4.14 ± 0.38	4.11 ± 0.38	0.002
Uric acid (mg/dL)	5.86 ± 1.71	5.90 ± 1.70	5.84 ± 1.65	5.82 ± 1.75	0.641
Urinary albumin creatinine ratio (mg/g)	69.72 (41.32–180.64)	76.84 (43.54–188.68)	69.55 (41.52–170.29)	74.17 (42.25–187.41)	0.229
eGFR (ml/min/1.73m^2)	103.0 ± 22.6	96.3 ± 23.8	94.9 ± 24.3	92.6 ± 24.2	< 0.001
Diseases history					
Taking insulin	51 (10.0%)	196 (13.6%)	309 (12.1%)	248 (12.2%)	0.186
Taking oral hypoglycemic agents	109 (21.5%)	363 (25.2%)	595 (23.3%)	478 (23.6%)	0.323
Had been diagnosed diabetes	151 (29.7%)	493 (34.3%)	839 (32.9%)	662 (32.6%)	0.308
Taking antihypertensive agents	244 (48.0%)	791 (55.0%)	1431 (56.1%)	1190 (58.6%)	< 0.001
Had been diagnosed hypertension	280 (55.1%)	854 (59.3%)	1532 (60.1%)	1281 (63.1%)	0.005
Taking lipid-lowering agents	117 (23.0%)	417 (29.0%)	785 (30.8%)	606 (29.9%)	0.006
Had been diagnosed hypercholesterolemia	209 (41.1%)	618 (42.9%)	1079 (42.3%)	845 (41.6%)	0.843
Had been diagnosed congestive heart failure	36 (7.1%)	132 (9.2%)	234 (9.2%)	218 (10.7%)	0.055
Had been diagnosed coronary heart disease	28 (5.5%)	136 (9.5%)	252 (9.9%)	221 (10.9%)	0.004
Had been diagnosed heart attack	40 (7.9%)	140 (9.7%)	269 (10.5%)	245 (12.1%)	0.020
Had been diagnosed stroke	39 (7.7%)	107 (7.4%)	231 (9.1%)	220 (10.8%)	0.004
Had been diagnosed cancer or malignancy	42 (8.3%)	194 (13.5%)	376 (14.7%)	295 (14.5%)	0.001

Table 1. Baseline characteristics of study participants by cluster. Data are expressed as mean ± standard deviation, number (percentage), or median (inter quartile range).

