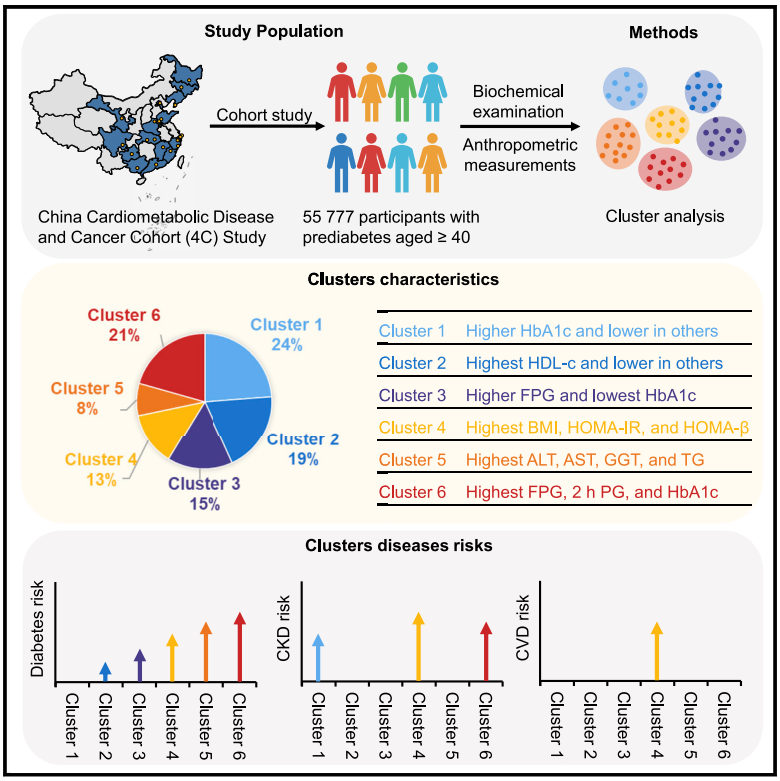
Cell Reports
Medicine

Article

Data-driven subgroups of prediabetes and the associations with outcomes in Chinese adults

Graphical abstract



Highlights

- Individuals with prediabetes have heterogeneity in metabolic features
- Prediabetes can be classified into six subgroups with different disease risks
- Prediabetes with obesity and insulin resistance has the highest CVD risk
- There are specific trends of transition from prediabetes clusters to diabetes clusters

Authors

Ruizhi Zheng, Yu Xu, Mian Li, ..., Jieli Lu, Yufang Bi, for the China Cardiometabolic Disease and Cancer Cohort (4C) Study Group

Correspondence

wqingw61@163.com (W.W.), jielilu@hotmail.com (J.L.), byf10784@rjh.com.cn (Y.B.)

In brief

Zheng et al. use data-driven clustering approaches to confirm the heterogeneity in individuals with prediabetes and explore their associations with major diseases. Individuals with prediabetes differ in metabolic features and risks of disease progression, raising the possibility of a practical, stratified approach for the prevention of diabetes and related diseases.

Zheng et al., 2023, Cell Reports Medicine *4*, 100958
March 21, 2023 © 2023
https://doi.org/10.1016/j.xcrm.2023.100958

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Data-driven subgroups of prediabetes and the associations with outcomes in Chinese adults

Ruizhi Zheng,[1,2,27] Yu Xu,[1,2,27] Mian Li,[1,2,27] Zhengnan Gao,[3,27] Guixia Wang,[4] Xinguo Hou,[5] Li Chen,[5] Yanan Huo,[6] Guijun Qin,[7] Li Yan,[8] Qin Wan,[9] Tianshu Zeng,[10] Lulu Chen,[10] Lixin Shi,[11] Ruying Hu,[12] Xulei Tang,[13] Qing Su,[14] Xuefeng Yu,[15] Yingfen Qin,[16] Gang Chen,[17] Xuejiang Gu,[18] Feixia Shen,[18] Zuojie Luo,[16] Yuhong Chen,[1,2] Yinfei Zhang,[19] Chao Liu,[20] Youmin Wang,[21] Shengli Wu,[22] Tao Yang,[23] Qiang Li,[24] Yiming Mu,[25] Jiajun Zhao,[26] Chunyan Hu,[1,2] Xiaojing Jia,[1,2] Min Xu,[1,2] Tiange Wang,[1,2] Zhiyun Zhao,[1,2] Shuangyuan Wang,[1,2] Hong Lin,[1,2] Guang Ning,[1,2] Weiqing Wang,[1,2,*] Jieli Lu,[1,2,*] Yufang Bi,[1,2,28,*] and for the China Cardiometabolic Disease and Cancer Cohort (4C) Study Group

[1]Department of Endocrine and Metabolic Diseases, Shanghai Institute of Endocrine and Metabolic Diseases, Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China
[2]Shanghai National Clinical Research Center for Metabolic Diseases, Key Laboratory for Endocrine and Metabolic Diseases of the National Health Commission of the PR China, Shanghai Key Laboratory for Endocrine Tumor, State Key Laboratory of Medical Genomics, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[3]Dalian Municipal Central Hospital, Dalian, China
[4]The First Hospital of Jilin University, Changchun, China
[5]Qilu Hospital of Shandong University, Jinan, China
[6]Jiangxi Provincial People's Hospital Affiliated to Nanchang University, Nanchang, China
[7]The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China
[8]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China
[9]The Affiliated Hospital of Southwest Medical University, Luzhou, China
[10]Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
[11]Affiliated Hospital of Guiyang Medical College, Guiyang, China
[12]Zhejiang Provincial Center for Disease Control and Prevention, China
[13]The First Hospital of Lanzhou University, Lanzhou, China
[14]Xinhua Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China
[15]Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
[16]The First Affiliated Hospital of Guangxi Medical University, Nanning, China
[17]Fujian Provincial Hospital, Fujian Medical University, Fuzhou, China
[18]The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China
[19]Central Hospital of Shanghai Jiading District, Shanghai, China
[20]Jiangsu Province Hospital on Integration of Chinese and Western Medicine, Nanjing, China
[21]The First Affiliated Hospital of Anhui Medical University, Hefei, China
[22]Karamay Municipal People's Hospital, Xinjiang, China
[23]The First Affiliated Hospital of Nanjing Medical University, Nanjing, China
[24]The Second Affiliated Hospital of Harbin Medical University, Harbin, China
[25]Chinese People's Liberation Army General Hospital, Beijing, China
[26]Shandong Provincial Hospital Affiliated to Shandong University, Jinan, China
[27]These authors contributed equally
[28]Lead contact
*Correspondence: wqingw61@163.com (W.W.), jielilu@hotmail.com (J.L.), byf10784@rjh.com.cn (Y.B.)
https://doi.org/10.1016/j.xcrm.2023.100958

SUMMARY

Prediabetes and its pathophysiology remain important issues. We aimed to examine the cluster characteristics of prediabetes and explore their associations with developing diabetes and its complications based on 12 variables representing body fat, glycemic measures, pancreatic β cell function, insulin resistance, blood lipids, and liver enzymes. A total of 55,777 individuals with prediabetes from the China Cardiometabolic Disease and Cancer Cohort (4C) were classified at baseline into six clusters. During a median of 3.1 years of follow-up, significant differences in the risks of diabetes and its complications between clusters were observed. The odds ratios of diabetes stepwisely increase from cluster 1 to cluster 6. Clusters 1, 4, and 6 have increased chronic kidney diseases risks, while the prediabetes in cluster 4, characterized by obesity and insulin resistance, confers higher risks of cardiovascular diseases compared with others. This subcategorization has potential value in developing more precise strategies for targeted prediabetes prevention and treatment.

INTRODUCTION

Prediabetes is a high-risk state for type 2 diabetes mellitus (T2DM), defined as levels of fasting plasma glucose (FPG), 2-h post-load plasma glucose (PG), or hemoglobin A1c (HbA1c) above their normal ranges but below diagnostic thresholds for diabetes.[1] Almost one-third of the Chinese population has prediabetes; similar numbers were reported worldwide, including in the United States.[2,3] Approximately, 5% to 10% of individuals with prediabetes will progress to diabetes per year.[4] Much emphasis has been placed on finding an effective method to delay or prevent incident T2DM among those with prediabetes. The Finnish Diabetes Prevention Study[5] and the Da Qing Diabetes Prevention Outcome Study[6] provided strong evidence that a combination of diet and exercise interventions was the most important factor that could halt the progression to T2DM and reduce the incidence of cardiovascular disease (CVD) events in patients with prediabetes. However, the questions surrounding prediabetes and its management have brought debate and remain important issues that warrant further investigation.[7] For example, the high-risk population is extremely large and the costs of prevention are substantial and rising. The term "prediabetes" has also been criticized because a substantial proportion of people with prediabetes regress to normoglycemia without treatment.[8] In such circumstances, it might imply that intervention is not necessary because no disease is present.

The dilemma between health benefits from the intervention and economic consideration in terms of cost-effectiveness calls for re-classification of prediabetes to enable precise and effective intervention in those at the greatest risk of T2DM and other complications. Wagner et al. used clustering analysis to classify Caucasian individuals at elevated risk for T2DM into six clusters with different metabolic features and disease risks.[9] Their results demonstrated that pathophysiological heterogeneity exists before the diagnosis of T2DM and highlighted groups differing in the risk for T2DM and its complications.[9] However, physiological features are different between Asian and Caucasian individuals. The Chinese population is more likely to have an impaired β-cell function and is more susceptible to the effects of overall obesity on metabolic factors.[10,11] It is unknown whether Chinese people with prediabetes can be classified into different subphenotypes.

We aimed to examine and validate whether measurements of metabolic parameters endorse the prediabetes clusters, and whether there are differences in disease risks between clusters. We postulated that specific cluster-based subphenotypes of prediabetes correlate with T2DM and complications differently, therefore individualized intervention is required.

RESULTS

Study population

A total of 193,846 adults aged 40 years or older were recruited from the China Cardiometabolic Disease and Cancer Cohort (4C) study at baseline, and 170,240 (87.8%) participants attended an in-person follow-up visit.[10,12] Among them, 105,590 participants were excluded because they had a major disease or did not have prediabetes at baseline; 7,416 participants were excluded because of missing data of baseline biochemical measurements and physical examination; and 1,457 participants were further excluded for having outlier values of the cluster variables. Finally, 55,777 participants with prediabetes were included in the analysis (Figure S1), and the total person-years of follow-up were 212,827 person-years.

Determination of cluster number

By visualizing the matrix heatmaps of the pairwise consensus for each cluster size (Figures S2A–S2G), the cumulative distribution functions (CDFs) (Figure S2H), and the proportion increase of the area under the CDFs (Figure S2I), the consensus clustering algorithm identified that $K = 6$ was the largest number of clusters best representing the data pattern of participants with prediabetes. For $K = 6$, the mean consensus scores were greater than 0.75 for all clusters, with a larger value indicating better stability of cluster membership (Figure S2J). The characteristics of participants in the six clusters are shown in Figure S3.

Distributions of glycemic status by clusters

We used K-means to cluster the overall prediabetes participants, and cluster stability was estimated as Jaccard means, which were greater than 0.77 for all clusters. As shown in Figure 1, the crosstabs of the participant's number between clusters and prediabetes definitions shows that most participants with prediabetes in cluster 1 and cluster 2 were diagnosed by HbA1c $\geq$5.7% only. Almost half of the participants in cluster 3 only had FPG $\geq$5.6 mmol/L; in cluster 4 and cluster 5, the participants with more than one abnormal glycemic criterion outweighed those with only one abnormal criterion; and almost 50% of those with prediabetes in cluster 6 had abnormal FPG, 2 h PG, and HbA1c, simultaneously.

Distribution of the clinical features by clusters

Demographic, anthropometric, and clinical data for the six clusters are shown in Table 1. The six clusters showed distinctive patterns displayed by standardized means of cluster variables (Figure 2, Table S1). Cluster 1, including 13,258 (23.8%) participants, was marked by relatively higher levels of HbA1c but lower levels of other features. Cluster 2 comprised 10,836 (19.4%) participants. These individuals had the highest levels of high-density lipoprotein cholesterol (HDL-c) than the other clusters and similar HbA1c levels as cluster 1. Cluster 3 constituted 8,664 (15.5%) participants. This group was characterized by higher levels of FPG and the lowest HbA1c level than the other clusters. Cluster 4, including 7,246 (13.0%) participants, was marked by the highest levels of body mass index (BMI), homoeostasis model assessment of insulin resistance (HOMA-IR), and homoeostasis model assessment β of cell function (HOMA-β). Cluster 5 comprised 4,310 (7.7%) participants with the highest levels of aspartate aminotransferase (AST), alanine transaminase (ALT), glutamyl transferase (GGT), and triglyceride (TG). Cluster 6, including 11,463 (20.6%) participants, was characterized by the highest levels of FPG, 2 h PG, and HbA1c. The pairwise comparisons of the clustering variables between clusters are shown in Figure 3. Most differences achieved Bonferroni adjusted statistical significance ($p < 0.003$), and half of them

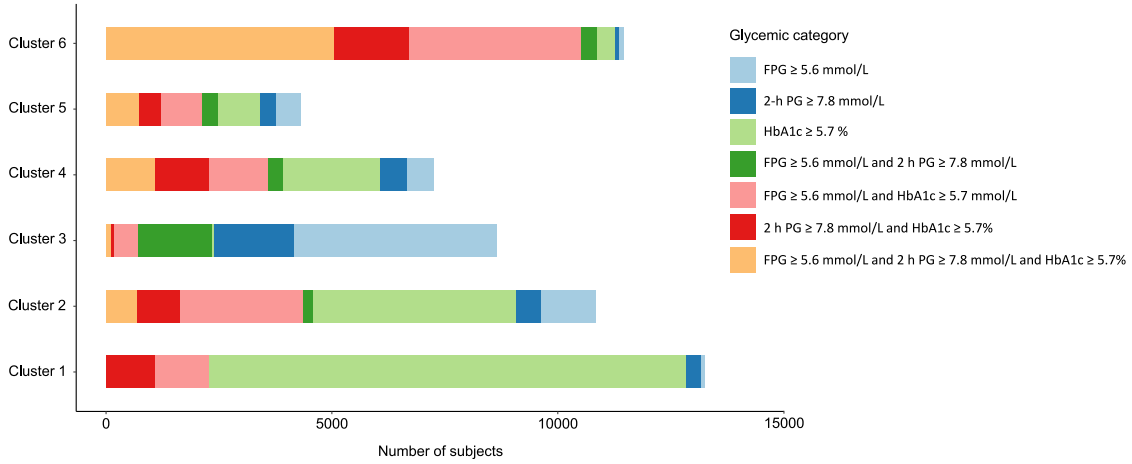


Figure 1. The number of participants in each group of glycemic status by clusters
FPG, fasting plasma glucose; 2 h PG, 2 h post-load plasma glucose.

showed small or negligible differences tested by using Cohen's d value (Table S2).

