Jia et al. Cardiovascular Diabetology (2026) 25:3
https://doi.org/10.1186/s12933-025-03008-9

Cardiovascular Diabetology

RESEARCH Open Access

Data-driven phenotypic profiling of prediabetes reveals heterogeneous cardiometabolic risks in Chinese adults

Xiaojing Jia[1,2†], Shuangyuan Wang[1,2†], Jinfeng Wang[1,2†], Yilan Ding[1,2†], Mian Li[1,2†], Yiting Lin[1,2†], Ruizhi Zheng[1,2], Feiyue Huang[1,2], Huapeng Wei[1,2], Chunyan Hu[1,2], Yu Xu[1,2], Hong Lin[1,2], Min Xu[1,2], Tiange Wang[1,2], Hong Qiao[3], Guijun Qin[4], Yingfen Qin[5], Xulei Tang[6], Zhen Ye[7], Ruying Hu[7], Lixin Shi[8], Qing Su[9], Xuefeng Yu[10], Li Yan[11], Qin Wan[12], Gang Chen[13], Zhengnan Gao[14], Guixia Wang[15], Feixia Shen[16], Xuejiang Gu[16], Zuojie Luo[5], Li Chen[17], Xinguo Hou[17], Qiang Li[3], Yanan Huo[18], Yinfei Zhang[19], Tianshu Zeng[20], Chao Liu[21], Youmin Wang[22], Shengli Wu[23], Tao Yang[24], Huacong Deng[25], Donghui Li[26], Shenghan Lai[27], Lulu Chen[20], Jiajun Zhao[28], Yiming Mu[29], Guang Ning[1,2], Yufang Bi[1,2*], Weiqing Wang[1,2*], Jieli Lu[1,2*] and for the 4C study author

Abstract

Background The heterogeneous and complex nature of prediabetes presents a major challenge in identifying individuals predisposed to developing incident diabetes and related complications. We aimed to identify phenotypic subgroups of prediabetes at risk and to explore their distinct associations with cardiometabolic outcomes.

Methods This study included 79,000 individuals with prediabetes from the three large-scale prospective cohorts in China. Phenotypic heterogeneity was identified using a soft-clustering algorithm based on the proximity network derived from uniform manifold approximation and projection (UMAP), combined with graph-clustering and Gaussian mixture models. Associations between phenotype probabilities and the incidence of type 2 diabetes (T2D), cardiovascular disease (CVD), and kidney events were assessed to evaluate risk differences across the identified profiles.

Results Six phenotypic profiles were identified, including five with distinct metabolic features (representing ~ 70% of the total population), and one without significant features. These profiles demonstrated substantial differences in both baseline cardiometabolic burden and future disease risk. For instance, individuals with a 20% higher probability of belonging to the hypertensive profile had a 9, 6, and 12% higher risk of T2D, CVD, and CKD, respectively, while the

†Xiaojing Jia, Shuangyuan Wang, Jinfeng Wang, Yilan Ding, Mian Li, Yiting Lin have contributed equally to this work.

*Correspondence:
Yufang Bi
byf10784@rjh.com.cn
Weiqing Wang
wqingw61@163.com
Jieli Lu
jielilu@hotmail.com

Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

profile with high lipids, creatinine, and liver enzyme was associated with an 10% increased risk of T2D and kidney events. Moreover, incorporating phenotypic probabilities into multivariable models significantly improved the prediction of disease risks (likelihood ratio test, $P < 0.05$).

Conclusions Prediabetes exhibits substantial phenotypic heterogeneity, and delineation of distinct metabolic profiles enables refined risk stratification and informs precision prevention strategies.

Keywords Heterogeneity, Prediabetes, Clustering, Diabetes, Cardiovascular disease, Kidney events

Research insights

What is currently known about this topic?

Prediabetes is a heterogeneous condition, but conventional clustering methods struggle to capture its complexity for effective risk stratification. There is a pressing need for more robust, reproducible, and clinically informative subtyping strategies, particularly in large prospective cohorts, to enhance risk stratification and enable personalized preventive interventions.

What is the key research question?

Can a soft-clustering approach identify clinically meaningful phenotypic subgroups of prediabetes with distinct cardiometabolic risk profiles?

What is new?

This study applied a UMAP-based soft-clustering method to identify six phenotypic profiles in a large Chinese cohort, revealing differential risks of diabetes, cardiovascular, and kidney outcomes. Incorporating phenotypic probabilities improved disease risk prediction and supports precision prevention strategies.

How might this study influence clinical practice?

The findings emphasize the heterogeneous nature of prediabetes and demonstrate that specific metabolic phenotypes are differentially associated with progression to diabetes and its complications. Refined phenotypic classification may improve early risk prediction and guide personalized preventive strategies for individuals with prediabetes.

Introduction

Prediabetes is a highly prevalent metabolic condition with substantial global burdens, projected to affect 638 million people with impaired glucose tolerance (IGT) and 414 million people with impaired fasting glucose (IFG) by 2045 [1]. The diagnostic thresholds of prediabetes variable across countries and institutions. Given the defination according to the American Diabetes Association (ADA) criteria, with fasting plasma glucose 5.6–6.9 mmol/L, 2 h plasma glucose 7.8–11.0 mmol/L or HbA1c 5.7–6.4% [2–4], was widely used in both international and national studies, we applied this in current study to ensure comparability with previous research and relevance to the Chinese population. In China, nearly one-third of the population has been reported to have prediabetes [3]. This condition is strongly associated with life-threatening conditions such as type 2 diabetes (T2D), cardiovascular diseases (CVD), and chronic kidney disease (CKD) [5, 6].

The heterogeneous and complex nature of prediabetes presents a major challenge in identifying individuals predisposed to developing incident T2D and its complications. Although early interventions could delay or prevent the onset of diabetes [7, 8], current glucose-based diagnostic criteria for prediabetes are inadequate to capture its pathophysiological heterogeneity or to predict disease progression and complications, with some individuals even persisting in a metabolically stable state over the long-term without progression to overt diabetes [9]. This clinical heterogeneity further implies that prediabetes is not a uniform condition, but rather comprises multiple biologically distinct subtypes, whose underlying mechanisms and long-term risks remain incompletely understood.

Conventional data-driven clustering approaches are often more challenging due to the model selection, analytical frameworks and parameter estimations [10]. These limitations are particularly evident in biological systems, where variables often exhibit nonlinear dependencies. Recent studies have proposed several novel classifications for prediabetic population [4, 6, 9, 11]. For example, using a k-means clustering method, Wagner et al. identified six distinct clusters based on clinical and genetic variables to reveal pathophysiological heterogeneity of prediabetes [6]. Building upon this, our previous work further stratified Chinese individuals with prediabetic into six subgroups with distinct disease risk profiles [4], underscoring the substantial heterogeneity within the prediabetic state. These findings suggest that multidimensional phenotypic stratification could enhance disease risk assessment [12]. However, conventional clustering methods remain limited in their ability to capture within-cluster differences, non-linear cluster boundaries, and the high-dimensional complexity in biological data [10, 13]. It is therefore critical to develop and validate more robust, reproducible, and clinically meaningful subtyping strategies, especially in large prospective cohorts, to better clarify individual risk profiles and facilitate personalized therapeutic prevention for prediabetes.

To address this gap, we applied a combination of unsupervised clustering approach integrating uniform

manifold approximation and projection (UMAP), graph-clustering techniques with Gaussian mixture models (GMM), to characterize phenotypic heterogeneity among individuals with prediabetes from a large-scale nationwide population-based Chinese cohort. Rather than categorically assigning individuals to a specific profile, each participant was characterized by a probability distribution across all profiles, reflecting the continuum of metabolic variation. We investigated the clinical and metabolic features of these profiles and their associations with the risk of T2D and related complications in the China Cardiometabolic Disease and Cancer Cohort (4C) study. The robustness and generalizability of the profiles were further confirmed in two independent validation cohorts. This study provides novel insights into the phenotypic architecture behind prediabetes, and establishes a conceptual and methodological framework to inform early risk stratification and personalized preventive strategies.