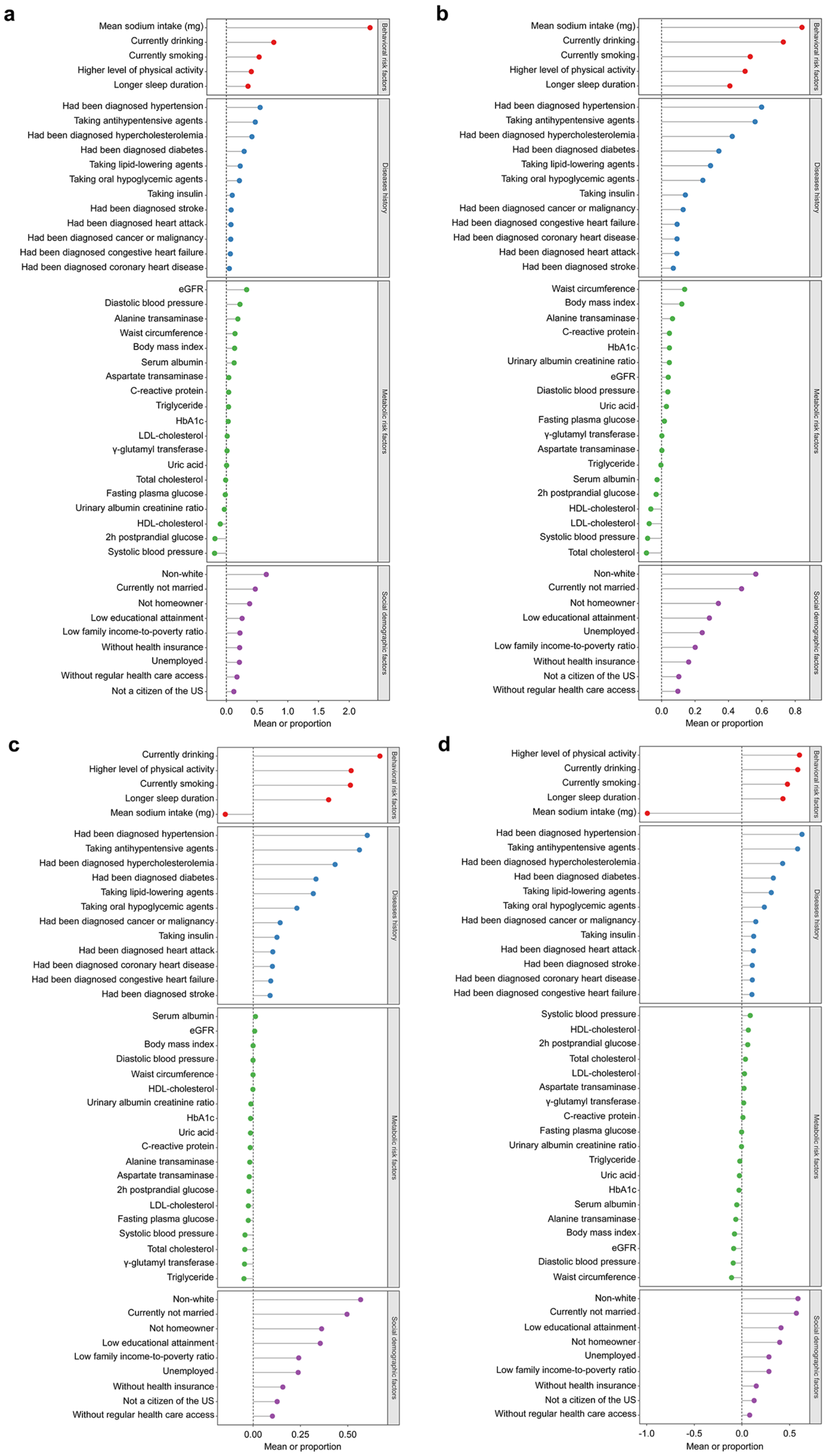

Fig. 3. Distribution of the cluster feature variables. All the values of metabolic risk factors and sodium intake were centered to a mean value of 0 and a standard deviation of 1, the other variables were presented by proportion. **(a)** Cluster 1; **(b)** Cluster 2; **(c)** Cluster 3; **(d)** Cluster 4.

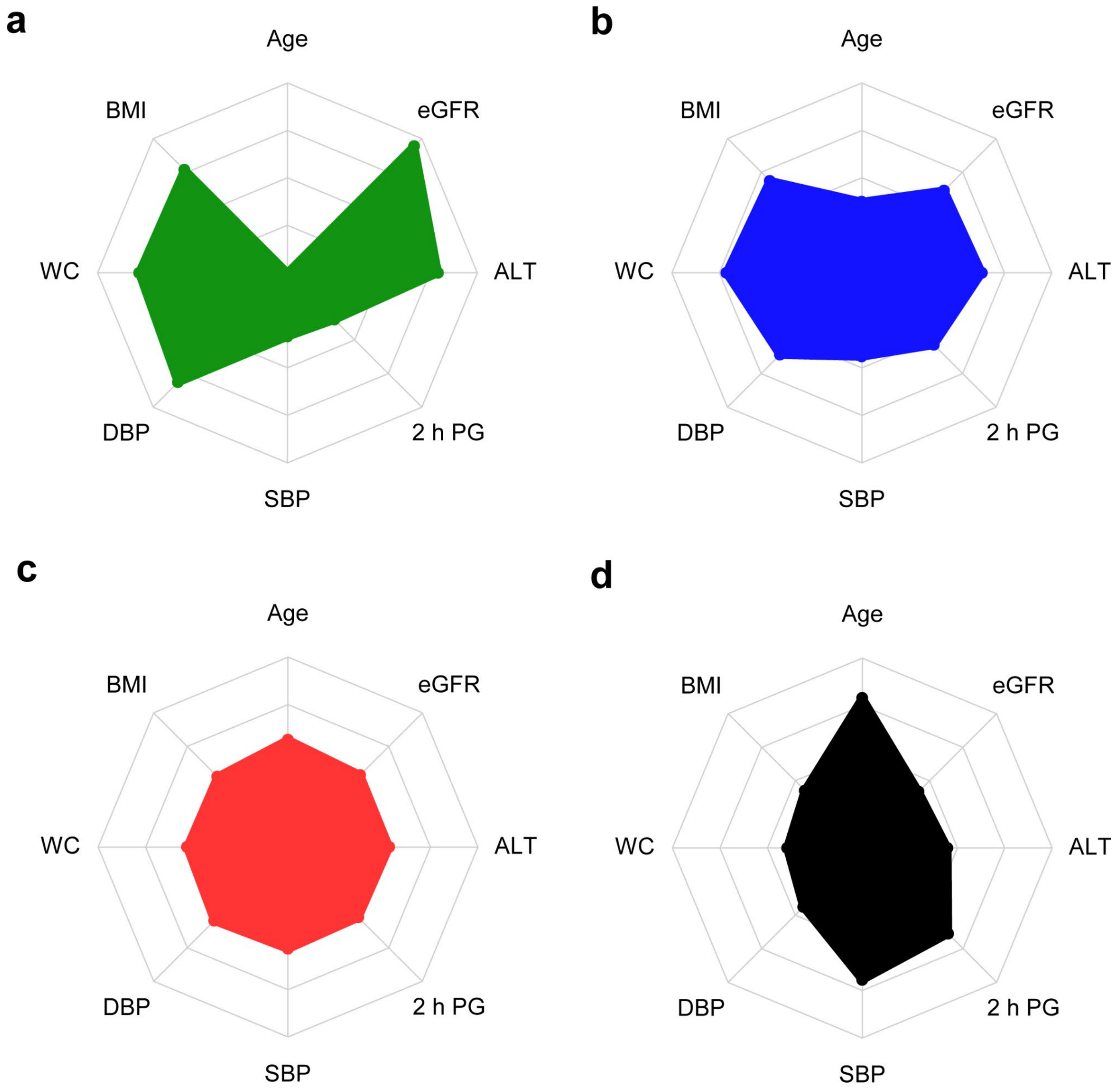

Fig. 4. Distribution of the metabolic risk factors feature variables that with larger difference. All the values of cluster feature were centered to a mean value of 0 and a standard deviation of 1. All the negative values were converted to positive value by added a fixed value to yield polygon areas related to adverse variable effects. The centre of the figure indicates 0. **(a)** Cluster 1; **(b)** Cluster 2; **(c)** Cluster 3; **(d)** Cluster 4. 2-h PG, 2-hour post-load glucose; WC, waist circumference; BMI, body mass index; ALT, alanine transaminase; SBP, systolic blood pressure, DBP, diastolic blood pressure.