Associations of prediabetes clusters with major diseases

During a median of 3.1 years of follow-up, 3,615 and 1,020 participants with prediabetes developed T2DM and chronic kidney disease (CKD), respectively. We found significant differences in T2DM incidence between clusters and the risks of T2DM were stepwisely increased from cluster 1 to cluster 6 (Figure 4, Table S3). Compared with cluster 1, which had the lowest incidence of T2DM, the odds ratios (ORs) for developing T2DM in cluster 2 to cluster 6 were 1.44 (95% confidence interval [CI], 1.24–1.66), 1.80 (95% CI, 1.55–2.08), 3.00 (95% CI, 2.61–3.45), 3.69 (95% CI, 3.18–4.28), and 4.93 (95% CI, 4.37–5.57) in adjusted model 2, respectively. The incidence of CKD between clusters was also different, and cluster 2 had the lowest incidence of CKD (Figure 4). Compared with cluster 2, cluster 1 (OR, 1.82; 95% CI, 1.50–2.27), cluster 4 (OR, 2.29; 95% CI, 1.80–2.91), and cluster 6 (OR, 2.22; 95% CI, 1.80–2.75) were at significantly higher risk for developing CKD. A total of 933 participants with prediabetes at baseline had incident CVD events during the follow-up. After adjusting for other covariates, cluster 4 had a significantly higher risk of incident CVD events compared with cluster 5 (Figure 4) and also other clusters (Table S3).

Cluster migration patterns from prediabetes clusters at baseline to normal glucose regulation, prediabetes clusters, and T2DM clusters at follow-up

We also performed clustering analysis using the same variables collected at the follow-up examination among the participants diagnosed as prediabetes at follow-up, and similar cluster features were observed (Figure S4). We found that, among the participants who had taken follow-up examination, there were 40.7%, 37.5%, 39.3%, 24.2%, 24.4%, and 13.3% of the participants with prediabetes in clusters 1 to 6 who regressed to normal glucose regulation (NGR), respectively. Among the participants who had migrated from prediabetes clusters at baseline to NGR, prediabetes clusters and T2DM at follow-up, 22.4%, 33.6%, 28.3%, 31.4%, 20.1%, and 25.7% of those with prediabetes in clusters 1 to 6 at baseline still maintained the original cluster type at follow-up (Figure S4, Table S4). We applied K-means clustering analysis among the participants with prediabetes at baseline and progression to T2DM at the follow-up visit using five variables collected at the follow-up, including age at diagnosis, BMI, FPG, HOMA-IR, and HOMA-β. Every patient with T2DM at follow-up was assigned to a predefined cluster named by Ahlqvist and colleagues, including mild age-related diabetes (MARD), mild obesity-related diabetes (MOD), severe insulin-resistant diabetes (SIRD), or severe insulin-deficient diabetes (SIDD) (Figure S5, Tables S4–S6). The Sankey diagram shows the patterns of cluster redistributions and migrations from prediabetes clusters at baseline to diabetes clusters at follow-up (Figure 5). We found a clear pattern of differences between clusters in clinical characteristics and were able to assign the same cluster as Ahlqvist and colleagues did (Table S4). In prediabetes cluster 1 who developed T2DM at follow-up, 99 (31.4%) and 124 (39.4%) of the participants progressed to T2DM MARD cluster and SIDD cluster, respectively. In cluster 2, 232 (64.3%) transformed into the T2DM SIDD cluster; 182 (46.3%) of individuals in prediabetes cluster 3 were assigned to the T2DM SIDD cluster. In individuals with prediabetes of cluster 4, 30.9% and 47.9% were transitioned to the T2DM MOD cluster and SIRD cluster, respectively; 37.3% of the participants in cluster 5 developed into the T2DM MOD cluster; and 34.5% of cluster 6 developed into the T2DM MARD cluster (Figure 5).

Sensitivity analysis

The sensitivity analysis was performed by conducting clustering analysis using the same variables among the participants with normoglycemia and those with prediabetes at baseline. The cluster features in each cluster and the differences in disease risks were similar to those only conducted in prediabetes (Figure S6).

Table 1. Baseline characteristics of study participants in six clusters

	Cluster 1 (n = 13,258, 23.8%)	Cluster 2 (n = 10,836, 19.4%)	Cluster 3 (n = 8664, 15.5%)	Cluster 4 (n = 7246, 13.0%)	Cluster 5 (n = 4310, 7.7%)	Cluster 6 (n = 11,463, 20.6%)
Female	8923 (67.3%)	7825 (72.2%)	5399 (62.3%)	5279 (72.9%)	2094 (48.6%)	7495 (65.4%)
Age (y)	56.08 ± 8.35	56.59 ± 8.45	54.42 ± 8.78	56.12 ± 8.56	54.75 ± 7.90	58.19 ± 8.23
High school or higher education	5032 (38.0%)	3611 (33.3%)	2731 (31.5%)	2722 (37.6%)	1531 (35.5%)	3987 (34.8%)
Married	12,096 (91.2%)	9812 (90.6%)	8036 (92.8%)	6616 (91.3%)	4004 (92.9%)	10,407 (90.8%)
Current smoking	2200 (16.6%)	1511 (13.9%)	1256 (14.5%)	760 (10.5%)	1033 (24.0%)	1525 (13.3%)
Current drinking	1046 (7.9%)	1244 (11.5%)	1161 (13.4%)	441 (6.1%)	869 (20.2%)	1155 (10.1%)
Moderate and vigorous physical activity	2156 (16.3%)	1546 (14.3%)	1118 (12.9%)	981 (13.5%)	540 (12.5%)	1820 (15.9%)
Current drinking tea	3350 (25.3%)	2313 (21.3%)	1894 (21.9%)	1755 (24.2%)	1291 (30.0%)	2957 (25.8%)
Healthy diet	5904 (44.5%)	4192 (38.7%)	3437 (39.7%)	3136 (43.3%)	1787 (41.5%)	5255 (45.8%)
Nighttime sleep duration (h)	7.75 ± 1.25	7.95 ± 1.38	8.00 ± 1.32	7.75 ± 1.26	7.85 ± 1.28	7.77 ± 1.26
Family history of diabetes	1426 (10.8%)	976 (9.0%)	768 (8.9%)	942 (13.0%)	515 (11.9%)	1416 (12.4%)
Taking antihypertensive medicine	1057 (8.0%)	529 (4.9%)	640 (7.4%)	970 (13.4%)	461 (10.7%)	1465 (12.8%)
Taking lipid-lowering medicine	55 (0.4%)	34 (0.3%)	18 (0.2%)	58 (0.8%)	38 (0.9%)	76 (0.7%)
Waist-to-hip ratio	0.88 ± 0.06	0.83 ± 0.06	0.87 ± 0.06	0.91 ± 0.06	0.91 ± 0.06	0.90 ± 0.06
BMI (kg/m^2)	24.12 ± 2.66	21.55 ± 2.36	23.88 ± 2.63	27.80 ± 3.17	25.79 ± 2.95	25.22 ± 2.72
SBP (mm Hg)	127.52 ± 19.04	128.00 ± 19.98	132.35 ± 19.61	135.70 ± 19.06	134.11 ± 19.10	133.28 ± 19.19
DBP (mm Hg)	76.25 ± 10.48	75.36 ± 10.83	79.15 ± 10.66	81.49 ± 10.64	81.58 ± 10.83	79.29 ± 10.31
FPG (mmol/L)	5.12 ± 0.37	5.51 ± 0.46	5.79 ± 0.40	5.53 ± 0.49	5.67 ± 0.49	5.97 ± 0.42
2 h PG (mmol/L)	6.09 ± 1.32	6.61 ± 1.53	7.28 ± 1.56	7.52 ± 1.58	7.52 ± 1.68	8.27 ± 1.44
HbA1c (%)	5.90 ± 0.20	5.82 ± 0.25	5.34 ± 0.27	5.85 ± 0.29	5.78 ± 0.31	6.01 ± 0.22
HOMA-β	76.83 (58.25, 98.80)	46.99 (34.62, 61.88)	56.87 (42.42, 73.88)	132.00 (112.03, 162.86)	78.18 (58.10, 100.00)	61.85 (47.58, 76.97)
HOMA-IR	1.40 (1.06, 1.77)	1.13 (0.82, 1.49)	1.65 (1.23, 2.13)	3.27 (2.69, 4.02)	2.08 (1.55, 2.68)	1.96 (1.52, 2.45)
HDL-c (mmol/L)	1.24 ± 0.26	1.75 ± 0.31	1.34 ± 0.29	1.20 ± 0.28	1.30 ± 0.30	1.22 ± 0.28
LDL-c (mmol/L)	2.83 ± 0.84	2.96 ± 0.82	2.79 ± 0.80	3.03 ± 0.87	3.05 ± 0.90	2.92 ± 0.86
Total cholesterol (mmol/L)	4.74 ± 1.07	5.30 ± 0.98	4.80 ± 1.01	5.13 ± 1.08	5.31 ± 1.08	4.93 ± 1.11
TG (mmol/L)	1.25 (0.92, 1.69)	0.95 (0.75, 1.24)	1.22 (0.90, 1.66)	1.84 (1.35, 2.58)	1.96 (1.37, 2.88)	1.49 (1.10, 2.06)
ALT (U/L)	12.10 (9.00, 17.00)	13.00 (10.00, 17.00)	14.00 (10.00, 18.00)	17.65 (13.00, 23.00)	36.00 (29.00, 45.00)	14.00 (11.00, 19.00)
AST (U/L)	19.37 ± 5.23	22.47 ± 6.08	20.52 ± 5.55	21.49 ± 6.02	35.76 ± 10.52	19.81 ± 5.10
GGT (U/L)	17.00 (13.00, 24.00)	17.00 (13.00, 24.00)	19.00 (14.00, 28.00)	25.00 (19.00, 36.00)	53.00 (33.00, 88.00)	20.00 (15.00, 29.00)
ACR mg/g	5.28 (3.44, 8.48)	5.82 (3.79, 9.50)	5.47 (3.45, 9.03)	5.87 (3.63, 10.11)	5.53 (3.52, 9.43)	5.66 (3.49, 9.41)
eGFR mL/min/ 1.73 m^2	97.45 (89.60, 103.60)	95.91 (88.33, 101.81)	97.93 (90.04, 104.58)	94.98 (86.10, 101.96)	95.71 (87.23, 102.05)	94.88 (86.46, 101.25)

Data are expressed as mean ± SD, number (percentage), or median (interquartile range).
ACR, albumin-to-creatinine ratio; ALT, alanine transaminase; AST, aspartate aminotransferase; BMI, body mass index; DBP, diastolic blood pressure; eGFR, estimated glomerular filtration rate; FPG, fasting glucose; GGT, glutamyl transferase; HDL-c, high-density lipoprotein cholesterol; HOMA-β, homoeostasis model assessment β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance; LDL-c, low density lipoprotein cholesterol; SBP, systolic blood pressure; TG, triglyceride; 2 h PG, 2-h post-load plasma glucose.

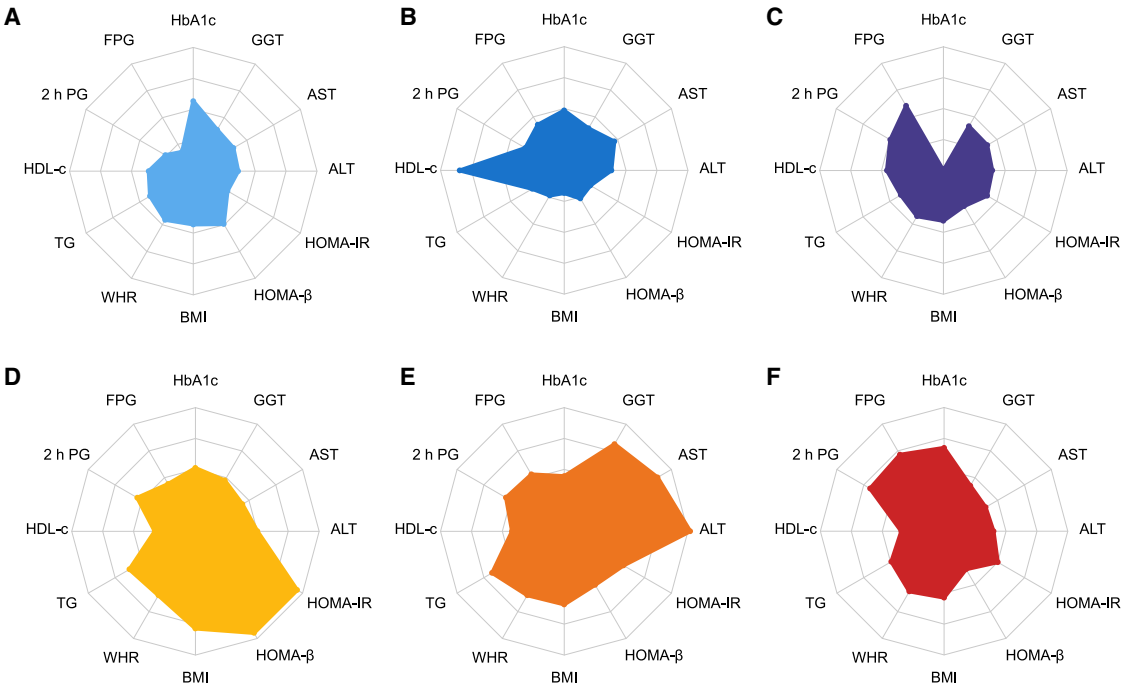


Figure 2. Distribution of the cluster feature variables by clusters

All the values of cluster features were centered to a mean value of 0 and an SD of 1. All the negative values were converted to positive values by adding a fixed value to yield polygon areas related to adverse variable effects. (A) Cluster 1; (B) Cluster 2; (C) Cluster 3; (D) Cluster 4; (E) Cluster 5; (F) Cluster 6. ALT, alanine transaminase; AST, aspartate aminotransferase; BMI, body mass index; FPG, fasting plasma glucose; GGT, γ-glutamyl transpeptidase; HDL-c, high-density lipoprotein cholesterol; HOMA-β, homoeostasis model assessment β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance; TG, triglyceride; WHR, waist-to-hip ratio; 2 h PG, 2-h post-load plasma glucose.

See also Table S1.