Methods

Ascertainment of baseline glucose tolerance status

Prediabetes was defined according to the 2010 American Diabetes Association criteria [2]. Among participants without diabetes, prediabetes was defined as having fasting plasma glucose (FPG) of 100–125 mg/dL (5.6–6.9 mmol/L), or 2 h postprandial plasma glucose (2 h-PG) level of 140–199 mg/dL (7.8–11.0 mmol/L), or HbA1c level of 5.7–6.4%.

Study subjects

To explore the heterogeneity of prediabetes, we used a nationwide data from the 4C Study as our discovery dataset, and two independent population-based cohorts: the Xinjiang Study (XJ_2011) and the Songnan Study (SN_2009) as validation datasets (Fig. 1).

Discovery dataset: The 4C is a nationwide, population-based, prospective cohort study. The study design of the 4C Study has been described in detail previously [14, 15]. A total of 193,846 adults aged 40 years or older were recruited during 2011–2012, and 170,240 (87.8%) participants attended an in-person follow-up visit between 2014 and 2016. For the current analysis, 88,998 participants with prediabetes at baseline were initially eligible. Participants with missing data on any of the nine clustering variables (N = 15,773) or with extreme outlier values (> 5 SD from the mean; N = 1345) were excluded. Ultimately, 71,880 individuals with prediabetes were included in the final analysis. The mean follow-up duration was 3.8 years.

Replication dataset 1: XJ_2011 study, a community-dwelling prospective cohort including 9834 individuals aged ≥ 40 years old, was established in Karamay region of Xinjiang, China, from March to June 2011 [16], and followed until November 2021. From this cohort, 5293 adults with prediabetes and complete baseline information were identified. The mean follow-up period was 10.4 years.

Replication dataset 2: SN_2009 study, a community-dwelling prospective cohort including 4012 individuals aged ≥ 40 years old was established in Songnan region of Shanghai, China, from June to August 2009 [17], and followed until November 2021. We identified 1827 adults with prediabetes and complete baseline information from this cohort. The mean follow-up was 11.8 years.

Each of study was approved by the Medical Ethics Committee of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. All study participants provided written informed consent.

Data collection

Trained technicians administered standardized questionnaires to collect demographic characteristics (age and sex), as well as smoking status, and medical history. All study participants had measured blood pressure, body weight, height, waist circumference, and hip circumference according to a standard protocol during the clinic visit. Body mass index (BMI) was calculated as the body weight (in kilograms) divided by height (in meters) squared. Three readings of systolic and diastolic blood pressures (SBP and DBP) were taken from the non-dominant arm of seated participants after at least 5-min rest, and the average was used for analysis. Hypertension was defined as a SBP ≥ 140 mmHg, DBP ≥ 90 mmHg, or treatment with antihypertensive medications.

Blood samples were collected at 0 and 2 h during a 75 g oral glucose tolerance test (OGTT) following an overnight fast of at least 10 h. In the 4C and XJ_2011 study, HbA1c was measured by high-performance liquid chromatography using the VARIANT II Hemoglobin Testing System (Bio-Rad Laboratories). Serum insulin, creatinine, alanine transaminase (ALT), total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C) and triglycerides (TG) were measured using an auto-analyzer (ARCHITECT ci16200, Abbott Laboratories, Chicago, IL, USA). In the SN_2009 study, serum creatinine, ALT, and lipids were measured with an autoanalyzer (ADVIAVR 1650 Chemistry System, Bayer Corporation, Leverkusen, Germany). Serum insulin was measured using an electrochemiluminescence assay (Roche Diagnostics, Basel, Switzerland). Homeostasis model assessment of insulin resistance (HOMA-IR) was calculated as fasting insulin (μIU/mL) × fasting glucose (mg/dL)/405 [18].

Follow-up and outcomes

In the 4C study, T2D was defined as FPG of 7.0 mmol/L or more, 2 h-PG of 11.1 mmol/L or more, or HbA1c of 6.5% or more, or a self-reported previous diagnosis of diabetes by healthcare professionals. In XJ_2011 and

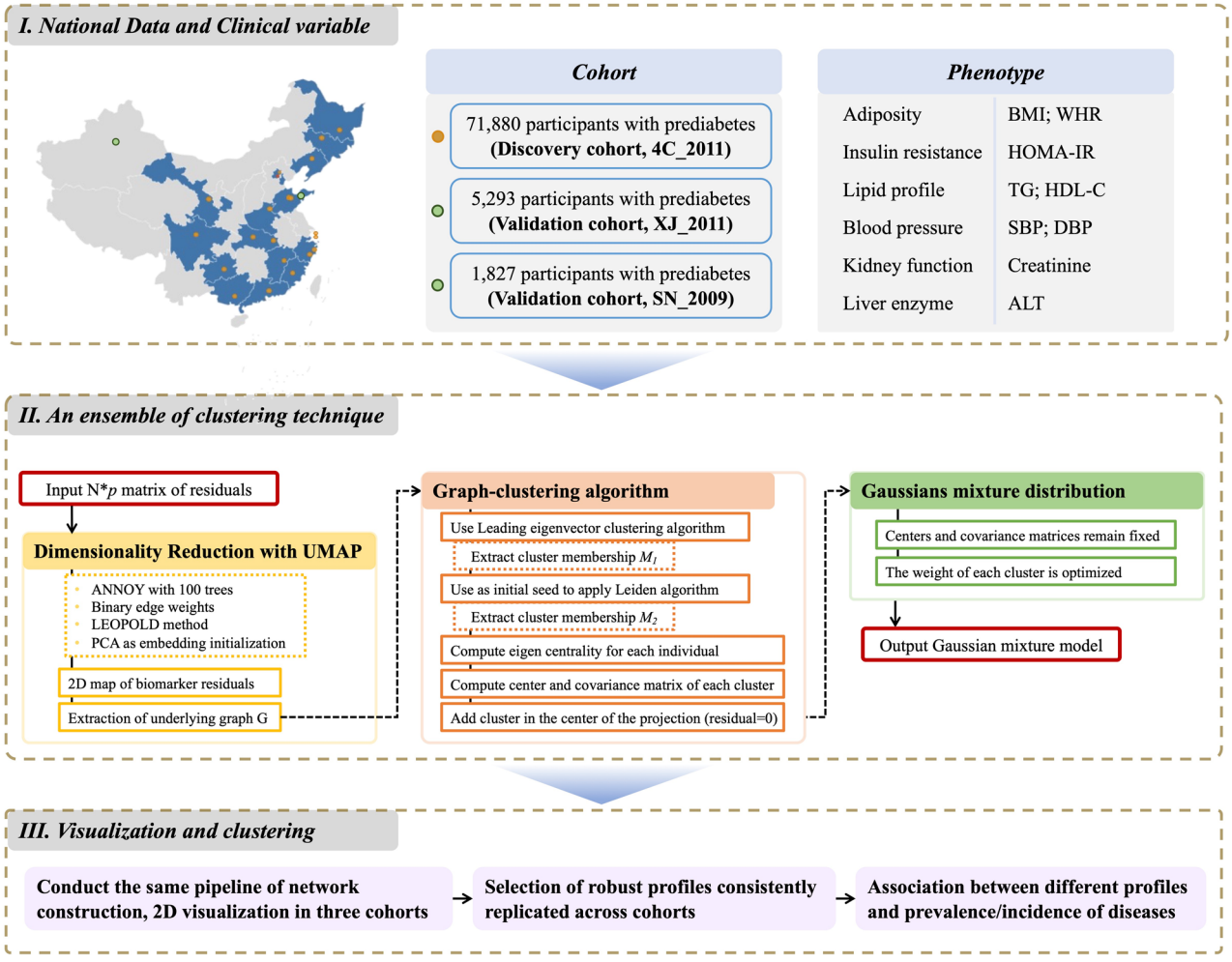


Fig. 1 Study flowchart and model structure. Three independent cohorts were included in the study. The discovery cohort consists of 71,880 individuals with prediabetes from various geographical regions in China. Validation cohort 1 (XJ_2011) included 5293 individuals with prediabetes from Karamay region in Xinjiang, China. Validation cohort 2 (SN_2009) included 1827 individuals with prediabetes from Songnan region in Shanghai, China. Nine clinical variables were retrieved as potential phenotypic biomarkers for the subclassification of prediabetes. The flowchart depicted the overall analytic steps, with details about the ensemble algorithms used to transform biomarker characteristics into probabilistic profiles. PCA, principal components analysis

SN_2009, incident diabetes was defined as physician-diagnosed diabetes based on National Health Insurance System.