The baseline characteristics of the four predefined clusters are presented in Supplementary Table 4. Similar distributions of characteristics were observed across the predefined clusters. Cluster 1, 2, 3, and 4 consisted of 129 (10.7%), 309 (25.7%), 427 (35.3%), and 342 (28.3%) participants, respectively. Cluster 1 was characterized by relatively higher levels of BMI, WC, eGFR, ALT, DBP, and sodium intake, as well as a younger age. Conversely, cluster 4 was marked by the highest levels of SBP and the lowest levels of eGFR, BMI, and WC. Moreover, participants in cluster 4 exhibited a higher prevalence of stroke, heart attack, and cancer or malignancy, as well as elder age, lower educational attainment, and the lowest sodium intake (Supplementary Table 4). The mortality risks were compared between each cluster. Even though the HRs for all-cause mortality presented increased risk trend by cluster, they did not reach statistical significance. Only cluster 4 had significantly higher risk of mortality due to other causes compared with cluster 1 (HR, 1.65; 95%CI, 1.03–1.81) (Supplementary Table 5).

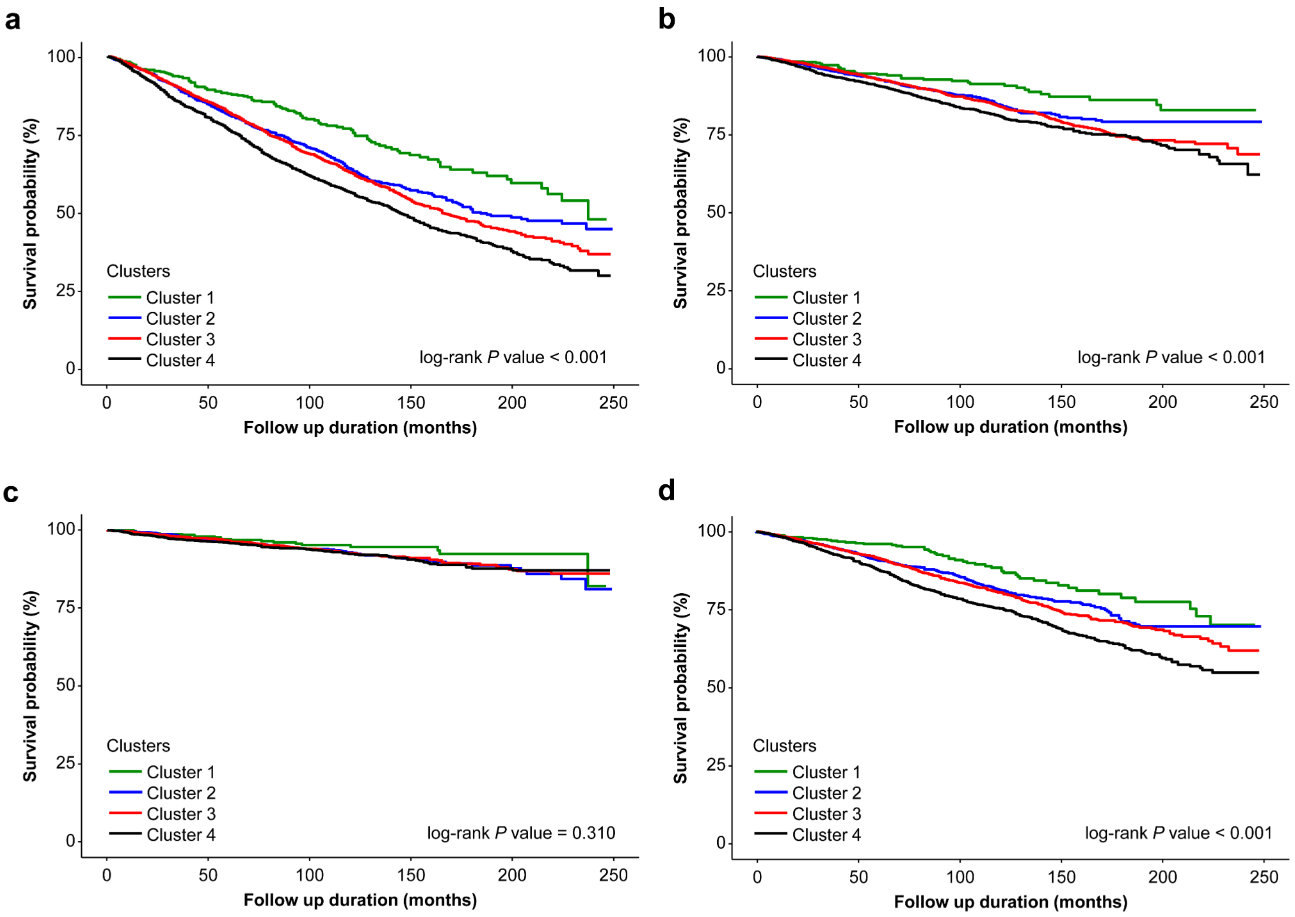

Fig. 5. Kaplan-Meier survival analysis illustrated the survival probability in different clusters. The log rank *P* values for comparisons were < 0.001 in (**a, b,** and **d**) and 0.310 in (**c**). (**a**) All-cause mortality; (**b**) cardiovascular disease mortality; (**c**) Cancer mortality 3; (**d**) Mortality due to other causes.