DISCUSSION

In this Chinese population with prediabetes, we discovered that the data-driven clusters were reproducible and could be classified into six clusters of distinct metabolic profiles. Future risks of diabetes varied between clusters. Three of the identified subphenotypes had increased CKD risk, including cluster 1 characterized by a single high level of HbA1c, cluster 4 characterized by obesity and insulin resistance, and cluster 6 characterized by high glycemic levels. But only prediabetes in cluster 4 conferred a higher risk of CVD compared with others. Nearly 40% of those with prediabetes in clusters 1, 2, and 3; 24% of those with prediabetes in clusters 4 and 5; and 13% of those with prediabetes in cluster 6 would reverse to NGR. Approximately 20%–30% of each prediabetes cluster would maintain the original prediabetes status during 3 years of follow-up. Finally, for participants with prediabetes who had developed T2DM during follow-up, there were apparent and specific trends of transitions from different clusters of prediabetes into the Ahlqvist classification of diabetes. Those with prediabetes in clusters 1, 2, and 3 mostly progressed to the T2DM MARD cluster and SIDD cluster, while cluster 4 and cluster 5 progressed to the T2DM MOD cluster and SIRD cluster.

The glucose tolerance test was reported to be more sensitive in identifying individuals who were at high risk for prediabetes and diabetes in Asian individuals.[13] Therefore, despite having a relatively higher level of HbA1c in those with prediabetes of cluster 1, the lowest levels of both FPG and 2 h PG might contribute to the lowest incidence of T2DM in cluster 1. Even though T2DM incidence was slightly higher in cluster 2 compared with cluster 1, the CKD risk was substantially lower in cluster 2. A high level of HDL-c might exert the protective effects of preventing CKD in cluster 2. Normal HDL-c is involved in maintaining endothelial function and nitric oxide production, which is critical in preserving tissue perfusion and preventing leukocyte adhesion and infiltration, while deficiency and dysfunction of HDL-c contribute to the severity of oxidative stress and inflammation and promote the progression of CKD.[14] Cluster 3 also had an increased incidence of T2DM compared with cluster 1 due to a higher level of FPG and 2 h PG, and the risk of incident CKD was significantly lower than that in cluster 1. The findings implied that the high HbA1c level was more closely associated with incident CKD than high levels of FPG and 2 h PG.

Previous studies have suggested individuals at the highest risk of developing diabetes are those with FPG of 6.1–6.9 mmol/L or HbA1c of 6.0%–6.4%, or women with a history of gestational diabetes mellitus.[15] Our study made further progress in identifying the individuals with a higher risk of T2DM who were not only with high glycemia levels (cluster 6) but also prediabetes with obesity and insulin resistance (cluster 4), and prediabetes

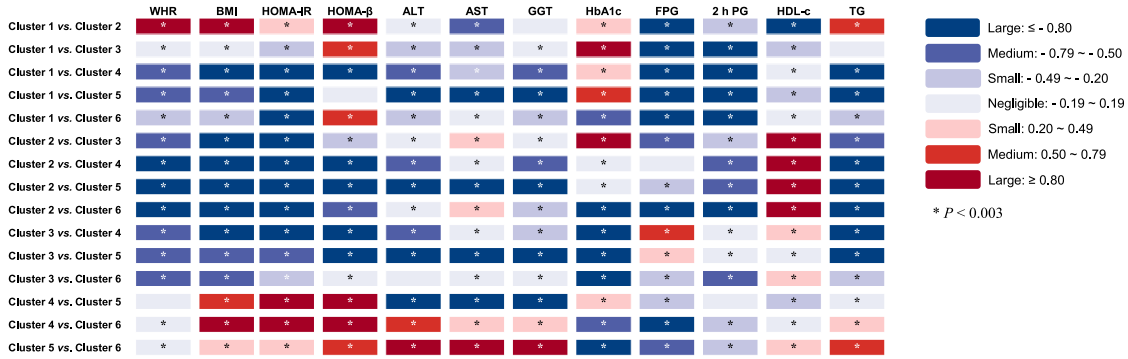


Figure 3. Pairwise comparisons of the cluster feature variables

Bonferroni correction was applied with p < 0.003 (0.05/15) as statistical significance. The blue or red color presented the Cohen's d values that indicated the standardized difference between the two means. FPG, fasting plasma glucose; 2 h PG, 2-h post-load plasma glucose; TG, triglyceride; HDL-c, high-density lipoprotein cholesterol; WHR, waist-to-hip ratio; BMI, body mass index; HOMA-β, homoeostasis model assessment β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance; AST, aspartate aminotransferase; ALT, alanine transaminase; GGT, γ-glutamyl transpeptidase.
See also Table S2.

with elevated liver enzymes and hypertriglyceridemia (cluster 5). Glucose does not seem to be the major driver of CKD and CVD in cluster 4. Prediabetes in cluster 4 represented an insulin-resistant phenotype, in which participants had obesity, and higher levels of HOMA-IR and HOMA-β. It may imply that there is a compensation for insulin resistance through elevated β-cell function in secreting insulin. Previous studies have suggested that even without the well-known risk factors of hypertension or diabetes, obesity per se might be harmful to the kidney by causing hyperfiltration.[16] Insulin resistance also deteriorates kidney function through alterations in hemodynamics, and podocyte and tubular function.[17] Cluster 4 had lower glycemia parameter levels than cluster 6, but presented a higher risk of incident CVD than cluster 6. This result highlighted the effects of insulin resistance, hyperinsulinemia, and attendant lipid disorders on promoting CVD.[18] Prediabetes in cluster 5 was associated with significantly worse liver function. Even though cluster 5 did not present a very high level of blood glucose, it was associated with a higher risk of diabetes. Roy Taylor proposed that T2DM was a result of excess liver fat causing an excess supply of fat to the pancreas with resulting dysfunction of both organs, and leading to diabetes.[19] Although we did not measure the liver fat in our study, the higher BMI and waist-to-hip ratio (WHR) might reveal the higher levels of visceral fat accumulation possessed by the participants in cluster 5.

Similar to previous study,[20,21] we also found a substantial number of individuals reverting from prediabetes to NGR. The Whitehall II study reported that most people with HbA1c-defined prediabetes had persistent prediabetes during 5 years of follow-up.[20] By contrast, reversion to NGR was frequent among people with FPG- or 2 h PG-defined prediabetes.[20] In the present study, we found different proportions of reversion by prediabetes clusters. Approximately 40% of those with prediabetes in clusters 1, 2, and 3 regressed to NGR, whereas the proportions were lower in clusters 4, 5, and 6, which might be because those with prediabetes in clusters 4, 5, and 6 were mostly diagnosed by more than one abnormal glycemic index that indicated worse glycemia metabolism. By performing re-clustering analysis among prediabetes at follow-up, we demonstrated that the prediabetes clusters were not stable, only 20%–30% of the prediabetes would keep the original prediabetes cluster types. The changes of cardiometabolic risk factors during follow-up may have had great impact on the re-classification of prediabetes clusters. In the current study, the data were collected at two time points, making it difficult to observe trends of cluster changes. Temporal trajectory data were needed to further explore the patterns of the cluster changes and the disease risk related to cluster changes.

We found specific trends of transitions from baseline prediabetes clusters to the incident T2DM of Ahlqvist classification. The strong connection between the prediabetes cluster and T2DM cluster, and risk of diabetes complications may suggest the potential value of targeted primary prevention in the management of subphenotypes of prediabetes. For example, a faster progression of renal disease and coronary events was also observed in T2DM clusters of SIRD and MARD in diabetes cluster studies.[22,23] In our study, more than 60% of individuals with prediabetes in cluster 4 who had developed T2DM transitioned to T2DM clusters of SIRD and MARD. Correspondingly, prediabetes in cluster 4 was also associated with a higher risk of incident CKD and CVD.

Prediabetes has been defined by international medical organizations for more than 10 years; however, there was still controversy about defining it as a distinct pathological condition. Prediabetes is characterized by a variety of pathophysiological abnormalities,[24] and the metabolic environments could lead to a broad range of glycemic fluctuations over a continuum with normoglycemia on one side and diabetes mellitus on the other, depending on the stage of process. In the present study, we only analyzed those with prediabetes in view of the priority of diabetes prevention in prediabetes. The early stages of metabolic prototypes could be excluded if only prediabetes was selected. Therefore, we also performed a sensitivity analysis by conducting the clustering analysis among the participants with

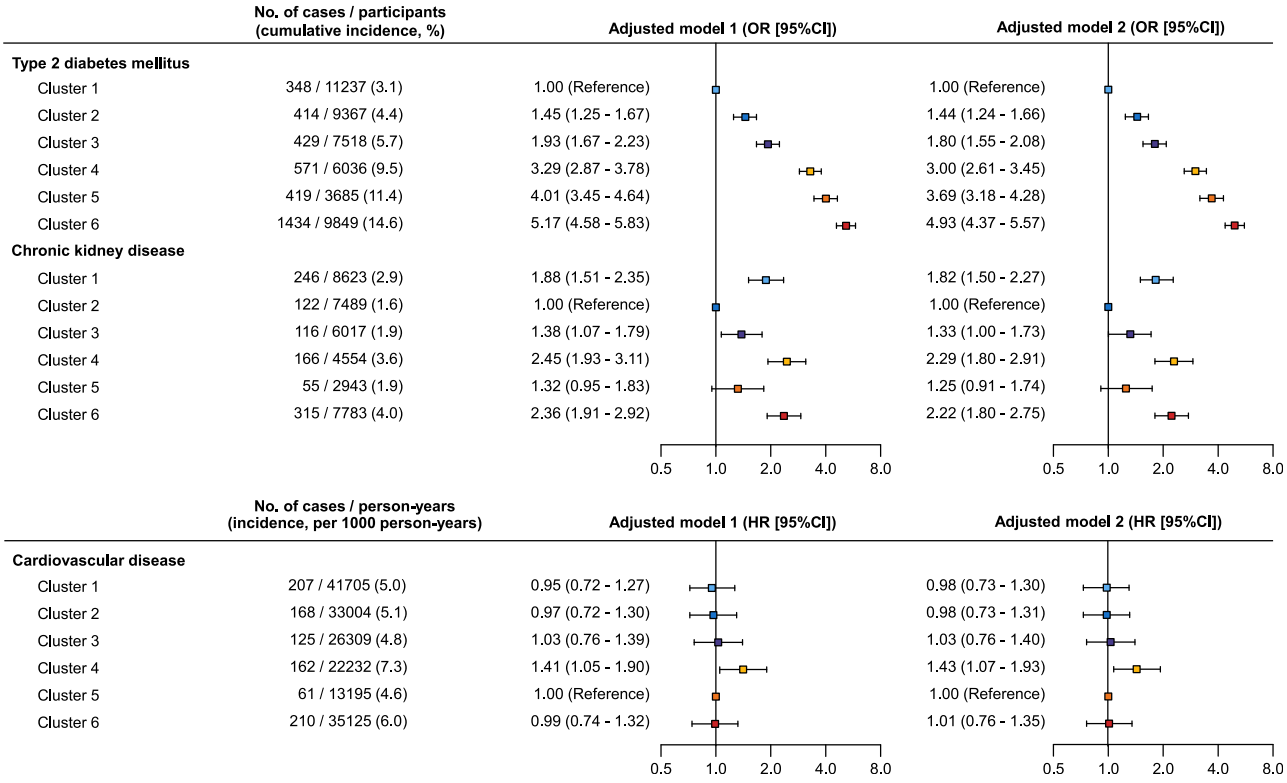


Figure 4. Comparisons of disease risks by clusters

The cumulative incidences of type 2 diabetes and chronic kidney disease were calculated by using the number of incident cases divided by the number of participants at risk in each cluster, the incidence of cardiovascular disease using the number of incident cases divided by the total observation duration (person-years) in each cluster. Incident type 2 diabetes and chronic kidney diseases between clusters were compared by logistic regression models and cardiovascular disease between clusters by Cox proportional hazards models. For each outcome, the cluster with the lowest incidence was used as the reference group. Model 1 was adjusted for age and gender. Model 2 was adjusted for age, gender, education, marital status, smoking status, alcohol consumption, drinking tea, healthy diet score, physical activity, family history of diabetes, nighttime sleep duration, systolic blood pressure, currently taking antihypertensive medication, and taking lipid-lowering agents.

See also Table S3.

normoglycemia and those with prediabetes, and similar results were observed to the findings in prediabetes alone.

Several landmark trials had proved that treating prediabetes with effective interventions could significantly alter the progression of T2DM. However, the effort to implement diabetes prevention in clinical practice has encountered some difficulties. The intervention is always resource intensive, reluctant, and declined if prediabetic individuals are unaware of their hyperglycemia condition or the effects are subtle.[25,26] Our new classification of prediabetes might be useful in identifying the metabolic heterogeneity prior to the onset of T2DM and offering hints for the potential therapeutic implications. For instance, prediabetes in cluster 4 and cluster 6 should be treated with priority due to high risks of T2DM and diabetes complications. The importance of diabetes prevention might be neglected for prediabetes in cluster 5 when risk-stratification concentrated on diabetes-related glycemic cutoffs. In addition, it should be noted that while the method based on widely accessible clinical variables may be informative, it may not be sufficiently precise. Prediabetes could shift from one cluster to another over time. To advance precision medicine for preventing T2DM, the factors associated with subphenotypes shifting should be explored by integrating multidimensional data, such as genetic and omics data, sensor-based behavioral monitoring, and phenomics.[27] Future studies may examine the types of intervention, such as aerobic exercise and dietary caloric restriction, that provide the greatest health advantages for people with prediabetes in various clusters.

Our study has several strengths. Our research with Chinese participants supports the novel prediabetes subgroups proposed by previous study,[9] suggesting a possible generalizability of this European-oriented metabolic classification to various ethnicities and populations. We used data from a large and nationwide prospective cohort of community adults selected from 20 study sites across mainland China, including both urban and rural areas. The large sample size has enabled adequate sample size in each prediabetes cluster. We used a two-step strategy of clustering approach to ensure the stability of the discovered cluster membership and validated the results.