Estimated glomerular filtration rate (eGFR) was calculated using the Chronic Kidney Disease-Epidemiology Collaboration (CKD-EPI) equation [19]. In the 4C study, the composite outcome of major adverse kidney events included incident kidney failure requiring dialysis or replacement therapy, death due to renal causes, a decline in eGFR defined as eGFR < 60 mL/min/1.73 m^2 at follow-up, or a certain drop in eGFR defined as a 25% or greater reduction from baseline. In XJ_2011 and SN_2009 study, incident kidney events were defined as death due to renal causes, an eGFR < 60 mL/min/1.73 m^2 at follow-up or a ≥ 25% reduction in eGFR from baseline.

Major CVD events included the first nonfatal myocardial infarction, nonfatal stroke, hospitalization or treatment for heart failure that occurred during follow-up, and cardiovascular death. Information on vital status and clinical outcomes was 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 Morbidity and Mortality Adjudication Committee independently adjudicated each clinical event.

Statistical analysis

Baseline characteristics of participants were described as means (standard deviations [SDs]) or medians (interquartile ranges [IQRs]) for continuous variables and numbers (proportions) for categorical variables.

Clustering feature: We selected nine phenotypic biomarkers routinely used in clinical practice for risk stratification, each representing distinct biological processes

known to influence the progression or remission of prediabetes: BMI and waist-to-hip ratio (WHR), representing adiposity; HOMA-IR, representing insulin resistance; HDL-C and TG, representing lipid profiles; SBP and DBP, representing blood pressure; serum creatinine, representing kidney function; and ALT, representing liver enzyme. For these nine clustering traits, participants with missing values or extreme outliers (>5 SD from the mean) were excluded. We further log-transformed variables with highly skewed distributions (HOMA-IR, TG, ALT, and creatinine) to further avoid the impact of extreme values. Each phenotype was residualized for age and sex using linear regression analysis. The resulting residuals were centered and scaled to a mean of zero and a standard deviation of one, and this standardized matrix was input into the clustering algorithm without any prior dimensionality reduction.

Dimensionality reduction with UMAP projection: To conduct dimensionality reduction, we used the implementation of umap from *uwot* R package, which projects high-dimensional data into a two-dimensional space. To better preserve both the global and local structures from the original high-dimensional data [20], we calculated the number of nearest neighbors (nn) based on the total number of participants in each cohort (N_{total}) through the following equation: $nn = \max(10, 10 + 15 \times (\log_{10}(N_{total}) - 4))$, as well as the "binary_edge_weights" parameter was set to "TRUE". Furthermore, we set the "dens_scale" parameter to 1 to ensure that individuals in densely connected regions appear closer together in the projection [21].

Graph-clustering on UMAP's underlying proximity network: Based on visually discernible subgroups in the UMAP projections, we applied a series of iterative graph-clustering techniques. We used the leading eigenvector algorithm [22] to identify initial seeds, which were then used to run the Leiden algorithm [23], implemented via the igraph R package (v.2.1.2). To identify strongly interconnected regions and optimize the modularity, we performed over 500 iterations of the Leiden algorithm. This process yielded hard partitions, wherein each individual was exclusively assigned to a single cluster. Additionally, for each individual, we computed the normalized eigen centrality scores within their respective clusters to reflect their relative importance in the local network structure.

Gaussian mixture distribution: We subsequently employed these centrality scores as weights in the calculating each cluster's center and covariance matrix, both of which served as parameters for a Gaussian mixture distribution. Notably, clusters located near the center of the low-dimensional projection—where residual values approximated zero—exhibited reduced stability across iterations compared to those at the boundaries. To mitigate this instability, we incorporated a "non-significant" distribution of residuals into the Gaussian mixture calculation, with a zero mean and identity covariance matrix, serving as a stabilizing reference distribution. Consequently, individuals whose profiles were not sufficiently distinct from the "non-significant" profile tended to have similar assignment probabilities across both profiles. As a result, their cluster assignments would lack consistency across iterations and would not be reliably reproduced (see 'Profile replication' section below). To enhance the robustness and interpretability of the final clustering partition, such individuals were reassigned to the "non-significant" profile. In the subsequent phase, we fixed the cluster center and covariance matrices while re-estimating only the cluster weights, which reflect the relative population proportions of each group. The final partition incorporated both "non-significant" and significant profiles, assigning to each individual a probability score for each profile, summing to a total of 1.

Profile replication: To assess the robustness and generalizability of the partitions identified in 4C_2011, we applied the same analytical pipeline to two independent cohorts: XJ_2011 and SN_2009. We evaluated whether individuals who had been confidently assigned to a particular profile in the 4C_2011 model (i.e., with a probability>75%) also exhibited similarly high assignment probabilities (>75%) to a profile in either of the two validation cohorts. A profile was considered successfully replicated if this condition was met in the 4C Study and in at least one validation cohort, thereby ensuring that only consistently observed clusters were retained in the final model. Following this filtering step, we re-estimated the weights of the replicated profiles and restricted all downstream analyses to these validated clusters. The final cluster labels were derived based on the average residual values of the included phenotypes within each cluster.

Connectivity and separation of profiles: To assess the connectivity within each profile, individuals were assigned to the profile for which they had the highest probability. For each resulting profile, we extracted the corresponding subgraph from the UMAP projection and computed the global transitivity index, with higher values indicating greater internal cohesion.

To evaluate the degree of separation between clusters, we calculated the relative entropy (also known as the Kullback–Leibler divergence) [24] of the final partition using the 4C_2011 data. Higher values indicate better separation, where individuals are clearly assigned to distinct clusters without overlap.

Profile-specific information at baseline: For each profile, we calculated summary statistics for all phenotypes by weighting individual values according to their respective profile probabilities. To determine the profile-specific distribution of glycemic status, we first calculated probability-weighted counts (i.e., each individual was weighted

by their profile probability), and then divided by the summing the probability-weighted counts for each glycemic status within the profile.

Profile-specific information at follow-up: For the prevalence of cardiometabolic diseases, we used the same probability-weighted approach, dividing the weighted case counts by the sum of profile probabilities. For incidence of T2D and kidney events, all participants with baseline T2D and kidney diseases were excluded. As the diagnosis of most T2D and kidney diseases cases was based on blood glucose and serum creatinine testing conducted simultaneously at baseline and follow-up visits, the exact time of onset was unavailable. Incidence rates were calculated by each probability-weighted cases and dividing by the total probability for that profile. For CVD, data were censored at the time of CVD diagnosis, death, or the end of follow-up, whichever occurred first. To estimate the incidence of CVD, the denominator was defined as the sum of the products between the follow-up time and the profile probability for each individual. Based on the estimated prevalence and incidence rates, we calculated odds ratios (ORs) and risk ratios (RRs) to compare each markable profile against the "non-significant" reference profile, employing binomial and Poisson regression models, respectively. We derived both study-specific and overall pooled estimates by applying fixed-effects and random-effects meta-analytic models [25]. Given their conservative nature and improved robustness, results from the random-effects models are reported as the primary findings.

Added predictive value of profiles: To evaluate the added predictive value of these profiles, we compared the performance of regression models with and without the inclusion of profile allocation probabilities. The risk of CVD was assessed using Cox proportional hazards models, while logistic regression models were applied for T2D and kidney disease outcomes.

For each outcome, two models were constructed: (1) a basic model incorporating a range of established risk factors, and (2) an extended model included all the risk factors from the base model and additionally including log-transformed ratios of profile allocation probabilities. Due to the inherent sum-to-one constraint of profile probabilities, we divided the allocation probabilities of markable profiles by that of the "non-significant" profile for each individual and used the logarithms of these quotients as predictors. For CVD, the base models incorporated predictors from PREVENT-base [26, 27], as well as phenotypes used in our clustering analysis but not included in PREVENT-base. For T2D and kidney disease, the base models included all variables used in our clustering analysis, along with current smoking status and use of antihypertensive and lipid-lowering medications known to influence the selected phenotypes. The full list of covariates is also presented in Supplementary Table 1. Participants with pre-existing disease were excluded from these models.