Another sensitivity analysis was conducted by conducting cluster analysis after excluding highly correlated variables (Supplementary Figs. 4 and 5). Six variables were excluded and 39 variables remained, the results of cluster analysis showed almost the same with the main analysis (Supplementary Table 6).

Discussion

Using unsupervised consensus clustering, we identified four distinct clusters based on 45 baseline variables. These clusters exhibited significant differences in the risk of mortality. Cluster 4, characterized by older age, higher levels of SBP, 2 h PG, lower levels of eGFR, and higher prevalence of CVD, had the highest risk of future mortality. Stratification by eGFR and UACR confirmed patient heterogeneity within the overall population. An understanding of which CKD phenotypes relate to mortality will guide both mechanistic research and promote precision medicine for CKD management.

In this study, the participants exhibited a high prevalence of obesity, as indicated by a BMI of ≥ 30 kg/m^2 and a WC of ≥ 103 cm. Cluster 1 has the highest levels of BMI and WC, whereas Cluster 4 shows the opposite pattern in our study. Existing evidence suggests a positive association between general adiposity (represented by BMI) and central adiposity (represented by WC) with the risk of all-cause mortality in the general population[18,19]. It has been observed that both higher and lower BMI values are associated with an increased risk of mortality, with obesity (BMI ≥ 30 kg/m^2) being responsible for the majority of the mortality burden[18]. However, interestingly, in our study, cluster 4, which had the lowest levels of BMI and WC, presented the higest with mortality risk. Previous studies have indicated that the association between BMI and mortality risk weakens substantially with increasing age[20,21]. Similar findings have been observed for waist circumference in participants aged 60 years and older[19]. This suggests that increased weight may actually confer a survival advantage in older individuals[20]. In our study, the participants were predominantly elderly, with a mean age of ≥ 59 years. This age composition may have attenuated the association between obesity and the risk of all-cause mortality. Overall, our findings highlight the complex relationship between obesity, age, and mortality risk, emphasizing the need for further investigation to better understand the impact of obesity on mortality in different age groups.

The patients in cluster 1 had much higher sodium intake than those in other clusters. The patients in cluster 1 are younger and more likely to be men. Generally, men and younger people have higher energy requirements than women and older people. A correlation between sodium and energy intake is evident. In addition, low

	No. of cases (mortality per 1000 PYs)	Unadjusted model	Adjusted model
All-cause mortality			
Cluster 1	116 (2.40)	1.00 (Reference)	1.00 (Reference)
Cluster 2	459 (2.52)	1.48 (1.21–1.82)	1.25 (1.02–1.54)
Cluster 3	881 (3.81)	1.60 (1.32–1.94)	1.25 (1.03–1.52)
Cluster 4	871 (6.71)	2.01 (1.65–2.43)	1.48 (1.22–1.81)
Cardiovascular disease mortality *			
Cluster 1	42 (0.87)	1.00 (Reference)	1.00 (Reference)
Cluster 2	166 (0.91)	1.48 (1.05–2.07)	1.23 (0.88–1.73)
Cluster 3	328 (1.42)	1.64 (1.19–2.26)	1.27 (0.92–1.75)
Cluster 4	314 (2.42)	1.99 (1.44–2.75)	1.45 (1.05–2.02)
Cancer mortality &			
Cluster 1	21 (0.43)	1.00 (Reference)	1.00 (Reference)
Cluster 2	83 (0.46)	1.48 (0.92–2.39)	1.25 (0.77–2.02)
Cluster 3	142 (0.61)	1.42 (0.89–2.25)	1.16 (0.73–1.83)
Cluster 4	122 (0.94)	1.55 (0.97–2.46)	1.24 (0.77–1.98)
Mortality due to other causes			
Cluster 1	53 (1.09)	1.00 (Reference)	1.00 (Reference)
Cluster 2	210 (1.15)	1.49 (1.10–2.01)	1.26 (0.94–1.71)
Cluster 3	411 (1.78)	1.64 (1.23–2.18)	1.27 (0.96–1.70)
Cluster 4	435 (3.35)	2.20 (1.65–2.92)	1.60 (1.20–2.14)

Table 2. Comparison of mortalities between clusters by cox proportional hazards model. The cluster 1 was the reference groups. Adjusted model was adjusted for age and gender. * Cardiovascular disease mortality including death due to diseases of heart (I00-I09, I11, I13, I20-I51) and cerebrovascular diseases (I60-I69). & Cancer mortality including death due to malignant neoplasms (C00-C97). PYs, person-months.

sodium intake may suggest a lower nutritional status, which is consistent with our findings that patients in cluster 1 have the lowest mortality risk. Another study also reported a correlation between higher sodium intake and lower all-cause mortality risk in American individuals[22].