Conclusions

In conclusion, the subphenotypes of prediabetes exhibited unique metabolic traits and were associated with different risks

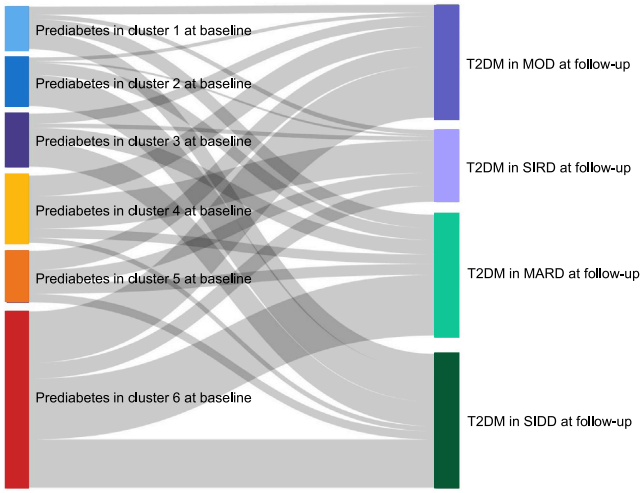


Figure 5. Cluster migration pattern from prediabetes clusters at baseline to those who developed type 2 diabetes of Ahlqvist-diabetes-classes at follow-up

MARD, mild age-related diabetes; MOD, mild obesity-related diabetes; SIDD, severe insulin-deficient diabetes; SIRD, severe insulin-resistant diabetes;. See also Figure S5, Tables S4–S6.

of developing diabetes and its complications. In particular, diabetes risks were significantly different from each other and stepwisely increased from cluster 1 to cluster 6. Three of the identified subphenotypes, including those with a single high HbA1c (cluster 1), obesity and insulin resistance (cluster 4), and high glycemic levels (cluster 6), had increased risks of developing CKD, but only the prediabetes in cluster 4 conferred a higher risk of CVD compared with others. This substratification of prediabetes might help to tailor and target those who would benefit most from early intervention on the risk factors, thereby providing the next step toward precision intervention.

Limitations of the study

Our study has several limitations. First, due to the relatively short period of follow-up duration, the risks of individual CVD components could not be estimated due to the small number of cases of each CVD component. Second, Wagner et al. used insulin secretion calculated by insulin or C-peptide level at 120 min after an oral glucose tolerance test as clustering variables.[9] However, they were not measured in our study. Third, due to a lack of follow-up measurements of the albumin-to-creatinine ratio (ACR), the incidence of CKD was only determined by the estimated glomerular filtration rate (eGFR) and medical history. Fourth, the clustering analysis was based on well-known biomarkers associated with T2DM. Other risk factors such as behavioral factors (smoking status, healthy diet score, and physical activity), genetic factors (polygenic risk score for T2DM), or family history of diabetes were not considered. Finally, the generalizability of our findings was limited to Chinese adults older than 40. The prevalence of prediabetes was also high in children and adolescents[28]; future research may explore the subphenotypes of prediabetes among them to identify the differences from adults.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Study design
  - Data collection
- METHOD DETAILS
  - Definitions of prediabetes
  - Ascertainment of incident outcomes
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2023.100958.

ACKNOWLEDGMENTS

This study was supported by grants from the National Natural Science Foundation of China (grant nos. 81970728, 81970691, 91857205, 82088102, 81930021, 81900741, and 82170819), the Shanghai Medical and Health Development Foundation (grant no. DMRFP_I_01), the Shanghai Outstanding Academic Leaders Plan (grant no. 20XD1422800), the Clinical Research Plan of SHDC (grant no. SHDC2020CR3064B, SHDC2020CR1001A), National Top Young Talents' program (Dr. Yu Xu) Innovative research team of high-level local universities in Shanghai, and the Shanghai Science and Technology Committee (Grant Nos. 20Y11905100 and 19MC1910100).

AUTHOR CONTRIBUTIONS

R.Z., J.L., M.L., Y.X., Y.B., and W.W. conceived and designed the study. R.Z., J.L., M.L., and Y.X. analyzed the data. R.Z. and J.L. drafted the manuscript. M.L., Y.X., and Y.B. revised the manuscript. T.W., Z.Z., M.X., Y.C., M.D., D.Z., X.H., T.Z., X.G., L.C., Y.H., G.Q., Y.L., Q.W., L.C., L.S., R.H., X.T., Q.S., X.Y., Y.Q., G.C., Z.G., G.W., F.S., Z.L., Y.C., Y.Z., C.L., Y.W., S.W., T.Y., Q.L., Y.M., and J.Z. collected the data and critically revised the manuscript for important intellectual content. All authors agreed to be held accountable for all aspects of this work and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 30, 2022
Revised: December 11, 2022
Accepted: February 4, 2023
Published: March 1, 2023

REFERENCES

1. American Diabetes Association (2010). Diagnosis and classification of diabetes mellitus. Diabetes Care *33*, S62–S69.
2. Menke, A., Casagrande, S., Geiss, L., and Cowie, C.C. (2015). Prevalence of and trends in diabetes among adults in the United States, 1988-2012. JAMA *314*, 1021–1029.
3. Li, Y., Teng, D., Shi, X., Qin, G., Qin, Y., Quan, H., Shi, B., Sun, H., Ba, J., Chen, B., et al. (2020). Prevalence of diabetes recorded in mainland China

using 2018 diagnostic criteria from the American Diabetes Association: national cross sectional study. BMJ *369*, m997.

4. Tabák, A.G., Herder, C., Rathmann, W., Brunner, E.J., and Kivimäki, M. (2012). Prediabetes: a high-risk state for diabetes development. Lancet *379*, 2279–2290.
5. Lindström, J., Louheranta, A., Mannelin, M., Rastas, M., Salminen, V., Eriksson, J., Uusitupa, M., and Tuomilehto, J.; Finnish Diabetes Prevention Study Group (2003). The Finnish Diabetes Prevention Study (DPS): lifestyle intervention and 3-year results on diet and physical activity. Diabetes Care *26*, 3230–3236.
6. Gong, Q., Zhang, P., Wang, J., Ma, J., An, Y., Chen, Y., Zhang, B., Feng, X., Li, H., Chen, X., et al. (2019). Morbidity and mortality after lifestyle intervention for people with impaired glucose tolerance: 30-year results of the Da Qing Diabetes Prevention Outcome Study. Lancet Diabetes Endocrinol. *7*, 452–461.
7. Piller, C. (2019). Dubious diagnosis. Science *363*, 1026–1031.
8. Gerstein, H.C., Santaguida, P., Raina, P., Morrison, K.M., Balion, C., Hunt, D., Yazdi, H., and Booker, L. (2007). Annual incidence and relative risk of diabetes in people with various categories of dysglycemia: a systematic overview and meta-analysis of prospective studies. Diabetes Res. Clin. Pract. *78*, 305–312.
9. Wagner, R., Heni, M., Tabák, A.G., Machann, J., Schick, F., Randrianarisoa, E., Hrabě de Angelis, M., Birkenfeld, A.L., Stefan, N., Peter, A., et al. (2021). Pathophysiology-based subphenotyping of individuals at elevated risk for type 2 diabetes. Nat. Med. *27*, 49–57.
10. Wang, T., Lu, J., Shi, L., Chen, G., Xu, M., Xu, Y., Su, Q., Mu, Y., Chen, L., Hu, R., et al.; China Cardiometabolic Disease and Cancer Cohort Study Group (2020). China Cardiometabolic Disease and Cancer Cohort Study Group. Association of insulin resistance and β-cell dysfunction with incident diabetes among adults in China: a nationwide, population-based, prospective cohort study. Lancet Diabetes Endocrinol. *8*, 115–124.
11. Zheng, R., Li, M., Xu, M., Lu, J., Wang, T., Dai, M., Zhang, D., Chen, Y., Zhao, Z., Wang, S., et al. (2021). Chinese adults are more susceptible to effects of overall obesity and fat distribution on cardiometabolic risk factors. J. Clin. Endocrinol. Metab. *106*, e2775–e2788.
12. Lu, J., He, J., Li, M., Tang, X., Hu, R., Shi, L., Su, Q., Peng, K., Xu, M., Xu, Y., et al. (2019). Predictive value of fasting glucose, postload glucose, and hemoglobin A1c on risk of diabetes and complications in Chinese adults. Diabetes Care *42*, 1539–1548.
13. Qiao, Q., Nakagami, T., Tuomilehto, J., Borch-Johnsen, K., Balkau, B., Iwamoto, Y., and Tajima, N.International Diabetes Epidemiology Group; DECODA Study Group (2000). Comparison of the fasting and the 2-h glucose criteria for diabetes in different Asian cohorts. Diabetologia *43*, 1470–1475.
14. Vaziri, N.D. (2016). HDL abnormalities in nephrotic syndrome and chronic kidney disease. Nat. Rev. Nephrol. *12*, 37–47.
15. Davidson, M.B. (2020). Metformin should not be used to treat prediabetes. Diabetes Care *43*, 1983–1987.
16. Chandie Shaw, P.K., Berger, S.P., Mallat, M., Frölich, M., Dekker, F.W., and Rabelink, T.J. (2007). Central obesity is an independent risk factor for albuminuria in nondiabetic South Asian subjects. Diabetes Care *30*, 1840–1844.
17. Artunc, F., Schleicher, E., Weigert, C., Fritsche, A., Stefan, N., and Häring, H.U. (2016). The impact of insulin resistance on the kidney and vasculature. Nat. Rev. Nephrol. *12*, 721–737.
18. Hill, M.A., Yang, Y., Zhang, L., Sun, Z., Jia, G., Parrish, A.R., and Sowers, J.R. (2021). Insulin resistance, cardiovascular stiffening and cardiovascular disease. Metabolism *119*, 154766.
19. Taylor, R. (2021). Type 2 diabetes and remission: practical management guided by pathophysiology. J. Intern. Med. *289*, 754–770.
20. Vistisen, D., Kivimäki, M., Perreault, L., Hulman, A., Witte, D.R., Brunner, E.J., Tabák, A., Jørgensen, M.E., and Færch, K. (2019). Reversion from prediabetes to normoglycaemia and risk of cardiovascular disease and mortality: the Whitehall II cohort study. Diabetologia *62*, 1385–1390.
21. Sallar, A., and Dagogo-Jack, S. (2020). Regression from prediabetes to normal glucose regulation: state of the science. Exp. Biol. Med. *245*, 889–896.
22. Dennis, J.M., Shields, B.M., Henley, W.E., Jones, A.G., and Hattersley, A.T. (2019). Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data. Lancet Diabetes Endocrinol. *7*, 442–451.
23. Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., Vikman, P., Prasad, R.B., Aly, D.M., Almgren, P., et al. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. *6*, 361–369.
24. Echouffo-Tcheugui, J.B., Kengne, A.P., and Ali, M.K. (2018). Issues in defining the burden of prediabetes globally. Curr. Diab. Rep. *18*, 105.
25. Herman, W.H., Hoerger, T.J., Brandle, M., Hicks, K., Sorensen, S., Zhang, P., Hamman, R.F., Ackermann, R.T., Engelgau, M.M., Ratner, R.E., et al. (2005). The cost-effectiveness of lifestyle modification or metformin in preventing type 2 diabetes in adults with impaired glucose tolerance. Ann. Intern. Med. *142*, 323–332.
26. Echouffo-Tcheugui, J.B., and Selvin, E. (2021). Prediabetes and what it means: the epidemiological evidence. Annu. Rev. Public Health *42*, 59–77.
27. Del Prato, S. (2019). Heterogeneity of diabetes: heralding the era of precision medicine. Lancet Diabetes Endocrinol. *7*, 659–661.
28. Han, C., Song, Q., Ren, Y., Chen, X., Jiang, X., and Hu, D. (2022). Global prevalence of prediabetes in children and adolescents: a systematic review and meta-analysis. J. Diabetes *14*, 434–441.
29. Levey, A.S., Stevens, L.A., Schmid, C.H., Zhang, Y.L., Castro, A.F., 3rd, Feldman, H.I., Kusek, J.W., Eggers, P., Van Lente, F., Greene, T., et al. (2009). CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A new equation to estimate glomerular filtration rate. Ann. Intern. Med. *150*, 604–612.
30. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group (2013). KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney Int. Suppl. *3*, 1–150.
31. Sundqvist, H., Heikkala, E., Jokelainen, J., Russo, G., Mikkola, I., and Hagnäs, M. (2022). Association of renal function screening frequency with renal function decline in patients with type 2 diabetes: a real-world study in primary health care. BMC Nephrol. *23*, 356.
32. Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach. Learn. *52*, 91–118.
33. Hennig, C. (2007). Cluster-wise assessment of cluster stability. Comput. Stat. Data Anal. *52*, 258–271.
34. Zou, X., Zhou, X., Zhu, Z., and Ji, L. (2019). Novel subgroups of patients with adult-onset diabetes in Chinese and US populations. Lancet Diabetes Endocrinol. *7*, 9–11.
35. Rosenthal, J.A. (1996). Qualitative descriptors of strength of association and effect size. J. Soc. Serv. Res. *21*, 37–59.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
R version 4.0.3	R-Project	https://www.r-project.org/
STAR Methods. Consensus clustering analysis: R package ConsensusClusterPlus version 1.62.0	Bioconductor	https://www.bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html
STAR Methods. K-means cluster analysis: R package fpc version 2.2-9	R Cran	http://mirrors.ustc.edu.cn/CRAN/
Biological samples		
Serum and urinary samples of the participants	4C BioBank	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Yufang Bi (byf10784@rjh.com.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The patient-level data reported in this study cannot be deposited in a public repository. No new code was generated in this study. Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Study design

The 4C study was a nationwide, multicenter, prospective, population-based study which was designed to examine the relationship between glycemic parameters and clinical outcomes, including diabetes, CVD, and cancer. The study design of the 4C Study has been described in detail previously.[12] From 2011 to 2012, a total of 193,846 adults aged over 40 were recruited from 20 communities located at different geographical regions in mainland China. A comprehensive set of questionnaires, clinical measurements, and laboratory examinations were carried out at the baseline visit. During 2014–2016, all participants were invited to attend an in-person follow-up visit and the same protocol for investigation was used. The study protocol and informed consent were approved by the Medical Ethics Committee of Ruijin Hospital affiliated to the Shanghai Jiaotong University School of Medicine, Shanghai, China. All participants signed the written informed consent.

Data collection

Data were collected from the local hospitals or community clinics at the baseline visit. Trained technicians used a standardized questionnaire to collect participants' demographic characteristics (age, gender, and education), dietary and lifestyle risk factors (including smoking status, alcohol drinking status, physical activity level, healthy diet score, and nighttime sleep duration), and medical history by personal interview. Body weight, height, and waist circumference, and blood pressure were measured by the trained staff. Blood samples were collected after an overnight fast of at least 10 h and the morning urine was also collected. The participants undertook a standard 75 g oral glucose tolerance test and post-load blood samples were collected at 2 h. All participants underwent measurements for FPG, 2 h PG, and HbA1c, TG, total cholesterol (TC), low-density lipoprotein-cholesterol (LDL-c), HDL-c, AST, ALT, GGT, urinary albumin, and urinary creatinine. The HOMA-IR, HOMA-β, and urine ACR were calculated. At the follow-up visit, the information on incident diseases and blood samples were collected, the FPG, 2 h PG, HbA1c, and serum creatinine were tested using the same protocol. Details of specimen processing and data collection were described in Methods S1.