To compare model performance within each cohort, we employed likelihood ratio tests [28] and assessed differences in C-statistics. We further quantified the proportion of additional variance explained by incorporating profile probabilities, expressed as a percentage of the total variance explained. This proportion was estimated using model likelihoods, a robust approach not influenced by arbitrary threshold selection.

Profile-specific risk estimates: In the regression models, the estimated effect of each profile reflects the expected change in the outcome resulting from a one-unit increase in the log ratio of that profile, assuming all other log ratios and covariates remain constant. Due to the sum-to-one constraint, this change is equivalent to increasing the probability of the selected profile by a certain factor, while proportionally reducing the probabilities of all other profiles—thus preserving the values of the remaining log ratios. However, changing the probability of a specific profile inevitably alters both the probabilities of the remaining profiles and the phenotype distributions, and vice versa. In this framework, profile log ratios act as interaction terms that modulate the relationship between phenotypes patterns and clinical outcomes, conditional on the pattern of phenotype characteristics. To robustly estimate the effect of redistributing profile probability from the "non-significant" to a selected markable profile—while maintaining all other markable profiles at their population levels—we incorporated into the model the full set of associated changes, including both biomarker variations and adjustments in the log-ratio terms. Profile-specific estimates were derived for each cohort and then pooled using both fixed-effects and random-effects meta-analyses [29].

Results

Study population

The study flowchart and model construction process were shown in Fig. 1. Baseline characteristics are shown in Supplementary Table 2 and Supplementary Fig. 1. Individuals with markable phenotypes tended to cluster into distinct subgroups. We conducted the same analysis in all three independent cohorts separately to identify stable and reproducible profiles (Supplementary Figs. 2 and 3).

Visualization and clustering of profiles

Based on profiles consistently replicated across cohorts (Supplementary Fig. 4), we defined a final partition comprising five profiles with markable phenotype characteristics and one without (Fig. 2). The relative entropy of this final partition was 0.64, indicating moderate separation

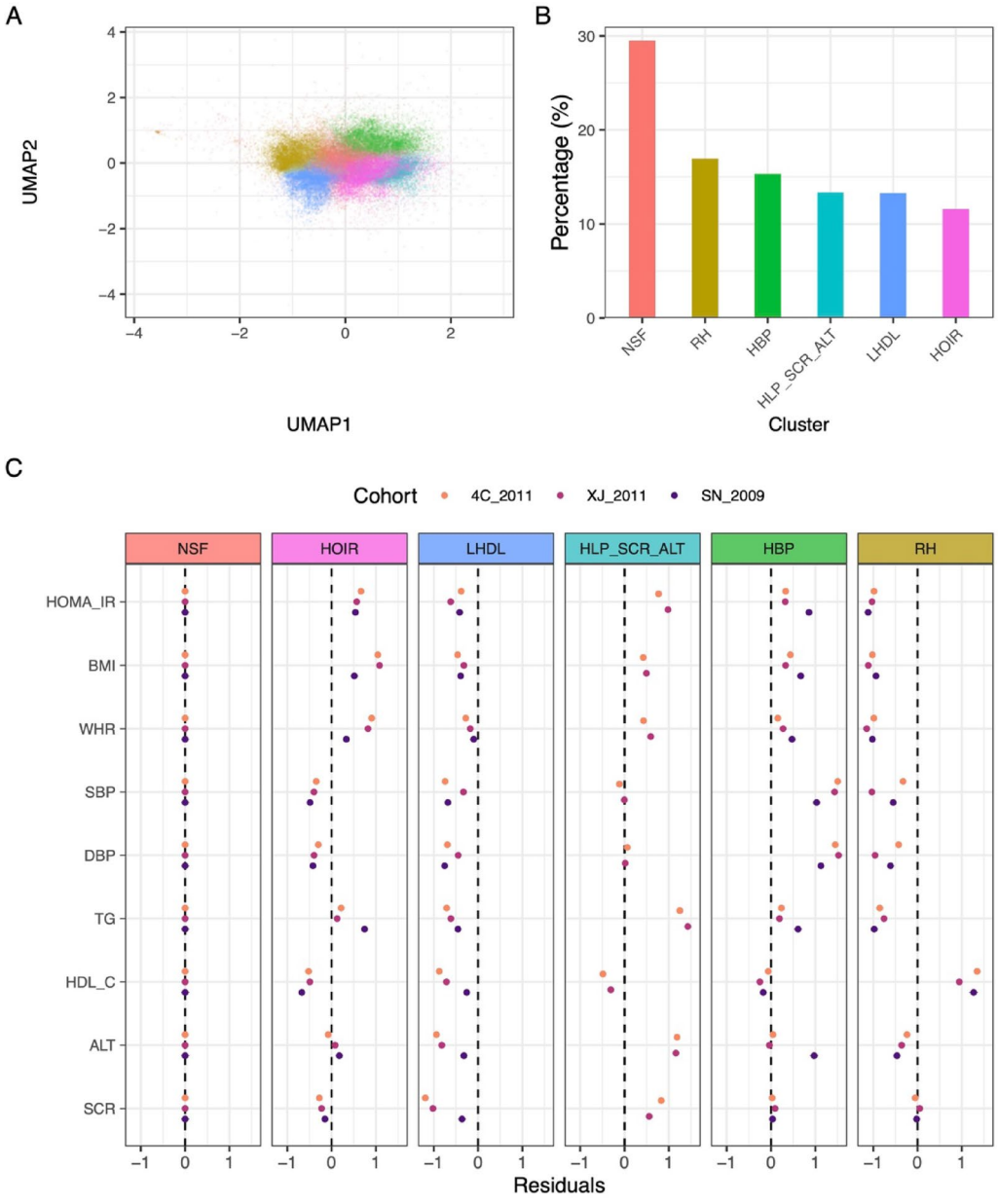


Fig. 2 Characteristics of the cluster profiles. Profiles were identified in the 4C_2011 and robustly replicated across the two independent cohorts. **A**, UMAP 2D projection. Colors denote profile allocations. **B**, Cluster weights. **C**, Forest plot of average phenotypic biomarker residuals characterizing each profile. Points and error bars represent the estimate mean residuals and 95% confidence intervals for each biomarker. The dashed line represents a residual of 0. NSF, non-significant features; HBP, high blood pressure; HLP_SCR_ALT, high lipids, serum creatinine, and ALT; HOIR, high obesity and insulin resistance; RH, relatively healthy; LHDL, low HDL-C

among the identified profiles. We observed that individuals with markable phenotypic characteristics exhibited multivariate profiles that differed substantially from the non-significant profile. These differences were reflected in the extent to which phenotype values deviated from age- and sex-specific expectations (Fig. 2C), in the absolute values of multiple phenotypic biomarkers (Supplementary Figs. 5 and 6), and in the correlation among these deviations (Supplementary Fig. 7).

Approximately 29.5% individuals exhibited no significant phenotypic abnormality, with biomarkers within the expected range for their age and sex, which we termed the 'non-significant feature' (NSF) profile (Fig. 2B). Distinct profiles were identified in the remaining individuals, representing ~ 70% of the total population. Around 15.3% of individuals showed a high blood pressure (HBP) profile, characterized by higher SBP and DBP values. A high lipids, serum creatinine, and ALT profile (HLP_SCR_ALT), defined by higher TG and lower HDL-C, as

well as elevated creatinine and ALT, were observed in 13.4% of individuals. Approximately 11.6% of individuals were assigned to the high obesity and insulin resistance (HOIR) profile, characterized by high BMI and HOMA-IR. Profiles of low HDL-C (LHDL), characterized by decreased HDL-C and other risk factors, accounted for 13.3% of individuals. About 16.9% of individuals had a relative healthy (RH) profile, characterized by generally lower phenotypic biomarker levels except for elevated HDL-C. Notably, individuals with the NSF profile formed a less cohesive subgroup, as indicated by a lower transitivity index (Supplementary Table 3), suggesting that profiles with significant phenotype features exhibited more structured and coordinated biomarker patterns.