As we are aware, hypertension is the most common comorbidity observed in patients with CKD[23]. In our present study, cluster 4 was characterized by a higher prevalence of hypertension and poorer kidney function. Uncontrolled hypertension in CKD patients can lead to detrimental clinical outcomes, such as myocardial infarction, acute coronary syndrome, ischemic stroke, heart failure, and even death[24]. Previous research has shown that SBP has a stronger association with adverse kidney outcomes than DBP in CKD patients[25]. In cluster 4, we observed higher levels of SBP and relatively lower levels of DBP, suggesting that SBP might be a more potent predictor of CKD progression and have a greater impact on future mortality in CKD patients compared to DBP. In conclusion, our findings emphasize the importance of effectively managing blood pressure in patients with CKD.

Our study focused on patients belonging to cluster 4, characterized by poorer kidney function (indicated by lower levels of eGFR) and a higher prevalence of CVD, including coronary heart disease and heart attacks. This particular group demonstrated an elevated risk of future mortality. It is widely acknowledged that eGFR and albuminuria are strongly and independently associated with various adverse outcomes, such as progression to kidney failure, cardiovascular events, and death[1]. Our study confirmed these previous findings by establishing a significant association between decreased eGFR and the risk of all-cause mortality. Prior research has consistently demonstrated that CVD is a leading cause of death among patients with CKD[1]. In line with these findings, our study revealed that higher-risk groups for cardiovascular events experienced higher overall mortality rates and an increased risk of CVD-related mortality. These results underscore the importance of reducing cardiovascular risk factors to slow down CKD progression and strive to prolong the lifespan of patients with CKD. By prioritizing strategies aimed at managing CVD risk factors in this population, we can potentially improve outcomes, reduce mortality rates, and enhance the overall quality of life for individuals with CKD.

Our study highlights the significant heterogeneity observed within the broad category of CKD, which encompasses various pathologies and etiologies. In order to explore this diversity, we employed consensus clustering using discrete data elements. The use of consensus clustering allowed us to identify distinct patient groups with specific clinical markers. Of particular interest is the stratification of patients based on their kidney function, distinguishing between those with relatively preserved or impaired kidney function. This stratification provides valuable insights for tailoring clinical interventions and treatments in a more precise manner. However, it is important to note that further clinical studies are required to validate the utility of this approach. These studies should investigate whether the identified patient groupings indeed lead to more targeted interventions and improved clinical outcomes, thus fulfilling the promise of precision medicine.

The study has several limitations that need to be considered. Firstly, it's important to acknowledge that cluster analysis is an exploratory data analysis method. It relies on the input of data, such as the 45 baseline variables used in our analysis, to identify clusters of individuals with similar characteristics. Therefore, if different patient characteristics were used as input variables, it is likely that different CKD subgroups would be identified. It's essential to interpret the results of cluster analysis within this context and understand that the identified subgroups are specific to the variables used in the analysis. Second, our clustering analysis relied mainly on well-known clinical traits associated with CKD, some other factors may also contribute to identifying sub-types of CKD. Future studies should explore the inclusion of novel biomarker data, such as inflammatory markers, genomics, metabolomics, and proteomics, to enhance our understanding of CKD heterogeneity. Finally, our study population was limited to the US population. Therefore, our findings may not be generalizable to other populations. It is essential to validate our results in a broader range of populations in the future.

By using the probably associated factors of CKD, our study demonstrates that substantial heterogeneity with sophisticated phenotyping and underlying disease pathologies exists within the broad category of "CKD". Our findings also added new information to the 2012 KDIGO CKD categorization guideline. The underlying heterogeneity in CKD staging by eGFR and urine ACR is not adequately captured. Patients with CKD can be subtyped into four separate subgroups based on 45 baseline features by applying data-driven clustering to multidimensional patient data, including demographics, biomarkers from blood and urine, health status and behaviors, and medication use. These subgroups are associated with future mortality risks. Identification of clinically subgroups among CKD patients provides an important step toward patient classification and management. Further research is needed to validate and refine these clusters, explore their reproducibility, and investigate underlying mechanisms. These findings contribute to a deeper understanding of CKD and have implications for improved patient outcomes.

Methods

Study population

The US National Health and Nutrition Examination Survey (NHANES) is an ongoing cross-sectional, national, stratified, multistage probability surveys of the civilian, noninstitutionalized US population[26]. About 10 thousand individuals in each survey for every 2 years are investigated to complete a household interview and underwent a physical examination. In present study, we used data from NHANES III 1999–2000 to 2017–2018 which included 10 cycles of survey. All nonpregnant participants with 20 years or older and CKD were included in the analysis. A detailed description of the NHANES database is publicly available (http://www.cdc.gov/nchs/nhanes.htm).

We combined ten consecutive survey cycles which included 101,316 participants. Participants were excluded for aged younger than 20 ($n=46{,}235$), being pregnant at examination or uncertain of the pregnancy status ($n=2{,}639$), and having received dialysis treatment in the past 12 months ($n=162$). Participants were also excluded due to missing data on mortality or outliers of cluster variables. Finally, a total of 6,526 eligible subjects with CKD were included in the analysis (Supplementary Fig. 1). The NHANES protocol was approved by the National Center for Health Statistics (NCHS) Institutional Review Board and written informed consent was obtained. Our study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline[27].