METHOD DETAILS

Definitions of prediabetes

Prediabetes was defined according to the American Diabetes Association 2010 criteria,[1] i.e. in participants without diabetes, FPG between 5.6 mmol/L to 6.9 mmol/L, or 2 h PG between 7.8 mmol/L to 11.0 mmol/L, or HbA1c between 5.7% and 6.4%.

Ascertainment of incident outcomes

Information on the incident diabetes, CKD, and CVD was collected at the follow-up visit. Diabetes was defined as FPG $\geq$7.0 mmol/L, 2 h PG after a 75g glucose load $\geq$11.1 mmol/L, or HbA1c $\geq$ 6.5%, or clinically ascertained diabetes (from a diagnosed history, or taking antidiabetic medications).[1] The eGFR was calculated using the chronic kidney disease epidemiology collaboration (CKD-EPI) equation.[29] The composite outcome of CKD included incident kidney failure requiring dialysis or replacement therapy, death due to renal causes, decline in eGFR defined as eGFR <60 mL/min/1.73 m^2 at follow-up or a certain drop in eGFR category (from eGFR $\geq$90 mL/min/1.73 m^2 at baseline to eGFR in 60–89 mL/min/1.73 m^2 at follow-up) accompanied by a 25% or greater reduction in eGFR from baseline.[30,31] Decline in eGFR was calculated as (eGFR at baseline - eGFR follow-up)/eGFR at baseline×100%. Major CVD events in the present study included the first nonfatal myocardial infarction, nonfatal stroke, hospitalization or treatment for heart failure that occurred during follow-up, and cardiovascular death. The occurrence of nonfatal CVD events was recorded and supporting medical documents were collected. Information on vital status and clinical outcomes was also collected from the local death and disease registries of the National Disease Surveillance Point System and the National Health Insurance System. Two members of the 4C Morbidity and Mortality Adjudication Committee independently adjudicated each clinical event.

QUANTIFICATION AND STATISTICAL ANALYSIS

We used a forward stepwise logistic regression model to select the set of most important biomarkers that were independently associated with T2DM based on the collected anthropometric measurements and biochemical detection. The remaining variables with all $p < 0.05$ were used as clustering variables, including BMI, waist-to-hip ratio (WHR), FPG, 2 h PG, HbA1c, HOMA-IR, HOMA-β, TG, HDL-c, ALT, AST, and GGT. Clustering analysis was done on values centered to a mean value of 0 and an SD of 1. Participants with outlier variables (absolute standardized levels >5) were excluded from the clustering procedure.

We applied a two-step clustering strategy, in which the first step estimated the optimal number of clusters by using consensus clustering analysis (Methods S2).[32] In the second step, we applied K-means clustering specifying six clusters to the overall participants, and each participant was assigned to a unique cluster. K-means clustering was done with a *K* value of 6 using the *kmeansruns* function (runs = 100) in the '*fpc*' package in R (version 4.0.3). To assess the cluster stability, we used bootstrap to perform 500 times random resampling of the overall data and the Jaccard similarities of the original clusters to the most similar clusters in the resampled data are computed. The computation was conducted with the *clusterboot* function from the '*fpc*' package.[33] Generally, stable clusters should yield a Jaccard similarity of greater than 0.75.[33] According to the cluster methods published by Ahlqvist et al.[23] and Zou et al.,[34] we applied the gender-specific cluster analysis among participants with prediabetes at baseline and progression to T2DM at follow-up based on five variables collected at follow-up, including age at diagnosis, BMI, FPG, HOMA-IR and HOMA-β.

Data are presented as the mean (standard deviation), median (interquartile range), or proportion (%). The comparison of mean values between clusters used ANOVA analysis. Skewed data were log-transformed before analysis. There were 15 times pairwise comparisons among 6 clusters and to account for multiple group comparisons, Bonferroni correction was applied with $p < 0.003$ (0.05/15) as statistical significance. We also calculated the Cohen′s effect size (d) to indicate the standardised difference between the two means by removing any influence from the study sample size.[35] The magnitude of the size effect was evaluated using the following cut-off points for classification: small: d = 0.20–0.49; medium: d = 0.50–0.79; large: d $\geq$ 0.80. The cohen′s d was obtained using the *cohen.d* function from the '*effsize*' package in R (version 4.0.3).

The diagnoses of most T2DM and CKD cases were based on the blood glucose and serum creatinine testing at the same time during the baseline and follow-up visits, the exact time of the incident T2DM and CKD was not available. The cumulative incidences of T2DM and CKD using the number of incident cases divided by the number of participants at risk. We compared the incidence of T2DM and CKD at follow-up between baseline clusters using logistic regression. The frequencies of endpoints related to CVD were calculated as the number of events divided by person-years 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 incident CVD by clusters. Odds ratios (ORs) (95%CI) and HRs (95%CI) of model 1 were adjusted for age and gender, and model 2 were adjusted for age, gender, education, marital status, smoking status, alcohol consumption, drinking tea, healthy diet score, physical activity, family history of diabetes, nighttime sleep duration, systolic blood pressure, currently taking antihypertensive medication, and taking lipid-lowering agents. To validate the reliability of the analysis, we also performed clustering analysis using the same method and variables collected at the follow-up examination among the participants diagnosed as prediabetes at follow-up. Another sensitivity analyses were performed by conducting a K-means clustering analysis among the baseline participants with normoglycemia and those with prediabetes, and the risks of diseases were also compared by clusters. All analyses were done using R (version 4.0.3).

Cell Reports Medicine, Volume *4*

Supplemental information

Data-driven subgroups of prediabetes and the associations with outcomes in Chinese adults

Ruizhi Zheng, Yu Xu, Mian Li, Zhengnan Gao, Guixia Wang, Xinguo Hou, Li Chen, Yanan Huo, Guijun Qin, Li Yan, Qin Wan, Tianshu Zeng, Lulu Chen, Lixin Shi, Ruying Hu, Xulei Tang, Qing Su, Xuefeng Yu, Yingfen Qin, Gang Chen, Xuejiang Gu, Feixia Shen, Zuojie Luo, Yuhong Chen, Yinfei Zhang, Chao Liu, Youmin Wang, Shengli Wu, Tao Yang, Qiang Li, Yiming Mu, Jiajun Zhao, Chunyan Hu, Xiaojing Jia, Min Xu, Tiange Wang, Zhiyun Zhao, Shuangyuan Wang, Hong Lin, Guang Ning, Weiqing Wang, Jieli Lu, Yufang Bi, and for the China Cardiometabolic Disease and Cancer Cohort (4C) Study Group

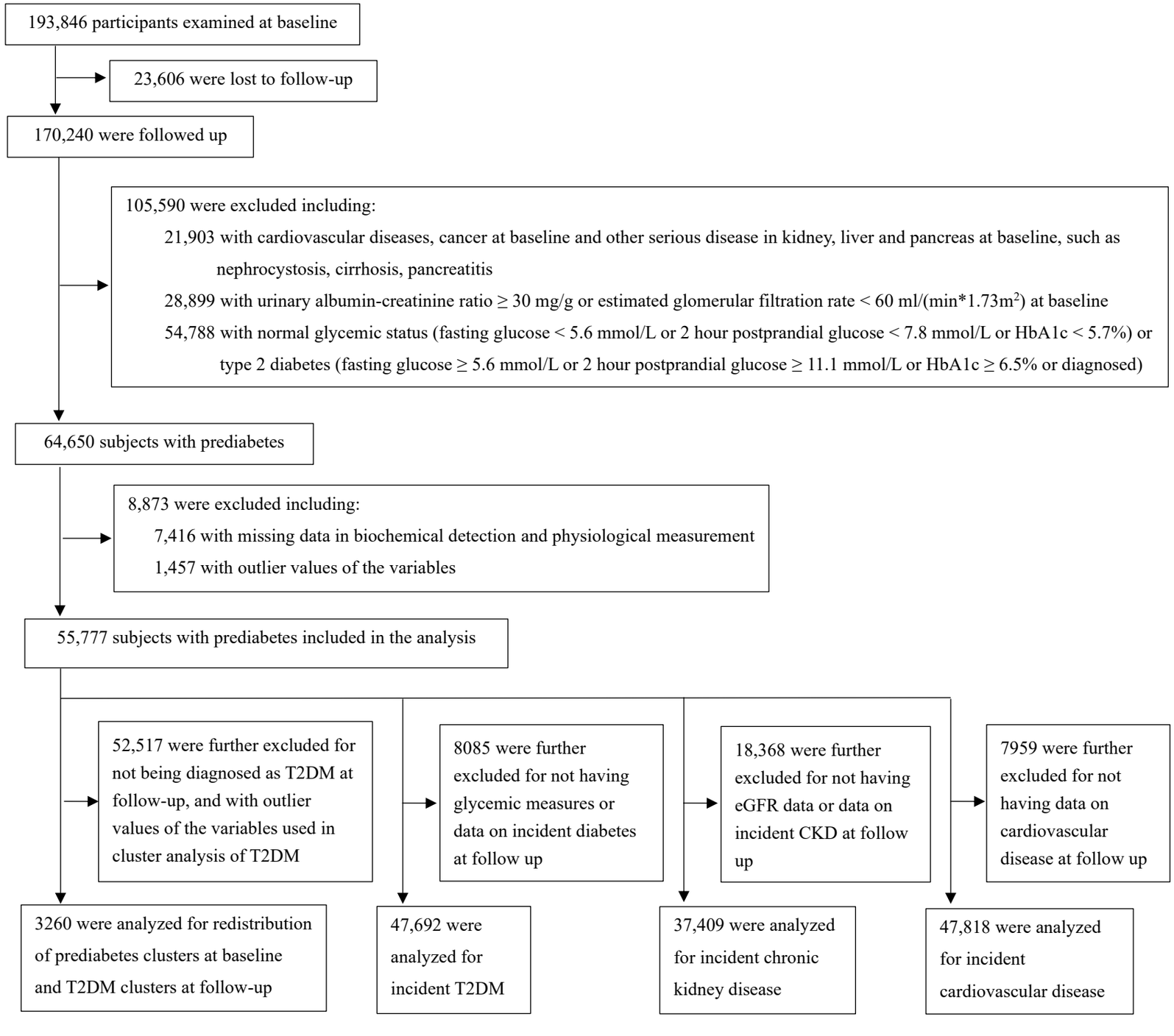
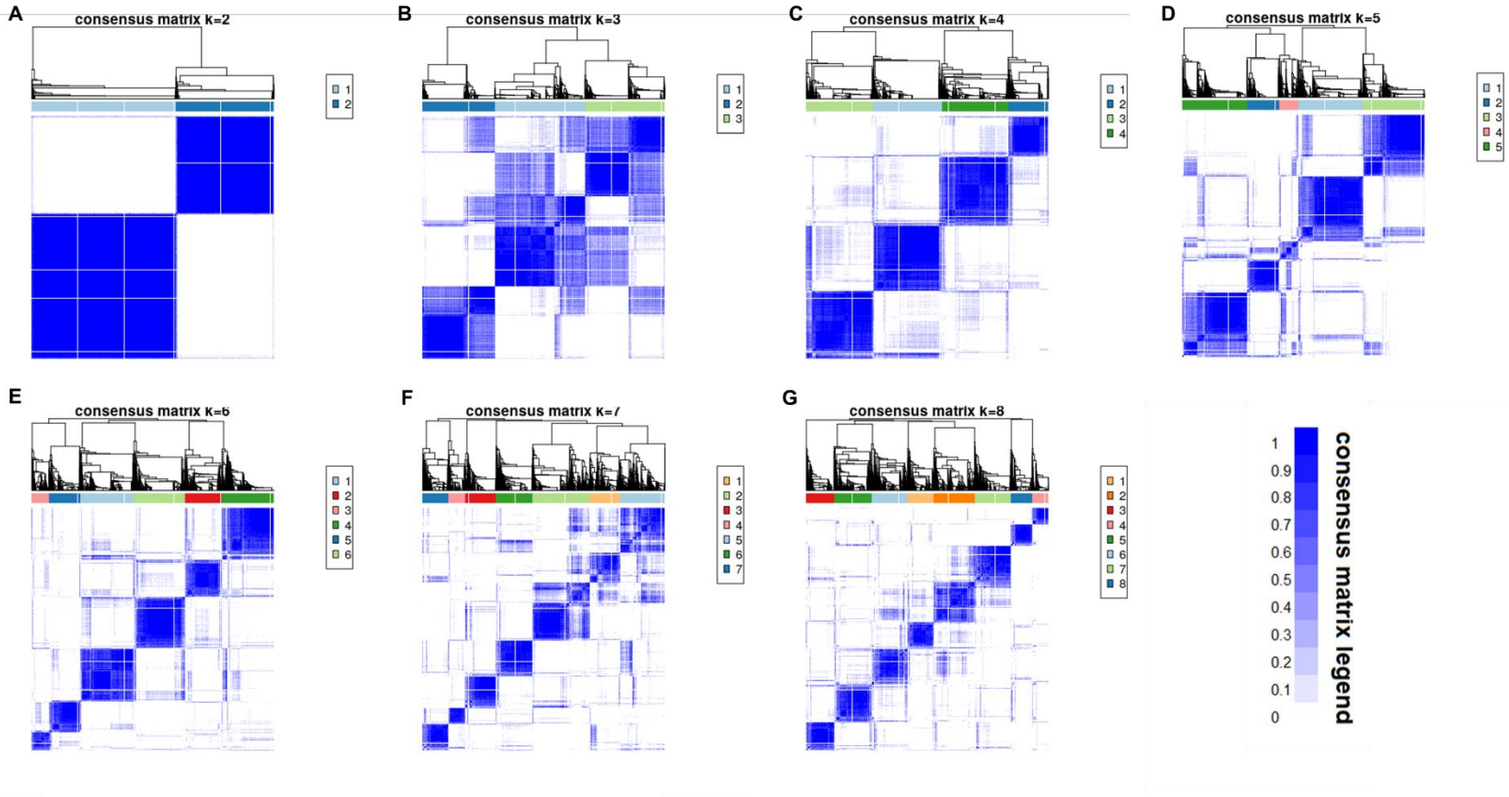
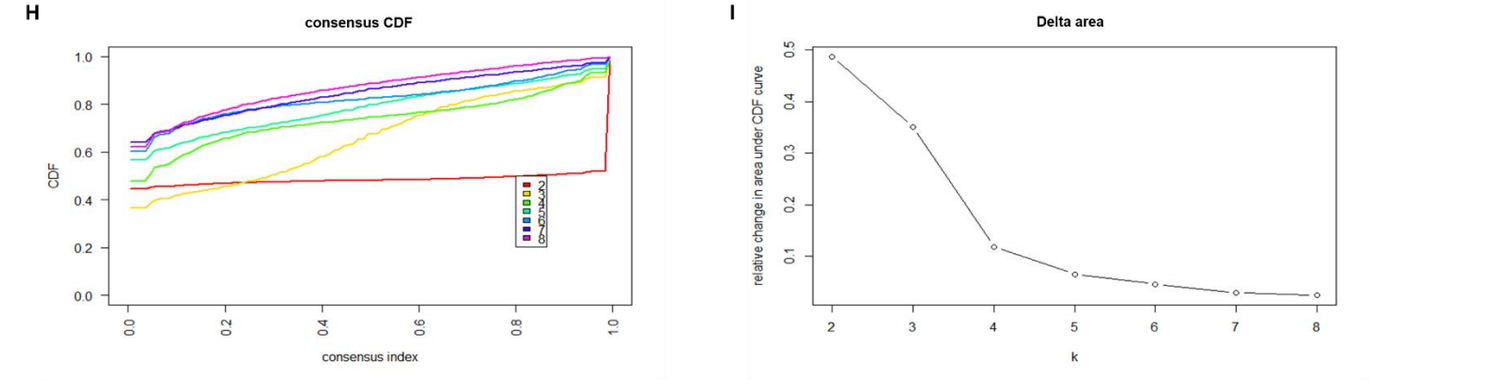