Distributions of glycemic status across different profiles

As shown in Fig. 3, the crosstabs of the participants' profile and prediabetes definitions reveal several distinct patterns. In the LHDL profile, approximately 42% of participants with prediabetes were identified solely by HbA1c ≥ 5.7%, while in the RH profile, nearly 16% were diagnosed based only on FPG ≥ 5.6 mmol/L. The HLP_SCR_ALT, HOIR, and HBP profiles showed a higher proportion (over 50%) of individuals meeting more than one glycemic criteria. Around 21% of individuals in HLP_SCR_ALT profile met all three glycemic thresholds simultaneously: elevated FPG, 2 h-PG, and HbA1c.

Different profiles and the prevalence of baseline cardiometabolic diseases

We estimated the profile-specific prevalence of various cardiometabolic comorbidities associated with the phenotypic biomarkers selected using allocation probabilities as weights (Fig. 4A). The prevalence of cardiometabolic diseases differed markedly between the NSF profile and profiles with markable phenotypic characteristics. Participants with the RH profile exhibited the lowest burden of disease and were significantly less likely to have coronary heart disease (CHD), myocardial infarction (MI), stroke, hypertension or kidney disease compared to those in the NSF profile. Meanwhile, individuals in the LHDL profile also showed decreased rates of diseases, except for MI. In contrast, CHD and stroke were more prevalent in HBP, and CHD was further enriched in HOIR profile.

Aside from the expected enrichment of hypertension cases in the HBP profile compared with NSF (> 12 times greater prevalence than in the NSF profile). Meanwhile, individuals within the HLP_SCR_ALT profile were approximately 2.5 times more likely to have kidney diseases compared to those in the NSF profile. Among participants free from cardiometabolic conditions (CHD, MI, stroke, hypertension and kidney disease), the distribution of profile allocations was largely similar, except for a lower probability of assignment to HBP profile (Fig. 4B).

We investigated medication use across profiles (Fig. 4C). The RH profile had significantly lower use of antihypertensive or lipid-lowering medication compared to the NSF profile (OR for antihypertensive medication: 0.36, 95% CI 0.33, 0.39; OR for lipid-lowering medication: 0.47, 95% CI 0.33, 0.66). In contrast, anti-hypertensive therapy was nearly twofold more frequent in the HBP profile than in the NSF profile. After adjustment for these medications, the associations between profiles and disease prevalence remained largely unchanged (Fig. 4A).

Different profiles and incidence estimate of cardiometabolic diseases

To evaluate the effect of distinct phenotypic profiles on the incidence of cardiometabolic diseases, we analyzed data from 4C_2011, XJ_2011 and SN_2009 studies, and excluded individuals with prevalent diseases at baseline. Crude incidence estimates of T2D, CVD and kidney events were calculated for each profile using the weighted approach described previously (Supplementary

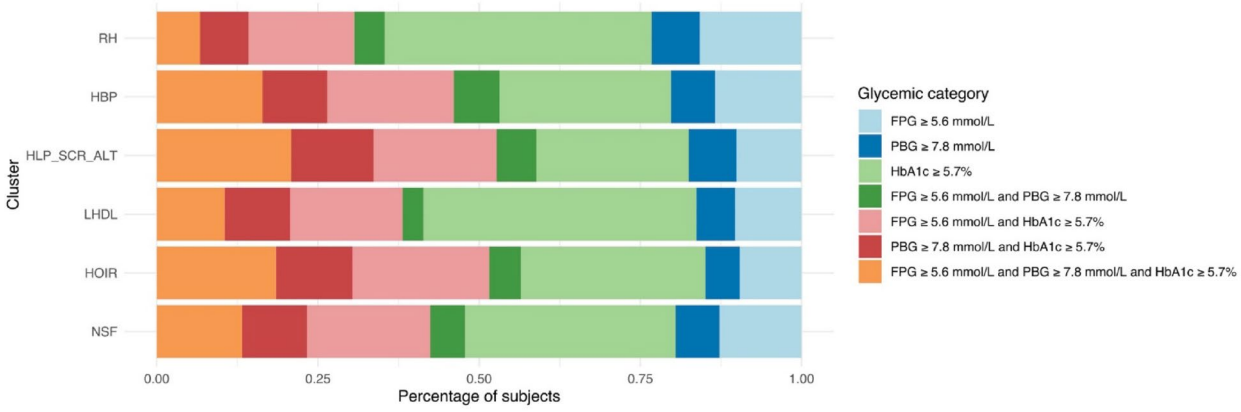


Fig. 3 Proportions of individuals meeting glycemic criteria within each phenotypic profile. NSF, non-significant features; HBP, high blood pressure; HLP_SCR_ALT, high lipids, serum creatinine, and ALT; HOIR, high obesity and insulin resistance; RH, relatively healthy; LHDL, low HDL-C. FPG, fasting plasma glucose; PBG, post-load blood glucose

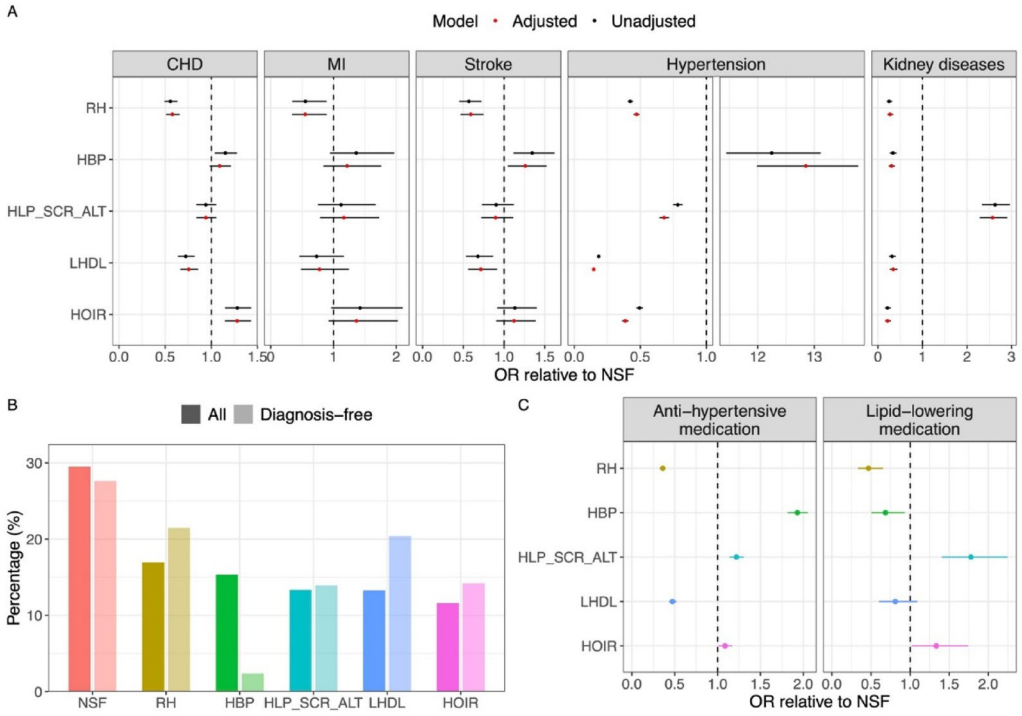


Fig. 4 Associations between phenotypic profiles, prevalent cardiometabolic comorbidities and medication use. **A**, Odds ratios (ORs) and 95% confidence intervals (CIs) for selected comorbidities across distinct phenotypic profiles relative to the non-significant feature (NSF) profile, presented before and after adjustment for lipid-lowering and antihypertensive medications. The dashed line represents the null association. **B**, Proportions of individuals with each profile among those without selected cardiometabolic conditions, compared with their distribution in the overall 4C population. **C**, ORs and 95% CIs for the use of selected medications across distinct profiles relative to the NSF profile. The dashed line represents the null association