Mortality data

NHANES III participant records were linked to mortality data from the National Death Index based on death certificate data (https://www.cdc.gov/nchs/data-linkage/mortality-public.htm). International Classification of Diseases, Tenth Revision codes were used to identify the cause of deaths. Cardiovascular death includes death due to diseases of heart (I00-I09, I11, I13, I20-I51) and cerebrovascular diseases (I60-I69). Cancer death was classified using codes C00-C97. Person-months were calculated in months from the date of interview to date of death or most recent vital status record.

Variables

Sociodemographic characteristics, behavioral risk factors and history of diseases were administered in the survey by trained interviewers using questionnaires. The physical examinations and laboratory tests in NHANES took place in a mobile examination center using standardized protocols and calibrated equipment, and details on the data collection are described on the website (https://wwwn.cdc.gov/nchs/nhanes/analyticguidelines.aspx). We selected 45 clinically available and novel factors from over all variables that were measured at NHANES study baseline. Variables were selected on the basis of literature review for those that are most clinically relevant to CKD[28–32]. We excluded variables with over 10% missing data or small variability (e.g., binary variable with $<5\%$). The 45 variables included variables of sociodemographic characteristics ($n=9$), behavioral risk factors ($n=5$), biomarkers of metabolic status ($n=19$), and history of diseases ($n=12$).

The sociodemographic characteristics included ethnicity, education level, family income, marriage status, citizenship status, housing, employment, health insurance, and regular health care access. The ethnicity was categorized into non-Hispanic white and non-white. Low education attainment was defined as attaining less than a high school education. The income-to-poverty ratio (annual family income divided by the poverty threshold adjusted for family size and inflation) was used as a measure of income. The low income-to-poverty ratio was defined as less than 100%. The marriage status was dichotomized as currently married and not married. The citizenship status includes two options, citizen by birth or naturalization and not a citizen of the US. For investigating housing status, the participants were asked "Is this home owned, being bought, rented, or occupied by some other arrangement by you or someone else in your family?" Employed status was dichotomized as unemployed and employed, student, or retired. The type of health insurance was also dichotomized as with and without health insurance. The participants were asked "Is there a place that you usually go when you are sick, or do you need advice about your health?" for investigating the health care access.

The behavioral risk factors included currently smoking status, currently drinking status, physical activity level, sleep duration and sodium intake. Current smoking was defined as having smoked at least 100 cigarettes in life and smoking at present. Current alcohol drinking was defined as taking at least 12 times drinks of any type of alcoholic beverage in the last 12 months. Physical activity was estimated using the form of the Global Physical Activity Questionnaire by asking questions about the intensity, duration, and frequency of physical activity. There were different types of physical activity assessment tools used in NHANES 1999–2000 to 2005–2006 and NHANES 2007–2008 to 2017–2018. In NHANES 1999–2000 to 2005–2006, the duration of the physical activity was not ascertained, each physical activity was assigned an intensity value (metabolic equivalent tasks) that represents the ratio of the energy expenditure of the activity to the basal metabolic rate. In NHANES 2007–2008 to 2017–2018, total metabolic equivalent minutes per week were calculated as the measurement of physical activity level for the subjects. A higher level of physical activity was defined as having a higher metabolic equivalent/week than the median levels of the metabolic equivalent/week by investigation cycles. The usual sleep duration at night was investigated and long sleep duration was considered as sleep longer than 8 h. The sodium intake was collected through dietary interview.

The physical examinations and laboratory tests of metabolic biomarkers were collected using standardized protocols and assays, including body mass index (BMI), waist circumference (WC), systolic blood pressure (SBP), diastolic blood pressure (DBP), HbA1c, fasting plasma glucose (FPG), 2 h postprandial glucose (2 h PG), alanine transaminase (ALT), aspartate transaminase (AST), γ-glutamyl transferase (GGT), triglyceride (TG), high-density lipoprotein cholesterol (HDL-cholesterol), total cholesterol, low-density lipoprotein cholesterol (LDL-cholesterol), C-reactive protein (CRP), serum albumin, uric acid (UA), Urinary albumin creatinine ratio (UACR), and eGFR.

The information on currently taking prescribed medicine for treating hypertension, diabetes, and hypercholesterolemia was investigated in the survey. The history of diseases, including congestive heart failure, coronary heart disease, heart attack, stroke, and cancer or malignancy was also collected.

Definition of CKD

The eGFR was calculated using the 2009 chronic kidney disease epidemiology collaboration (CKD-EPI) equation with considering the sex, serum creatinine level and race[33]. Albuminuria was calculated using urinary albumin divided by the urinary creatinine based on morning spot urine. CKD was defined as eGFR level < 60 ml/min/1.73 m^2 or UACR ≥ 30 mg/g. Equations expressed for specified sex and serum creatinine level were showed in the Supplementary Table 1.

Statistical analysis

We employed multiple imputation with arbitrary missing patterns to correct for response bias under the assumption of missing at random, and to maximally utilize existing risk factor data. The continuous variables with skewed distributions were log-transformed to normal distributions in the imputation process. The linear regression method was used to impute missing values for continuous variables, and the logistic regression method for variables having binary or ordinal responses. For each variable with missing data, we used the other variables to impute.