Figure S1. Participants selection flow diagram, related to Table 1


A
consensus matrix k=2
B
consensus matrix k=3
C
consensus matrix k=4
D
consensus matrix k=5
E
consensus matrix k=6
F
consensus matrix k=7
G
consensus matrix k=8
consensus matrix legend


H
consensus CDF
CDF
consensus index
I
Delta area
relative change in area under CDF curve
k

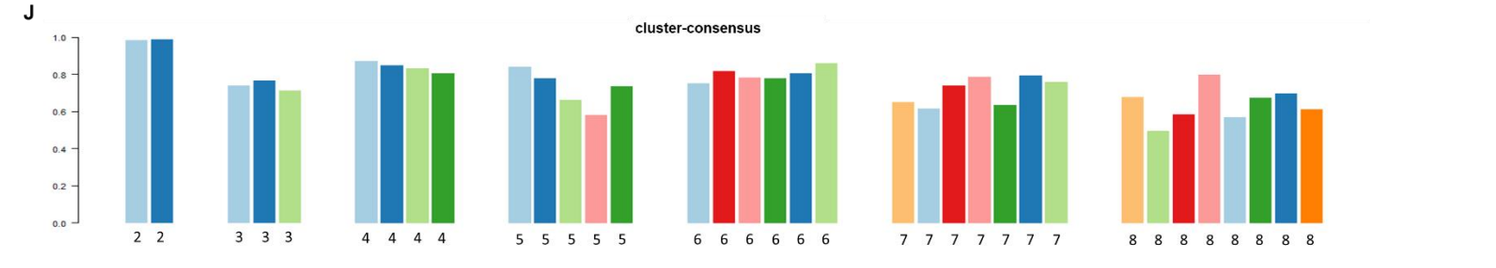

J
cluster-consensus

Figure S2 A-J. Consensus matrix heatmaps using diabetes related factors, consensus cumulative distribution function and cluster consensus score to determine at what number of clusters, related to STAR Methods

(A) *K*=2; (B) *K*=3; (C) *K*=4; (D) *K*=5; (E) *K*=6; (F) *K*=7; (G) *K*=8; (H) The lines by colors indicating the cumulative distribution functions (CDF) of the consensus matrix for each number of clusters; (J) The mean consensus score for different numbers of clusters (*K* ranges from 2 to 8). For *K* = 6, the mean consensus score was 0.75 for cluster 1, 0.82 for cluster 2, 0.78 for cluster 3, 0.78 for cluster 4, 0.80 for cluster 5, and 0.86 for cluster 6.

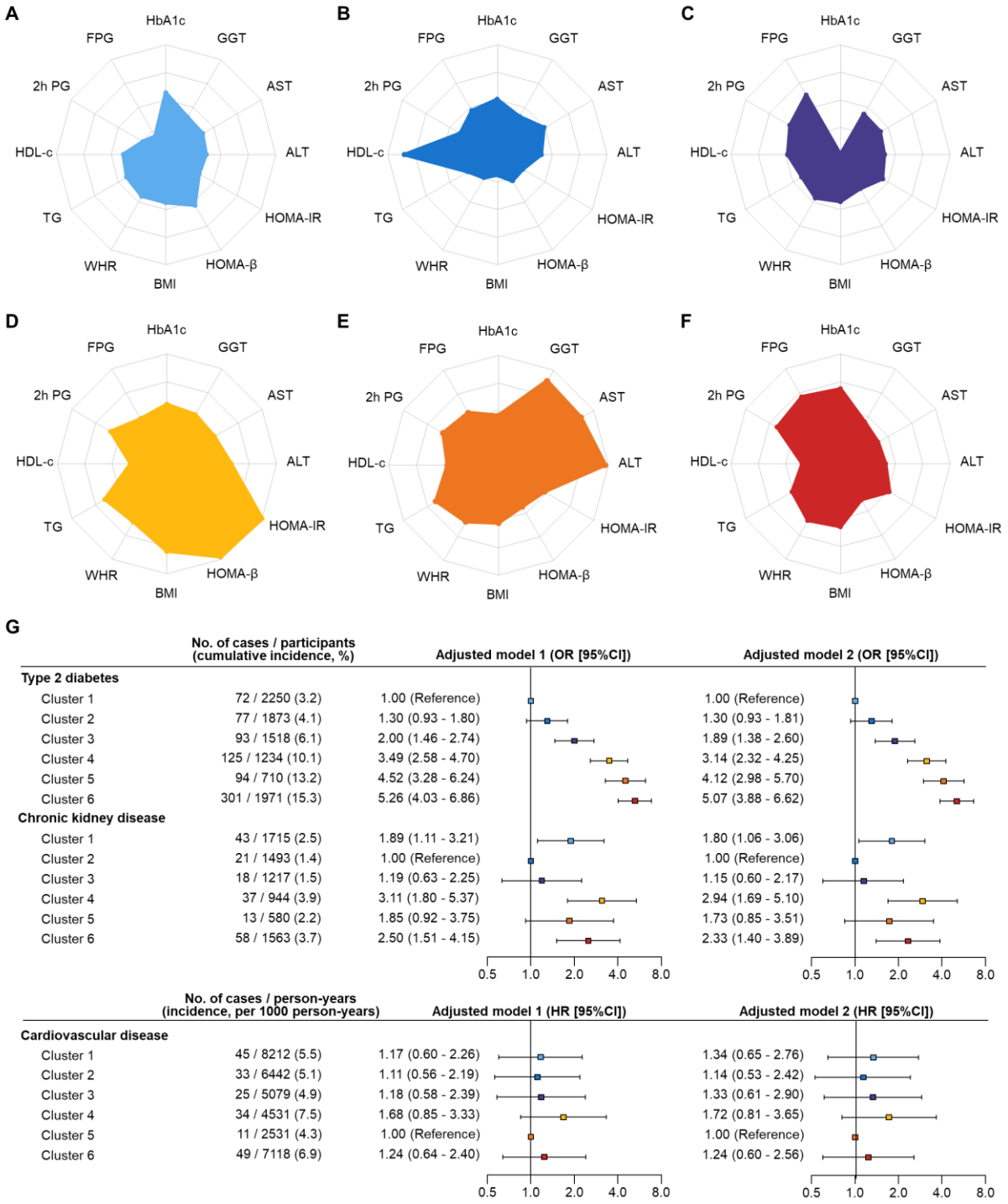


Figure S3 A-G. Characteristics and disease risks of the random sample of participants in diabetes at baseline used to perform consensus clustering algorithm by clusters (n = 11 155), related to STAR Methods

(A) cluster 1; (B) cluster 2; (C) cluster 3; (D) cluster 4; (E) cluster 5; (F) cluster 6; (G) Comparison of incident type 2 diabetes, chronic kidney diseases, and cardiovascular diseases between clusters.

BMI, body mass index; WHR, waist-to-hip ratio; FPG, fasting glucose; 2 h PG, 2-hour post-load plasma glucose; HOMA-β, homoeostasis model assessment β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance; HDL-c, high density lipoprotein cholesterol; TG, triglyceride; AST, aspartate aminotransferase; ALT, alanine transaminase; GGT, γ-glutamyl transpeptidase

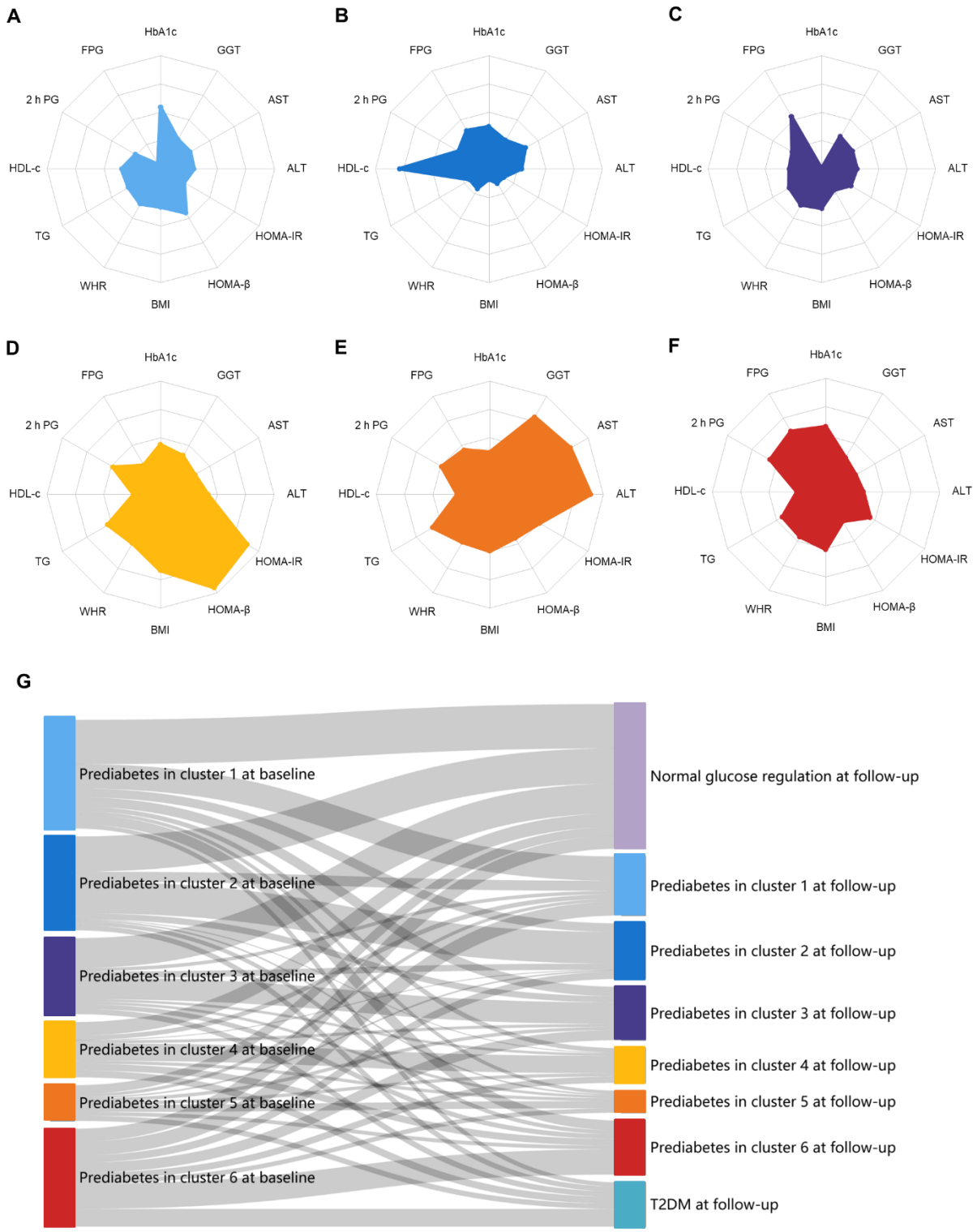


Figure S4 A-G. Distribution of the cluster feature variables among the individuals with prediabetes at the follow-up (A-F) and cluster migration pattern from prediabetes clusters at baseline to normal glucose regulation, prediabetes clusters and type 2 diabetes clusters at follow-up (G), related to Figure 5

(A) cluster 1; (B) cluster 2; (C) cluster 3; (D) cluster 4; (E) cluster 5; (F) cluster 6

BMI, body mass index; WHR, waist-to-hip ratio; FPG, fasting glucose; 2 h PG, 2-hour post-load plasma glucose;

HOMA-β, homoeostasis model assessment of β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance; HDL-c, high density lipoprotein cholesterol; TG, triglyceride; AST, aspartate aminotransferase; ALT, alanine transaminase; GGT, γ-glutamyl transpeptidase