Table 1 Model comparison without and with profile allocation probabilities

Outcome	Cohort	LRT P value	Added variance explained (%)	C-statistic baseline model	C-statistic baseline + profiles	C-statistic difference	Difference *P* value
Type 2 diabetes	4C_2011	$\mathbf{1.25\times10^{-3}}$	0.97	0.642	0.643	0.001	0.064
Type 2 diabetes	XJ_2011	0.689	5.44	0.637	0.637	< 0.001	0.963
Type 2 diabetes	SN_2009	0.059	9.57	0.702	0.744	0.042	0.214
Incident cardiovascular events	4C_2011	$\mathbf{2.84\times10^{-4}}$	2.27	0.748	0.750	0.002	0.052
Incident cardiovascular events	XJ_2011	0.927	1.58	0.854	0.855	0.001	0.778
Incident cardiovascular events	SN_2009	0.188	7.48	0.699	0.704	0.006	0.366
Incident kidney events	4C_2011	$\mathbf{1.31\times10^{-67}}$	16.82	0.718	0.736	0.019	$\mathbf{2.67\times10^{-11}}$
Incident kidney events	XJ_2011	0.177	9.10	0.724	0.742	0.018	$\mathbf{1.37\times10^{-2}}$
Incident kidney events	SN_2009	$\mathbf{3.14\times10^{-3}}$	21.48	0.849	0.902	0.053	0.088

Likelihood ratio tests, fraction of additional variance explained, and C-statistics comparing models with versus without profile allocation probabilities. LRT, likelihood ratio test. For CVD outcome, we applied Cox proportional hazards model, whereas for T2D and kidney diseases we used logistic regression models. The bold values indicated statistical significance

Fig. 8). The RH profile was associated with lower risks of diseases compared with NSF profile. In contrast, the HBP, HLP_SCR_ALT, and HOIR profiles were associated with an increased risk of T2D relative to NSF. Additionally, the HBP profile was associated with a higher risk of CVD compared with NSF (RR 1.41, 95% CI 1.23, 1.62). Individuals with the LHDL profile showed higher risk of incident CKD compared with NSF.

Added predictive value of phenotypic profiles for cardiometabolic diseases

To assess the added predictive value of phenotypic profiles, we incorporated profile allocation probabilities into base models. In the 4C_2011 study, the inclusion of profile information significantly improved model performance, as shown by likelihood ratio tests and increase in C-statistics (Table 1). The additional explained variance in T2D, CVD, and kidney events prediction explained by the profiles ranged from 0.97 to 16.82%. Although

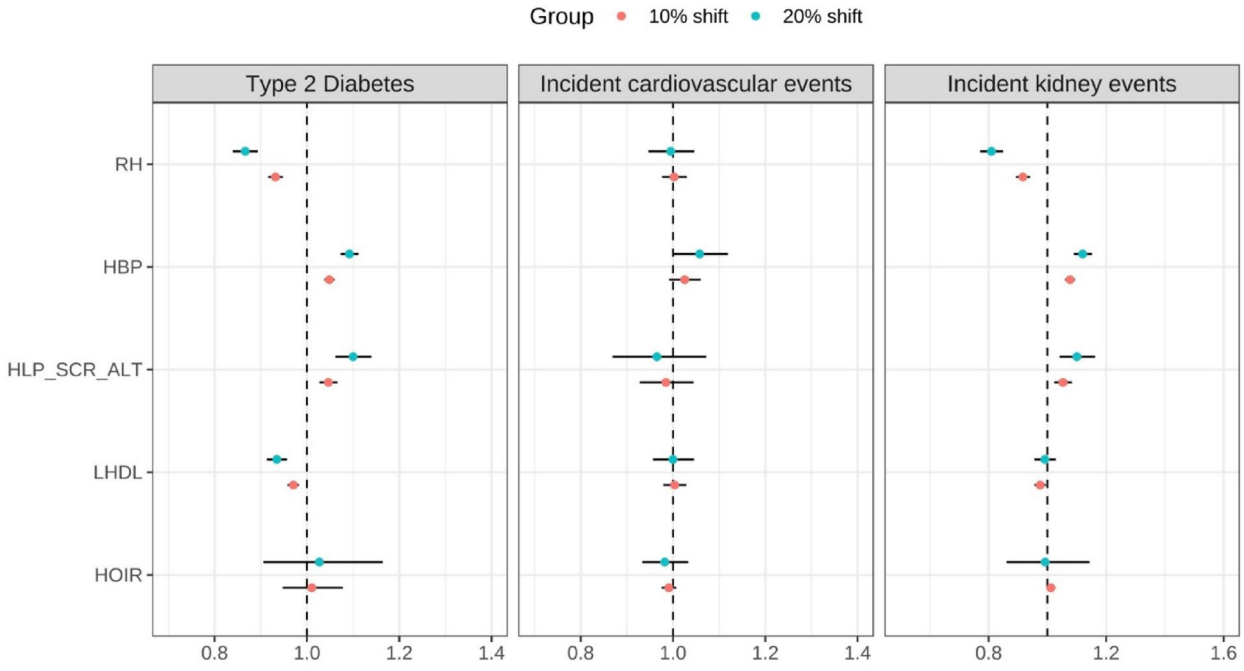


Fig. 5 Associations between phenotypic profile probabilities and future risk of cardiometabolic diseases. Odds ratios (ORs) or hazard ratios (HRs) with 95% confidence intervals (CIs) represent the estimated risk associated with a 10 or 20% increase in the probability of assignment to each distinct phenotypic profile, offset by a corresponding decrease in the probability of the non-significant features (NSF) profile. Estimates were derived from random-effects meta-analyses across cohorts

profile information also explained additional variance in XJ_2011 and SN_2009, the improvements were not statistically significant based on likelihood ratio tests.

Because profile allocation is inherently derived from phenotypic biomarker patterns, the model-derived estimates reflect complex interactions between profiles and individual biomarkers. These interactions modify the associations between phenotypic biomarkers and disease risk in a profile-dependent manner (Supplementary Fig. 9). To illustrate this, we estimated the change in disease risks in a hypothetical 55-year-old male with prediabetes, assuming a 10 or 20% increase in the allocation probability and corresponding decrease in NSF probability (Fig. 5, Supplementary Figs. 10 and 11). A 20% increase in the probability of belonging to the HBP or HLP_SCR_ALT profile was associated with higher risk of T2D compared with NSF profile, with ORs (95% CI) of 1.09 (1.07–1.11) and 1.10 (1.06–1.14), respectively (Fig. 5). Similarly, a 20% increase in the HBP probability was associated with a higher risk of CVD (HR 1.06, 95% CI 1.00, 1.12). Higher probabilities of HBP and HLP_SCR_ALT profiles were associated with a higher risk of incident kidney events compared with NSF, with ORs (95% CIs) of 1.12 (1.09–1.15) and 1.10 (1.04–1.16), respectively. Conversely, the RH profile was associated with lower risks of both incident T2D and kidney events compared with NSF (T2D OR 0.87, 95% CI 0.84, 0.89; kidney events OR 0.81, 95% CI 0.77, 0.85). Meanwhile, the risk of diseases in a 55-year-old female with prediabetes reaffirmed the findings observed in the male subgroup (Supplementary Fig. 12).

Discussion

In this study, using unsupervised clustering methods on routinely available clinical indicators, we identified five distinct phenotypic profiles and one NSF profile among Chinese individuals with prediabetes. These data-driven profiles revealed substantial phenotypic diversity and were associated with markedly different risks for T2D, CVD, and kidney events. Notably, the RH profile was consistently associated with lower disease risks, while HLP_SCR_ALT and HBP profiles indicated varying degree of higher cardiometabolic burden. The clustering structure proved robust and was successfully validated in two independent external cohorts. Furthermore, the incorporation of profile allocation probabilities significantly improved disease prediction models, suggesting the added clinical utility of phenotypic profiling beyond conventional biomarkers. We have demonstrated here how this approach enhance performance in ensuring phenotypic robustness, and improving clinical interpretability.