We performed consensus clustering analysis on the participants with chronic kidney disease and the continuous values were centered to a mean value of 0 and a standard deviation of 1. The clustering algorithm is to maintain high cluster consensus while maximizing the number of clusters[34]. With the prespecified setting a number of clusters K = 2, 3, …, 7, the consensus clustering algorithm generated a random subset that contained 80% of the data records without replacement and repeated 100 times for each number of clusters. For each random subset, the K-means (Euclidean distance-based) algorithm was conducted while each individual was assigned to one of the clusters. The frequencies of any pair of two individuals were calculated after 100 iterations, which were grouped together under each scenario of K and constructed a matrix of participants' pairwise consensus value[34]. In the consensus matrix, consensus values ranged from 0 (never clustered together) to 1 (always clustered together) were marked by white to bright blue. For each number of cluster analysis, the cluster memberships are marked by colored rectangles. The consensus matrix is ordered by the consensus clustering which is displayed as a dendrogram atop the heatmap.

The optimal number of clusters was ascertained by observing the consensus matrix heat map, the within-cluster consensus scores, and the cumulative distribution function (CDF) (range 0–1) plot[34]. The CDF plot showed the area under the CDFs for each K, and for a specific number of clusters, the CDF reached an approximate maximum, thus consensus and cluster confidence was at a maximum at this K. The relative change in area under the CDF curve comparing K and K − 1 was also used to determine the optimal number of clusters. The cluster consensus score, ranged between 0 and 1, was defined as the average consensus value for all pairs of individuals belonging to the same cluster. A value approached to 1 indicated better cluster stability[34].

For continuous variables with normal distribution, we calculated mean and standard deviation; for continuous variables with skewed distributions, we calculated median and interquartile range; and for categorical variables, we presented count and percentage. To present the cluster profiles of the 45 variables, we graphically displayed the standardized means of continuous variables (metabolic risk factors and sodium intake) and proportions for categorical variables (the other variables) by cluster. The frequencies of endpoints related to death were calculated as the number of events divided by personmonths of observation censored at the date of event occurrence, death, or follow-up visit, whichever came first. Adjusted Cox proportional hazard models were used and hazard ratios (HRs) with 95% confidence intervals (CIs) were calculated to estimate the risks for all-cause mortality, CVD mortality, cancer mortality and mortality due to other causes by cluster.

All the statistical analysis was conducted using the R version 4.2.3 (http://www.r-project.org). Consensus clustering analysis was done using the *ConsensusClusterPlus* function (minimum K = 2, maximum K = 7, replication = 100, proportion of random subset = 0.8, Euclidean distance-based K-means algorithm) in the '*ConsensusClusterPlus*' package in R version 4.2.3 (http://www.r-project.org).

Data availability

All data are fully available on request from the corresponding author.

Received: 8 January 2024; Accepted: 25 November 2024
Published online: 02 December 2024