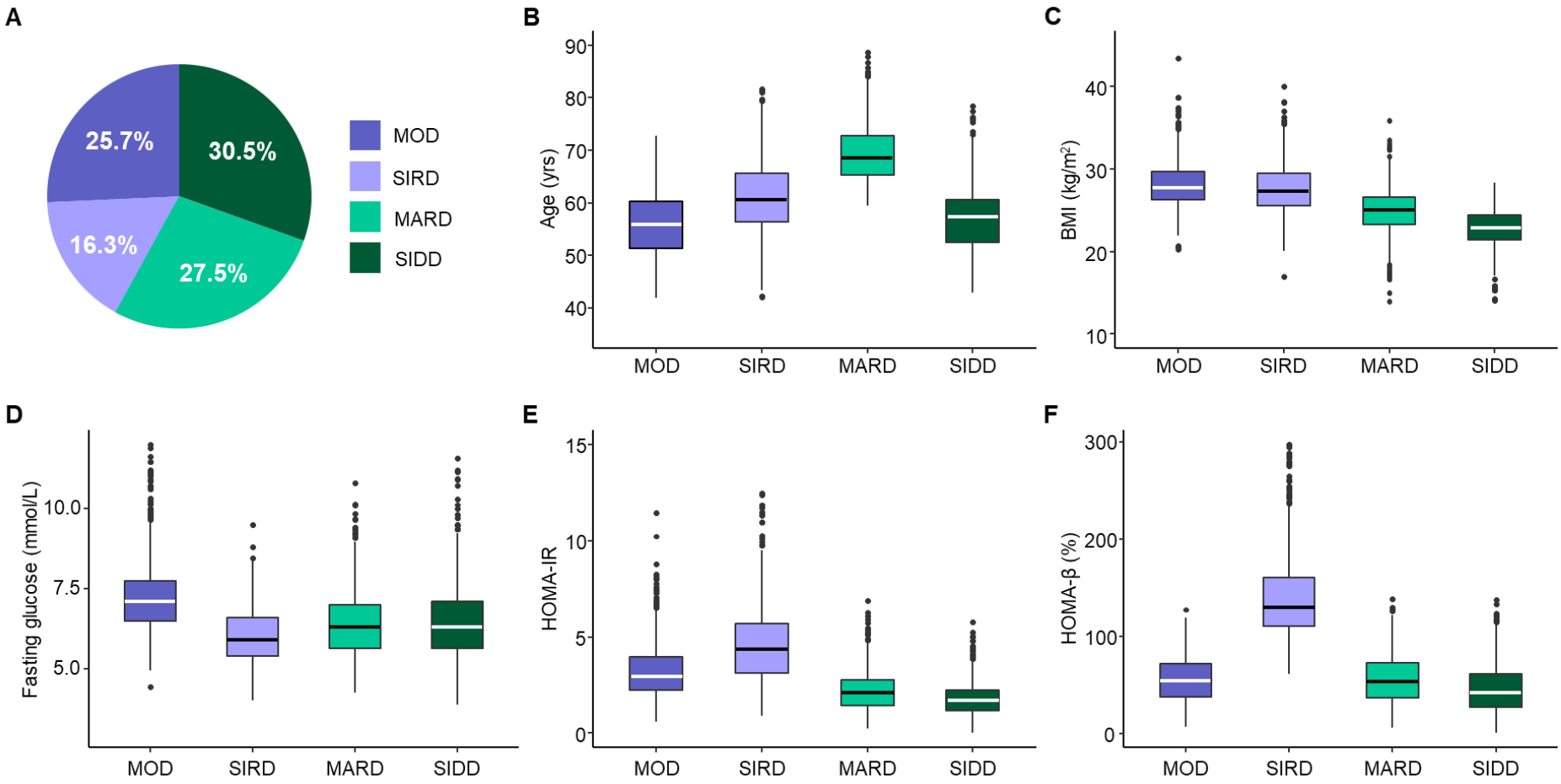


Figure S5 A-F. Characteristics of the subjects developed type 2 diabetes at follow-up clustered by using Ahlqvist-diabetes-classes, related to Figure 5

MARD, mild age-related diabetes; MOD, mild obesity-related diabetes; SIRD, severe insulin-resistant diabetes; SIDD, severe insulin-deficient diabete

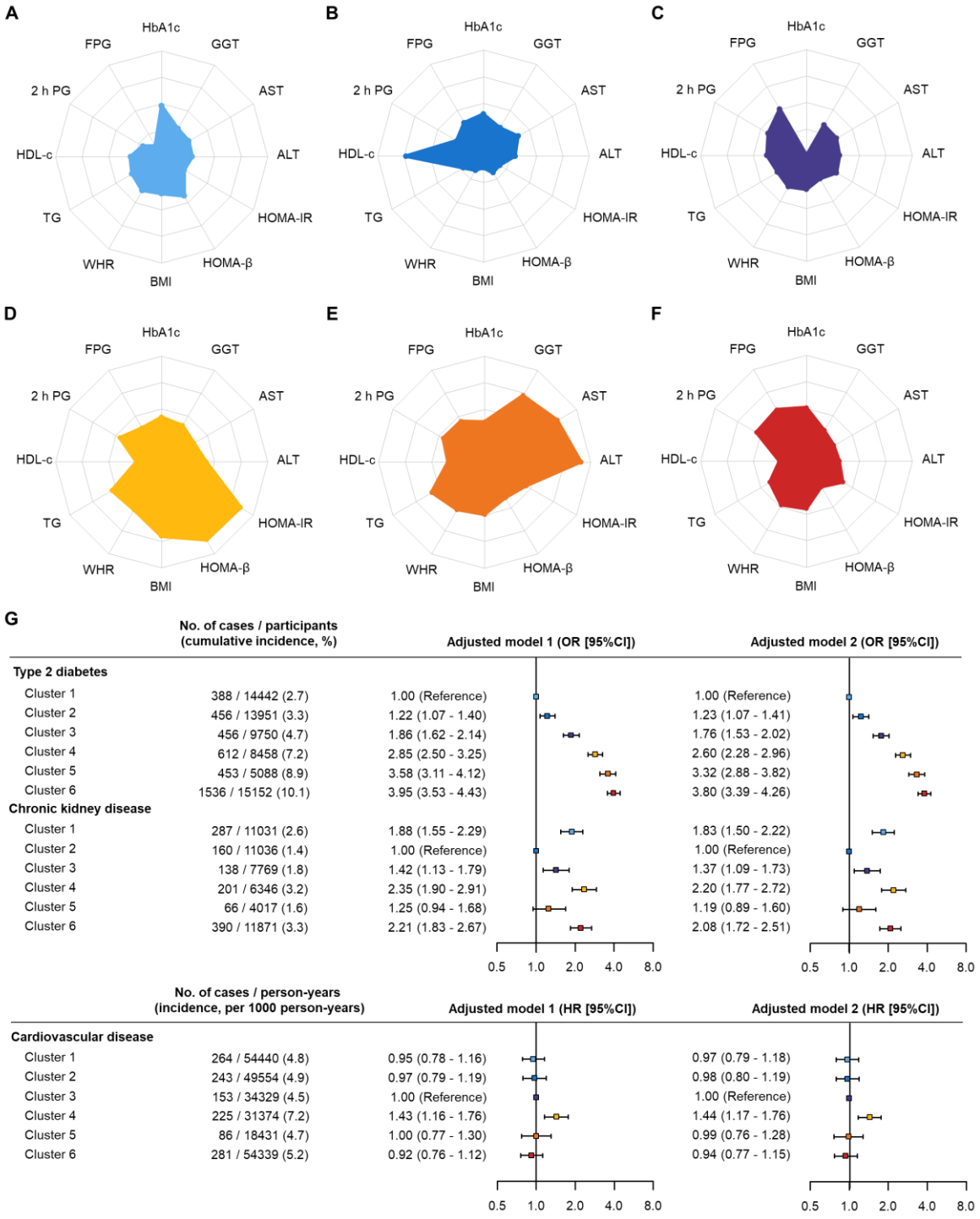


Figure S6 A-G. Characteristics and disease risks of the participants without diabetes defined as fasting plasma glucose < 7.0 mmol/L, 2-hour post-load plasma glucose < 11.1 mmol/L, HbA1c < 6.5% and having not been diagnosed with diabetes at baseline by clusters, related to Figure 2 and 4

(A) cluster 1; (B) cluster 2; (C) cluster 3; (D) cluster 4; (E) cluster 5; (F) cluster 6; (G) Comparison of incident type 2 diabetes, chronic kidney diseases and cardiovascular diseases between clusters.

BMI, body mass index; WHR, waist-to-hip ratio; FPG, fasting glucose; 2 h PG, 2-hour post-load plasma glucose; HOMA-β, homoeostasis model assessment β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance; HDL-c, high density lipoprotein cholesterol; TG, triglyceride; AST, aspartate aminotransferase; ALT, alanine transaminase; GGT, γ-glutamyl transpeptidase

Table S1. Cluster centers using K-means clustering by analysis datasets, related to Figure 2

	WHR	BMI	FPG	2 h PG	HbA1c	HOMA-IR	HOMA-β	ALT	AST	GGT	TG	HDL-c
Clustering analysis in participants with prediabetes at baseline												
Cluster 1	-0.033	-0.103	-0.855	-0.618	0.301	-0.437	0.085	-0.307	-0.273	-0.238	-0.186	-0.316
Cluster 2	-0.695	-0.849	-0.116	-0.307	0.057	-0.664	-0.612	-0.252	0.009	-0.217	-0.504	1.138
Cluster 3	-0.115	-0.175	0.412	0.091	-1.396	-0.172	-0.397	-0.211	-0.168	-0.157	-0.206	-0.027
Cluster 4	0.396	0.961	-0.071	0.232	0.139	1.453	1.449	0.109	-0.08	0.03	0.451	-0.411
Cluster 5	0.396	0.378	0.197	0.235	-0.06	0.263	0.104	1.659	1.217	1.045	0.622	-0.13
Cluster 6	0.289	0.214	0.747	0.675	0.627	0.108	-0.331	-0.197	-0.233	-0.122	0.08	-0.359
Clustering analysis in participants without diabetes at baseline												
Cluster 1	-0.047	-0.098	-0.788	-0.533	0.299	-0.402	0.137	-0.274	-0.252	-0.223	-0.181	-0.260
Cluster 2	-0.682	-0.815	-0.055	-0.271	0.043	-0.620	-0.604	-0.247	-0.011	-0.217	-0.484	1.054
Cluster 3	-0.119	-0.185	0.374	0.110	-1.143	-0.163	-0.386	-0.208	-0.169	-0.155	-0.200	-0.022
Cluster 4	0.419	0.965	-0.038	0.219	0.121	1.441	1.409	0.120	-0.082	0.044	0.479	-0.436
Cluster 5	0.415	0.346	0.188	0.205	-0.009	0.212	0.030	1.583	1.230	1.028	0.579	-0.115
Cluster 6	0.282	0.202	0.536	0.487	0.371	0.045	-0.304	-0.212	-0.248	-0.137	0.046	-0.385
Clustering analysis in participants with prediabetes at the follow-up												
Cluster 1	-0.081	-0.145	-1.009	-0.385	0.481	-0.433	0.194	-0.270	-0.246	-0.244	-0.174	-0.120
Cluster 2	-0.545	-0.870	0.018	-0.221	-0.035	-0.700	-0.711	-0.281	-0.033	-0.278	-0.574	1.203
Cluster 3	-0.050	-0.103	0.449	-0.256	-1.149	-0.260	-0.497	-0.215	-0.211	-0.166	-0.158	-0.297
Cluster 4	0.335	0.860	-0.263	0.295	0.169	1.502	1.693	0.105	-0.099	0.034	0.453	-0.447
Cluster 5	0.305	0.351	0.204	0.317	-0.038	0.371	0.186	1.516	1.306	1.212	0.598	-0.293
Cluster 6	0.221	0.373	0.698	0.552	0.573	0.199	-0.252	-0.159	-0.244	-0.111	0.177	-0.384

BMI, body mass index; WHR, waist-to-hip ratio; FPG, fasting glucose; 2 h PG, 2-hour post-load plasma glucose; HOMA-β, homoeostasis model assessment β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance; HDL-c, high density lipoprotein cholesterol; TG, triglyceride; AST, aspartate aminotransferase; ALT, alanine transaminase; GGT, γ-glutamyl transpeptidase

Table S2. Cohen's d estimates of cluster comparisons of variables in prediabetes at baseline, related to Figure 3

Comparison groups	WHR	BMI	HOMA-IR	HOMA-β	ALT	AST	GGT	HbA1c	FPG	2 h PG	HDL-c	TG
Cluster 1 vs Cluster 2	0.841	1.020	0.466	1.100	-0.114	-0.550	-0.051	0.364	-0.948	-0.371	-1.790	0.618
Cluster 1 vs Cluster 3	0.103	0.094	-0.486	0.730	-0.200	-0.215	-0.195	2.460	-1.740	-0.842	-0.375	0.032
Cluster 1 vs Cluster 4	-0.535	-1.290	-2.710	-1.680	-0.763	-0.384	-0.623	0.228	-0.990	-1.010	0.125	-0.831
Cluster 1 vs Cluster 5	-0.541	-0.609	-1.180	-0.026	-2.980	-2.370	-1.880	0.518	-1.380	-1.020	-0.244	-1.030
Cluster 1 vs Cluster 6	-0.402	-0.409	-0.966	0.671	-0.233	-0.085	-0.283	-0.518	-2.150	-1.580	0.056	-0.383
Cluster 2 vs Cluster 3	-0.734	-0.938	-0.884	-0.423	-0.084	0.332	-0.129	1.860	-0.644	-0.434	1.350	-0.609
Cluster 2 vs Cluster 4	-1.370	-2.310	-2.930	-2.920	-0.637	0.161	-0.511	-0.102	-0.050	-0.585	1.810	-1.400
Cluster 2 vs Cluster 5	-1.390	-1.670	-1.520	-1.220	-2.740	-1.750	-1.670	0.144	-0.356	-0.580	1.440	-1.620
Cluster 2 vs Cluster 6	-1.240	-1.440	-1.340	-0.596	-0.115	0.475	-0.210	-0.800	-1.040	-1.110	1.780	-0.957
Cluster 3 vs Cluster 4	-0.633	-1.360	-2.020	-2.470	-0.551	-0.168	-0.382	-1.810	0.572	-0.151	0.476	-0.814
Cluster 3 vs Cluster 5	-0.641	-0.699	-0.615	-0.804	-2.560	-2.010	-1.510	-1.550	0.263	-0.151	0.125	-0.981
Cluster 3 vs Cluster 6	-0.502	-0.503	-0.438	-0.136	-0.030	0.135	-0.078	-2.740	-0.428	-0.658	0.415	-0.401
Cluster 4 vs Cluster 5	<-0.001	0.652	1.290	1.520	-1.860	-1.780	-1.200	0.219	-0.289	-0.003	-0.347	-0.162
Cluster 4 vs Cluster 6	0.132	0.889	1.710	2.590	0.545	0.307	0.320	-0.642	-0.962	-0.498	-0.066	0.428
Cluster 5 vs Cluster 6	0.133	0.204	0.221	0.775	2.700	2.270	1.580	-0.909	-0.663	-0.491	0.286	0.600

BMI, body mass index; WHR, waist-to-hip ratio; FPG, fasting glucose; 2 h PG, 2-hour post-load plasma glucose; HOMA-β, homoeostasis model assessment β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance; HDL-c, high density lipoprotein cholesterol; TG, triglyceride; AST, aspartate aminotransferase; ALT, alanine transaminase; GGT, γ-glutamyl transpeptidase

Table S3. Comparisons of incident type 2 diabetes, chronic kidney diseases, and cardiovascular disease between clusters by using different clusters as reference groups, related to Figure 4