Understanding the heterogeneity of prediabetes is crucial for accurate risk stratification and targeted intervention. Within our phenotyping framework, the HBP subtype demonstrated a consistent association with the

development of T2D, CVD and CKD. Hypertension and diabetes share common pathophysiological risk factors and frequently co-occur. Elevated blood pressure markedly increases the risk of progression from prediabetes to overt T2D. A recent meta-analysis confirms that hypertension is among the more consistent comorbidity predictors of progression from prediabetes to T2D [30]. Meanwhile, prediabetes concomitant with hypertension was associated with an increased risk for CVD in the Jackson Heart Study (JHS) of 3313 black adults [31]. One of the most important complications of both diabetes and hypertension is loss of renal function and eventually leading to kidney events. In individuals with prediabetes, the coexistence of elevated blood pressure has been linked to a modestly increased risk of incident kidney disease [32]. The consistent association of the HBP phenotype with multiple disease risks suggests that metabolic derangements in this subgroup may drive disease progression via multiple converging pathways. Mechanistically, elevated blood pressure may exacerbate insulin resistance, endothelial dysfunction, increased oxidative stress, and low-grade inflammation, all of which may impair glucose homeostasis and accelerate β-cell dysfunction [33]. Clinically, this suggests that among individuals with prediabetes, assessment of blood pressure status should be considered not only for cardiovascular risk but also as a marker of diabetes and kidney disease risk.

Furthermore, the HLP_SCR_ALT profile, characterized by elevated lipids, creatinine, and ALT, showed a robust association with incident T2D and CKD, reinforcing prior evidence that even high-normal ALT levels are independently predictive of diabetes and kidney event risks [34–36]. This observation aligns with the growing body of evidence that liver dysfunction, particularly hepatic insulin resistance, plays a pivotal role in the early stages of metabolic disease. Elevated ALT level, a marker of liver injury and metabolic stress, may serve as a surrogate indicator of impaired hepatic glucose metabolism, which accelerates the development of T2D when coupled with insulin resistance [37]. Moreover, recent studies have suggested that elevated ALT levels may reflect systemic inflammation, oxidative stress, and a dysregulated metabolic environment, all of which contribute to insulin resistance, β-cell dysfunction, and renal function decline [36, 38–40]. In line with the observed role of organ-specific dysfunction in metabolic disease progression, our findings also demonstrated that high serum creatinine was significantly associated with increased risks of incident kidney and T2D events. These findings indicate that impaired renal function, as reflected by creatinine, may accelerate kidney and glucose tolerance damage in the prediabetic population. The combination of these factors highlights the need for early identification and targeted interventions to prevent the progression of complications in individuals at risk for diabetes.

Compared with previous studies, our approach offers several methodological and translational advantages. First, we accounted for age- and sex-related confounding by deriving residuals from linear models for each phenotype prior to clustering. This adjustment enhances the biological interpretability of phenotype differences by minimizing demographic influences known to affect both biomarker levels and cardiometabolic outcomes. Second, unlike conventional k-means clustering, which enforces hard assignment and may misclassify individuals near cluster boundaries, we applied a combination of unsupervised soft clustering learning techniques with Gaussian mixture models, that assign probabilistic membership. This strategy allows more sophisticated identification of borderline individuals and reduces classification errors [41]. Third, traditional algorithms assume that data variability is captured by linear relationships, which limits its ability to depict the essential nonlinear characteristics of clinical data and confuses cluster boundaries of distinguishing heterogeneity [42, 43]. By contrast, UMAP builds a topological and meaningful low-dimensional representation preserved distances and neighborhood relationships as much as possible in the original high-dimensional space [44]. It therefore improves more robust and stable prediction performance and enhances the overall biological interpretability of the model. Subsequent combination of UMAP and Gaussian mixture model [45] facilitates clear visualization and identifies more reproducible subtypes. In addition, to further enhance the model's robustness, we evaluated the consistency of profile assignment by examining whether individuals were allocated to specific profiles in the 4C_2011 study and other two validation cohorts with high certainty (probability > 75%), resulting in five robust feature clusters and one remaining cluster we termed NSF. This strategy demonstrated the advantages of reproducibility and generalizability for subclassification among prediabetes population in large cohorts.

It is important to emphasize that in this study, we aimed to identify clinically meaningful subtypes of prediabetes by leveraging the multidimensional relationships between biomarkers. While single biomarkers such as blood pressure or blood glucose levels are valuable in clinical practice, they offer limited capacity to characterize the complex pathophysiological interplay underlying metabolic disease heterogeneity. Our data-driven clustering approach offers a more comprehensive view of prediabetes, revealing stable subtypes that emerge from the interactions between multiple clinical markers. These subtypes reflect underlying disease patterns that may provide valuable insights into disease progression and risk prediction.

To enhance clinical applicability, we constructed the clustering model using routinely available clinical indicators to ensure the six profiles could be feasibly identified in real-world settings. Early identification of high-risk individuals during the prediabetic stage is crucial—particularly for the HBP and HLP_SCR_ALT subtypes, which exhibit the highest T2D and kidney event risks and may benefit most from early interventions such as blood pressure and fatty liver management [46, 47]. This model demonstrates strong scalability and adaptability, offering a promising tool for precision medicine. It can facilitate risk stratification, predict intervention responses in clinical settings, and support mechanistic investigations in metabolic disease research.

Some limitations of our study should be acknowledged. First, the findings of this study were based on a middle-aged and elderly Chinese population, and the generalizability of these findings to other ethnicities remains uncertain. Second, the clustering is based on limited set of biomarkers chosen for their pathological relevance and community applicability. Third, although we carefully controlled for confounding factors in the analyses, bias from unmeasured or residual confounders cannot be fully ruled out. Fourth, differences in follow-up protocols across cohorts may affect comparability; nevertheless, similar risk patterns for T2D, CKD, and CVD were observed across cohorts. In addition, single glycemia measurements may not fully capture long-term risk. Fifth, the relatively short follow-up duration may limit the statistical power for evaluating individual risks of complications. Finally, we wish to emphasize that while data-driven profiles complement conventional approaches by providing a more comprehensive and holistic view of prediabetes heterogeneity, their clinical utility requires further robust evaluation [48, 49], and as such, future studies utilizing individual-level data from clinical trials are needed to determine their utility in guiding targeted treatment strategies [50].

Conclusions

This study demonstrates the utility of an integrated unsupervised clustering framework in delineating the complex heterogeneity inherent in prediabetes. Based on a nationwide cohort of individuals with prediabetes in China, we identified five distinct phenotypic spectra with clearly defined clinical characteristics using unsupervised clustering approaches. These phenotypes differed significantly from the group without distinct features in terms of cardiometabolic risk factor profiles, as well as the prevalence and incidence of cardiometabolic diseases. The probability-based assignment of individuals to specific profiles, derived from routinely available clinical indicators, significantly improved the prediction of future cardiometabolic events. These findings provide novel insights into the pathophysiological diversity of prediabetes while establishing a scalable framework for risk stratification and targeted intervention, thereby advancing the paradigm of precision prevention in metabolic disease.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12933-025-03008-9.

Supplementary Material 1

Acknowledgements

We thank all the participants of the 4C study.

Author contributions

L.J., B.Y., and W.W. had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. J.X., W.S., W.J., D.Y., L.M., and L.Y. contributed equally to this work. Conceptualization: L.J., B.Y., W.W., G.N. Data curation: Z.R., X.Y., X.M., W.T., Q.H., Q.G., Q.Y, T.X., Y.Z., H.R., S.L., S.Q., Y.X., Y.L., W.Q., C.G., G.Z., W.G., S.F., G.X., L.Z., C.L., H.X., H.Y., L.Q., Z.Y., Z.T., L.C., W.Y., W.S., Y.T., D.H., L.D., L.S., C.L., Z.J., M.Y., N.G., L.J., B.Y., and W.W. Formal analysis: J.X., D.Y., L.H., W.J. Investigation: J.X., W.S., D.Y., L.M. Methodology: J.X., D.Y., L.H., B.Y., W.W., and L.J. Software: J.X., D.Y., L.H., H.F., W.H. Writing—original draft: J.X., D.Y., and L.J. Writing—review & editing: W.S., W.J., L.M., L.Y., H.C., B.Y., and L.J. Supervision: L.J., B.Y., W.W.