References

1. Kalantar-Zadeh, K., Jafar, T. H., Nitsch, D., Neuen, B. L. & Perkovic, V. Chronic kidney disease. *Lancet* **398** (10302), 786–802 (2021).
2. Astor, B. C. et al. Lower estimated glomerular filtration rate and higher albuminuria are associated with mortality and end-stage renal disease. A collaborative meta-analysis of kidney disease population cohorts. *Kidney Int.* **79** (12), 1331–1340 (2011).
3. Chronic Kidney Disease Prognosis Consortium;et al et al. Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. *Lancet* **375** (9731), 2073–2081 (2010).
4. Gansevoort, R. T. et al. Lower estimated GFR and higher albuminuria are associated with adverse kidney outcomes. A collaborative meta-analysis of general and high-risk population cohorts. *Kidney Int.* **80** (1), 93–104 (2011).
5. GBD Chronic Kidney Disease Collaboration. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the global burden of Disease Study 2017. *Lancet* **395** (10225), 709–733 (2020).
6. Xie, Y. et al. Analysis of the Global Burden of Disease study highlights the global, regional, and national trends of chronic kidney disease epidemiology from 1990 to 2016. *Kidney Int.* **94** (3), 567–581 (2018).
7. Tervaert, T. W. et al. Pathologic classification of diabetic nephropathy. *J. Am. Soc. Nephrol.* **21** (4), 556–563 (2010).
8. Cheung, A. K. et al. Effects of intensive BP Control in CKD. *J. Am. Soc. Nephrol.* **28** (9), 2812–2823 (2017).
9. Kurts, C., Panzer, U., Anders, H. J. & Rees, A. J. The immune system and kidney disease: basic concepts and clinical implications. *Nat. Rev. Immunol.* **13** (10), 738–753 (2013).
10. Parsa, A. et al. APOL1 risk variants, race, and progression of chronic kidney disease. *N Engl. J. Med.* **369** (23), 2183–2196 (2013).
11. Carrero, J. J. et al. Comparison of nutritional and inflammatory markers in dialysis patients with reduced appetite. *Am. J. Clin. Nutr.* **85** (3), 695–701 (2007).
12. Soderland, P., Lovekar, S., Weiner, D. E., Brooks, D. R. & Kaufman, J. S. Chronic kidney disease associated with environmental toxins and exposures. *Adv. Chronic Kidney Dis.* **17** (3), 254–264 (2010).
13. Wright, J. T. Jr et al. Effect of blood pressure lowering and antihypertensive drug class on progression of hypertensive kidney disease: results from the AASK trial. *JAMA* **288** (19), 2421–2431 (2002).
14. Xie, Y. et al. Proton Pump inhibitors and risk of Incident CKD and Progression to ESRD. *J. Am. Soc. Nephrol.* **27** (10), 3153–3163 (2016).
15. Cacoub, P., Desbois, A. C., Isnard-Bagnis, C., Rocatello, D. & Ferri, C. Hepatitis C virus infection and chronic kidney disease: time for reappraisal. *J. Hepatol.* **65** (1 Suppl), S82–S94 (2016).
16. Soria, D. et al. A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients. *Comput. Biol. Med.* **40** (3), 318–330 (2010).
17. Zheng, R. et al. Data-driven subgroups of prediabetes and the associations with outcomes in Chinese adults. *Cell. Rep. Med.* **4** (3), 100958 (2023).
18. Bhaskaran, K., Dos-Santos-Silva, I., Leon, D. A., Douglas, I. J. & Smeeth, L. Association of BMI with overall and cause-specific mortality: a population-based cohort study of 3·6 million adults in the UK. *Lancet Diabetes Endocrinol.* **6** (12), 944–953 (2018).
19. Jayedi, A., Soltani, S., Zargar, M. S., Khan, T. A. & Shab-Bidar, S. Central fatness and risk of all cause mortality: systematic review and dose-response meta-analysis of 72 prospective cohort studies. *BMJ* **370**, m3324 (2020).
20. Lv, Y. et al. The obesity paradox is mostly driven by decreased noncardiovascular disease mortality in the oldest old in China: a 20-year prospective cohort study. *Nat. Aging*. **2** (5), 389–396 (2022).
21. Ng, T. P. et al. Age-dependent relationships between body mass index and mortality: Singapore longitudinal ageing study. *PLoS One*. **12** (7), e0180818 (2017).
22. Liu, D. et al. Sodium, potassium intake, and all-cause mortality: confusion and new findings. *BMC Public. Health*. **24**, 180 (2024).
23. Kim, H. et al. Baseline Cardiovascular characteristics of adult patients with chronic kidney disease from the KoreaN Cohort Study for outcomes in patients with chronic kidney Disease (KNOW-CKD). *J. Korean Med. Sci.* **32** (2), 231–239 (2017).
24. James, P. A. et al. 2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the Eighth Joint National Committee (JNC 8). *JAMA* **311** (5), 507–520 (2014).
25. Lee, J. Y. et al. Association of blood pressure with the progression of CKD: findings from KNOW-CKD Study. *Am. J. Kidney Dis.* **78** (2), 236–245 (2021).
26. Curtin, L. R. et al. The National Health and Nutrition Examination Survey: Sample Design, 1999–2006. *Vital Health Stat. 2* ;(155):1–39. (2012).
27. von Elm, E. et al. The strengthening the reporting of Observational studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann. Intern. Med.* **147** (8), 573–577 (2007).
28. Kronenberg, F. Emerging risk factors and markers of chronic kidney disease progression. *Nat. Rev. Nephrol.* **5** (12), 677–689 (2009).
29. Kao, H. Y. et al. Associations between sex and risk factors for Predicting chronic kidney disease. *Int. J. Environ. Res. Public. Health*. **19** (3), 1219 (2022).
30. Geylis, M., Coreanu, T., Novack, V. & Landau, D. Risk factors for childhood chronic kidney disease: a population-based study. *Pediatr. Nephrol.* **38** (5), 1569–1576 (2023).
31. Kareem, S. et al. Epidemiology and risk factors of chronic kidney Disease in Rural areas 4 (Badin) of Sind, Pakistan. *J. Pak Med. Assoc.* **73** (7), 1399–1402 (2023).
32. Xie, Y. & Chen, X. Epidemiology, major outcomes, risk factors, prevention and management of chronic kidney disease in China. *Am. J. Nephrol.* **28** (1), 1–7 (2008).
33. Stevens, P. E., Levin, A. & Kidney Disease: Improving Global Outcomes Chronic Kidney Disease Guideline Development Work Group Members. Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline. *Ann. Intern. Med.* **158** (11), 825–830 (2013).
34. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52**, 91–118 (2003).

Acknowledgements

The authors would like to thank all the participants of the NHANES. We thank the individual patients who provided the sample that made data available; without them the study would not have been possible.

Author contributions

SW designed and supervised the study and acted as the corresponding author and guarantor. YQ and LX served as first authors and contributed to the study implementation, data analysis, and manuscript preparation. ZW, YD and BL contributed to result interpretation, revisions, and finalizing the manuscript. All authors reviewed and approved the manuscript's final version.

Funding

This work was supported by grants from the Youth Program for Shandong Province Science and Technology of Traditional Chinese Medicine (No. Q-2022117); the National Nature Science Foundation of China (No. 82200968); the Provincial Nature Science Foundation of Shandong (No. ZR2022QH295).

Declarations

Ethics approval and consent to participate

The NHANES is a public-use dataset available through the website. The NHANES protocol was approved by the institutional review board of the Centers for Disease Control and Prevention (https://www.cdc.gov/nchs/nhanes/ irba98. htm). NHANES has obtained written informed consent from all participants.

Competing interests

The authors declare no competing interests.

Compliance with ethics guidelines

All authors declare that they have no conflict of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-81208-1.

Correspondence and requests for materials should be addressed to S.W.