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Type 2 diabetes						
Reference: cluster 1	1.00 (reference)	1.44 (1.24 - 1.66)	1.80 (1.55 - 2.08)	3.00 (2.61 - 3.45)	3.69 (3.18 - 4.28)	4.93 (4.37 - 5.57)
Reference: cluster 2	0.70 (0.60 - 0.81)	1.00 (reference)	1.25 (1.09 - 1.44)	2.09 (1.83 - 2.39)	2.57 (2.20 - 2.97)	3.44 (3.06 - 3.85)
Reference: cluster 3	0.56 (0.48 - 0.64)	0.80 (0.69 - 0.92)	1.00 (reference)	1.67 (1.46 - 1.90)	2.05 (1.78 - 2.36)	2.74 (2.45 - 3.08)
Reference: cluster 4	0.33 (0.29 - 0.38)	0.48 (0.42 - 0.55)	0.60 (0.53 - 0.68)	1.00 (reference)	1.23 (1.07 - 1.41)	1.64 (1.48 - 1.82)
Reference: cluster 5	0.27 (0.23 - 0.32)	0.39 (0.34 - 0.45)	0.49 (0.42 - 0.56)	0.81 (0.71 - 0.93)	1.00 (reference)	1.34 (1.19 - 1.51)
Reference: cluster 6	0.20 (0.18 - 0.23)	0.29 (0.26 - 0.33)	0.37 (0.33 - 0.41)	0.61 (0.55 - 0.68)	0.75 (0.66 - 0.84)	1.00 (reference)
Chronic kidney disease						
Reference: cluster 1	1.00 (reference)	0.55 (0.44 - 0.69)	0.73 (0.58 - 0.92)	1.26 (1.01 - 1.54)	0.69 (0.51 - 0.93)	1.22 (1.00 - 1.45)
Reference: cluster 2	1.82 (1.50 - 2.27)	1.00 (reference)	1.33 (1.00 - 1.73)	2.29 (1.80 - 2.91)	1.25 (0.91 - 1.74)	2.22 (1.81 - 2.75)
Reference: cluster 3	1.36 (1.10 - 1.71)	0.75 (0.58 - 0.99)	1.00 (reference)	1.72 (1.30 - 2.19)	0.94 (0.68 - 1.30)	1.66 (1.32 - 2.07)
Reference: cluster 4	0.79 (0.65 - 0.98)	0.44 (0.34 - 0.56)	0.58 (0.46 - 0.74)	1.00 (reference)	0.55 (0.40 - 0.75)	0.97 (0.80 - 1.18)
Reference: cluster 5	1.45 (1.08 - 1.96)	0.80 (0.58 - 1.10)	1.06 (0.77 - 1.47)	1.83 (1.34 - 2.49)	1.00 (reference)	1.77 (1.32 - 2.37)
Reference: cluster 6	0.82 (0.69 - 0.97)	0.45 (0.36 - 0.56)	0.60 (0.48 - 0.75)	1.03 (0.85 - 1.25)	0.57 (0.42 - 0.76)	1.00 (reference)
Cardiovascular disease						
Reference: cluster 1	1.00 (reference)	1.00 (0.81 - 1.23)	1.05 (0.84 - 1.31)	1.47 (1.19 - 1.81)	1.02 (0.77 - 1.36)	1.04 (0.86 - 1.26)
Reference: cluster 2	1.00 (0.82 - 1.23)	1.00 (reference)	1.05 (0.83 - 1.33)	1.47 (1.18 - 1.83)	1.03 (0.76 - 1.38)	1.04 (0.85 - 1.28)
Reference: cluster 3	0.95 (0.76 - 1.19)	0.95 (0.75 - 1.20)	1.00 (reference)	1.40 (1.10 - 1.77)	0.97 (0.72 - 1.32)	0.99 (0.79 - 1.24)
Reference: cluster 4	0.68 (0.56 - 0.84)	0.68 (0.55 - 0.85)	0.72 (0.57 - 0.91)	1.00 (reference)	0.70 (0.52 - 0.94)	0.71 (0.58 - 0.87)
Reference: cluster 5	0.98 (0.73 - 1.30)	0.98 (0.73 - 1.31)	1.03 (0.76 - 1.41)	1.43 (1.07 - 1.93)	1.00 (reference)	1.01 (0.76 - 1.35)
Reference: cluster 6	0.96 (0.79 - 1.17)	0.96 (0.78 - 1.18)	1.01 (0.81 - 1.27)	1.41 (1.15 - 1.74)	0.99 (0.74 - 1.31)	1.00 (reference)

The comparisons of incident type 2 diabetes and chronic kidney diseases between clusters used logistic regression and odds ratios were presented. The comparisons of cardiovascular disease between clusters used cox proportional hazards model and hazard ratios were presented.

Table S4. Transitions from prediabetes clusters at baseline to normal glucose regulation, prediabetes clusters and Ahlqvist-diabetes-classes at follow-up, related to Figure 5

Prediabetes at baseline	NGR at follow-up	Prediabetes at follow-up						T2DM at follow-up
		Cluster 1 at follow-up	Cluster 2 at follow-up	Cluster 3 at follow-up	Cluster 4 at follow-up	Cluster 5 at follow-up	Cluster 6 at follow-up	
Cluster 1 at baseline	3263 (40.7%)	1791 (22.4%)	641 (8.0%)	697 (8.7%)	338 (4.2%)	246 (3.1%)	718 (9.0%)	315 (3.9%)
Cluster 2 at baseline	2596 (37.5%)	741 (10.7%)	2324 (33.6%)	451 (6.5%)	54 (0.8%)	106 (1.5%)	288 (4.2%)	361 (5.2%)
Cluster 3 at baseline	2200 (39.3%)	253 (4.5%)	455 (8.1%)	1586 (28.3%)	180 (3.2%)	190 (3.4%)	343 (6.1%)	393 (7.0%)
Cluster 4 at baseline	987 (24.2%)	326 (8.0%)	41 (1.0%)	213 (5.2%)	1279 (31.4%)	229 (5.6%)	495 (12.2%)	501 (12.3%)
Cluster 5 at baseline	643 (24.4%)	220 (8.4%)	108 (4.1%)	242 (9.2%)	235 (8.9%)	528 (20.1%)	274 (10.4%)	381 (14.5%)
Cluster 6 at baseline	962 (13.3%)	1014 (14.1%)	528 (7.3%)	748 (10.4%)	471 (6.5%)	327 (4.5%)	1854 (25.7%)	1309 (18.1%)

Prediabetes at baseline	T2DM at follow-up			
	MOD	SIRD	MARD	SIDD
Cluster 1 at baseline	59 (18.7%)	33 (10.5%)	99 (31.4%)	124 (39.4%)
Cluster 2 at baseline	26 (7.2%)	14 (3.9%)	89 (24.7%)	232 (64.3%)
Cluster 3 at baseline	73 (18.6%)	31 (7.9%)	107 (27.2%)	182 (46.3%)
Cluster 4 at baseline	155 (30.9%)	240 (47.9%)	70 (14.0%)	36 (7.2%)
Cluster 5 at baseline	142 (37.3%)	97 (25.5%)	80 (21.0%)	62 (16.3%)
Cluster 6 at baseline	383 (29.3%)	117 (8.9%)	452 (34.5%)	357 (27.3%)

The number of prediabetes who had taken follow-up examination were 8009, 6921, 5600, 4071, 2631 and 7213 in prediabetes cluster 1 to cluster 6 after excluding those for having missing data in biochemical detection and physiological measurement, and outlier values of the variables.

MARD, mild age-related diabetes; MOD, mild obesity-related diabetes; SIRD, severe insulin-resistant diabetes; SIDD, severe insulin-deficient diabetes, NGR, normal glucose regulation.

Table S5. Ahlqvist-diabetes-classes for the subjects with type 2 diabetes at follow up, related to Figure 5

Clusters	Men		Women		Overall	
	N	%	N	%	N	%
MOD	358	29.8	480	23.3	838	25.7
SIRD	200	16.7	332	16.1	532	16.3
MARD	336	28.0	561	27.2	897	27.5
SIDD	306	25.5	687	33.3	993	30.5

MARD, mild age-related diabetes; MOD, mild obesity-related diabetes; SIRD, severe insulin-resistant diabetes; SIDD, severe insulin-deficient diabetes

Table S6. Cluster center using K-means clustering in subjects with type 2 diabetes at follow up, related to Figure 5

	Age	BMI	FPG	HOMA-IR	HOMA-β
Men					
MOD	-0.831	0.428	0.430	0.098	-0.287
SIRD	-0.144	0.718	-0.356	0.936	1.619
MARD	1.032	0.086	-0.178	-0.301	-0.231
SIDD	-0.070	-1.076	-0.327	-0.643	-0.521
Women					
MOD	-0.396	0.935	0.398	0.289	-0.209
SIRD	0.081	0.524	-0.455	0.893	1.731
MARD	1.064	-0.334	-0.164	-0.343	-0.271
SIDD	-0.635	-0.644	-0.061	-0.497	-0.478

MARD, mild age-related diabetes; MOD, mild obesity-related diabetes; SIRD, severe insulin-resistant diabetes; SIDD, severe insulin-deficient diabetes; BMI, body mass index; FPG, fasting plasma glucose; HOMA-β, homoeostasis model assessment β of cell function; HOMA-IR, homoeostasis model assessment of insulin resistance

Methods S1. Details of specimen testing and data collection, related to Table 1

Fasting and post-load glucose concentrations were measured at local hospitals using the glucose oxidase or hexokinase method. Triglyceride (TG), total cholesterol (TC), low density lipoprotein-cholesterol (LDL-c), high density lipoprotein-cholesterol (HDL-c), aspartate aminotransferase (AST), alanine transaminase (ALT), and glutamyl transferase (GGT) were tested using an auto-analyser (ARCHITECT ci16200, Abbott Laboratories, Abbott Park, IL) at the central laboratory in the Shanghai Institute of Endocrine and Metabolic Diseases (certified by the National Glycohemoglobin Standardization Program and the College of American Pathologists Laboratory Accreditation Program). Finger capillary whole-blood samples were collected using the Hemoglobin Capillary Collection System (Bio-Rad Laboratories, Hercules, CA, USA) and were shipped and stored at 2°C to 8°C. HbA1c was measured within 4 weeks of blood collection by high-performance liquid chromatography using the VARIANT II Hemoglobin Testing System (BioRad Laboratories) at the central laboratory Serum insulin was measured by an autoanalyser (ARCHITECT ci16200, Abbott Laboratories, Chicago, IL, USA). Urinary albumin concentrations were measured at the central laboratory by immunonephelometry using Siemens BNII nephelometers (Siemens Healthcare Diagnostics, Marburg, Germany). The lower limit of detection is 2.13 mg/L. The intra-assay and inter-assay coefficients of variation for urinary albumin were 2.1% and 2.3%, respectively. Urinary creatinine concentrations were measured at the central laboratory by an enzymatic method (ADVIA Chemistry XPT System; Siemens Healthcare, Erlangen, Germany). The intra-assay and inter-assay coefficients of variation for urinary creatinine were 1.1% and 1.3%, respectively. Albuminuria was assessed using albumin-to-creatinine ratio (ACR) based on morning spot urine. Insulin resistance was estimated by the homoeostasis model assessment of insulin resistance (HOMA-IR) index: fasting insulin (μU/mL) × fasting glucose (mmol/L)/22.5 [1]. β-cell function was estimated by the homoeostasis model assessment β of cell function (HOMA-β) index: (20 × fasting insulin [μU/mL])/(fasting glucose [mmol/L] – 3.5) [1].

Blood pressure was measured by the trained staff. Before blood pressure measurement, participants were advised to avoid alcohol, coffee, tea, smoking, and exercise at least 30 minutes. The appropriate cuff was used depending on the subject's arm circumference. An automated electronic device (OMRON Model HEM-725 FUZZY, Omron Company, Dalian, China) was used to measure blood pressure of seated participants three times consecutively at 1-min intervals after a ≥ 5-min rest. The three readings were averaged for analysis.

Standardized questionnaires were used to collect participants' demographic characteristics, dietary, and lifestyle risk factors. Education level was classified by using 9 years of education as the cutoff. Current smoking

was defined as having smoked at least 1 cigarette per day for the past 6 months. Current drinking was defined as drinking alcohol at least once a week for the past 6 months. The International Physical Activity Questionnaire was used to assess physical activity [2]. Moderate and vigorous physical activity was defined as ≥ 150 min/week of moderate-intensity physical activity, or 75 min/week of vigorous aerobic activity, or an equivalent combination of moderate-intensity and vigorous aerobic activities. A food frequency questionnaire was used to collect habitual dietary intake by asking the consumption frequency and portion size of typical food items during the previous 12 months, and a dietary quality score was categorized as high (≥ 4.5 cups per day) and low (< 4.5 cups per day) based on the intake of fruits and vegetables. Nighttime sleep duration was defined as the time space between bedding and waking up.

Methods S2. Consensus clustering algorithm, related to STAR Methods

Consensus clustering is a method of unsupervised cluster has been widely used for high-dimensional data [3]. The clustering algorithm is to maximize the number of clusters meanwhile maintaining high cluster consensus. We set a prespecified number of clusters K=2, 3, …, 8, for each number of clusters, the consensus clustering algorithm created a random subset that included 80% of the data records without replacement and repeated 100 times. For each random subset, K-means (Euclidean distance based) algorithm was performed and each individual was assigned to one of the clusters. After running 100 times, the frequencies of any pair of two individuals were calculated, which were clustered together under each scenario of K and constructed a $N \times N$ matrix of participants' pairwise consensus value (N is the sample size). The cluster membership was determined by applying a hierarchical clustering algorithm using the consensus matrix as a measure of similarity. In the consensus matrix, consensus values range from 0 (never clustered together) to 1 (always clustered together) were marked by white to bright blue. The consensus matrix is ordered by the consensus clustering which is displayed as a dendrogram atop the heatmap. The cluster memberships are marked by colored rectangles between the dendrogram and heatmap with a legend above the graphic.

The optimum number of clusters was ascertained by reviewing the consensus matrix heatmap, cumulative distribution function (CDF) (range 0 – 1) plot and the within-cluster consensus scores. The CDF plot showed the area under the CDFs for each K, and at what 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 also provide the suggestions of the optimum 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 closer to one indicated better cluster stability.

We performed consensus clustering analysis on the random sample containing 20% of the whole participants in prediabetes (n = 11 155). Consensus clustering analysis was done using the ConsensusClusterPlus function (maximum K = 8, replication = 100, proportion of random subset = 0.8, Euclidean distance-based K-means algorithm) in the ‘ConsensusClusterPlus’ package in R version 4.0.3 (http://www.r-project.org).

References

[1] Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia 1985; 28: 412-9.

[2] Craig CL, Marshall AL, Sjostrom M, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003; 35: 1381-95.

[3] Stefano MP, Tamayo; Jill, Mesirov; Todd, Golub. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning 2003; 52: 91-118.