Funding

The research was funded by National Major Science and Technology Projects (2023ZD0508402, 2023ZD0508401, 2023ZD0508902, 2023ZD0508901, 2024ZD0536700), the National Natural Science Foundation of China (grant nos.82170819, 82370810, 82372347, 82200998, and 82570943), National Key Research and Development Program of China (grant nos. 2023YFC2506700, 2022YFC2505202, and 2025YFA1805800), Science and Technology Commission of Shanghai Municipality (grant nos. 22Y31900300, 23Y11908400, 23JS1400900, and 24QA2708000), Shanghai Municipal Health Commission (202340084), Innovative research team of high-level local universities in Shanghai. The sponsor had no role in the design and conduct of the study, in the collection, management, analysis, and interpretation of data; in the preparation, review, or approval of the manuscript; nor in the decision to submit the manuscript for publication.

Data availability

Data collected in our study, including individual data and a data dictionary defining each field in the set, is not available to others.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Endocrine and Metabolic Diseases, Shanghai Institute of Endocrine and Metabolic Diseases, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[2]Shanghai National Clinical Research Center for Endocrine and Metabolic Diseases, Key Laboratory for Endocrine and Metabolic Diseases of the National Health Commission of the PR China, Shanghai Digital Medicine Innovation Center, Shanghai National Center for Translational Medicine, Ruijin Hospital, Shanghai Jiao-Tong University School of Medicine, Shanghai, China
[3]The Second Affiliated Hospital of Harbin Medical University, Harbin, China

[4]The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China
[5]The First Affiliated Hospital of Guangxi Medical University, Nanning, China
[6]The First Hospital of Lanzhou University, Lanzhou, China
[7]Zhejiang Provincial Center for Disease Control and Prevention, Hangzhou, China
[8]The Affiliated Hospital of Guiyang Medical College, Guiyang, China
[9]Xinhua Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China
[10]Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
[11]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China
[12]The Affiliated Hospital of Southwest Medical University, Luzhou, China
[13]Fujian Provincial Hospital, Fujian Medical University, Fuzhou, China
[14]Department of Endocrinology, Central Hospital of Dalian University of Technology, Dalian, China
[15]The First Hospital of Jilin University, Changchun, China
[16]The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China
[17]Qilu Hospital of Shandong University, Jinan, China
[18]Jiangxi Provincial People's Hospital Affiliated to Nanchang University, Nanchang, China
[19]Central Hospital of Shanghai Jiading District, Shanghai, China
[20]Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
[21]Jiangsu Province Hospital on Integration of Chinese and Western Medicine, Nanjing, China
[22]The First Affiliated Hospital of Anhui Medical University, Hefei, China
[23]Karamay Municipal People's Hospital, Xinjiang, China
[24]The First Affiliated Hospital of Nanjing Medical University, Nanjing, China
[25]The First Affiliated Hospital of Chongqing Medical University, Chongqing, China
[26]Department of Gastrointestinal Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA
[27]Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, Maryland, USA
[28]Shandong Provincial Hospital affiliated to Shandong University, Jinan, China
[29]Chinese People's Liberation Army General Hospital, Beijing, China

Received: 24 June 2025 / Accepted: 4 November 2025
Published online: 02 December 2025

References

1. Rooney MR, Fang M, Ogurtsova K, Ozkan B, Echouffo-Tcheugui JB, Boyko EJ, et al. Global prevalence of prediabetes. Diabetes Care. 2023;46(7):1388–94.
2. Diagnosis and classification of diabetes mellitus. Diabetes care. 2010;33 Suppl 1(Suppl 1):S62–9.
3. Li Y, Teng D, Shi X, Qin G, Qin Y, Quan H, et al. Prevalence of diabetes recorded in mainland China using 2018 diagnostic criteria from the American diabetes association: national cross sectional study. BMJ. 2020;369:m997.
4. Zheng R, Xu Y, Li M, Gao Z, Wang G, Hou X, et al. Data-driven subgroups of prediabetes and the associations with outcomes in Chinese adults. Cell Rep Med. 2023;4(3):100958.
5. Tabák AG, Herder C, Rathmann W, Brunner EJ, Kivimäki M. Prediabetes: a high-risk state for diabetes development. Lancet. 2012;379(9833):2279–90.
6. Wagner R, Heni M, Tabák AG, Machann J, Schick F, Randrianarisoa E, et al. Pathophysiology-based subphenotyping of individuals at elevated risk for type 2 diabetes. Nat Med. 2021;27(1):49–57.
7. Li G, Zhang P, Wang J, Gregg EW, Yang W, Gong Q, et al. The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing diabetes prevention study: a 20-year follow-up study. Lancet. 2008;371(9626):1783–9.
8. Tuomilehto J, Lindström J, Eriksson JG, Valle TT, Hämäläinen H, Ilanne-Parikka P, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J Med. 2001;344(18):1343–50.
9. Prystupa K, Delgado GE, Moissl AP, Kleber ME, Birkenfeld AL, Heni M, et al. Clusters of prediabetes and type 2 diabetes stratify all-cause mortality in a cohort of participants undergoing invasive coronary diagnostics. Cardiovasc Diabetol. 2023;22(1):211.
10. Gao CX, Dwyer D, Zhu Y, Smith CL, Du L, Filia KM, et al. An overview of clustering methods with guidelines for application in mental health research. Psychiatry Res. 2023;327:115265.
11. Li Y, Chen GC, Moon JY, Arthur R, Sotres-Alvarez D, Daviglus ML, et al. Genetic subtypes of prediabetes, healthy lifestyle, and risk of type 2 diabetes. Diabetes. 2024;73(7):1178–87.
12. Stefan N, Fritsche A, Schick F, Häring HU. Phenotypes of prediabetes and stratification of cardiometabolic risk. Lancet Diabetes Endocrinol. 2016;4(9):789–98.
13. Pantanowitz L, Pearce T, Abukhiran I, Hanna M, Wheeler S, Soong TR, et al. Nongenerative artificial intelligence in medicine: advancements and applications in supervised and unsupervised machine learning. Mod Pathol. 2025;38(3):100680.
14. Wang T, Lu J, Shi L, Chen G, Xu M, Xu Y, et al. 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. 2020;8(2):115–24.
15. Lu J, He J, Li M, Tang X, Hu R, Shi L, et al. Predictive value of fasting glucose, postload glucose, and hemoglobin A(1c) on risk of diabetes and complications in Chinese adults. Diabetes Care. 2019;42(8):1539–48.
16. Wu S, Yi F, Zhou C, Zhang M, Zhu Y, Tuniyazi Y, et al. HbA1c and the diagnosis of diabetes and prediabetes in a middle-aged and elderly Han population from northwest China (HbA1c). J Diabetes. 2013;5(3):282–90.
17. Lin L, Peng K, Du R, Huang X, Lu J, Xu Y, et al. Metabolically healthy obesity and incident chronic kidney disease: The role of systemic inflammation in a prospective study. Obesity (Silver Spring). 2017;25(3):634–41.
18. 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(7):412–9.
19. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF 3rd, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009;150(9):604–12.
20. Huang H, Wang Y, Rudin C, Browne EP. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. Commun Biol. 2022;5(1):719.
21. Narayan A, Berger B, Cho H. Assessing single-cell transcriptomic variability through density-preserving data visualization. Nat Biotechnol. 2021;39(6):765–74.
22. Newman ME. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E Stat Nonlin Soft Matter Phys. 2006;74(3 Pt 2):036104.
23. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9(1):5233.
24. Jedidi K, Ramaswamy V, Desarbo WS. A maximum likelihood method for latent class regression involving a censored dependent variable. Psychometrika. 1993;58(3):375–94.
25. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67(1):1–48.
26. Khan SS, Coresh J, Pencina MJ, Ndumele CE, Rangaswami J, Chow SL, et al. Novel prediction equations for absolute risk assessment of total cardiovascular disease incorporating cardiovascular-kidney-metabolic health: a scientific statement from the American heart association. Circulation. 2023;148(24):1982–2004.
27. Khan SS, Matsushita K, Sang Y, Ballew SH, Grams ME, Surapaneni A, et al. Development and validation of the American heart association's PREVENT equations. Circulation. 2024;149(6):430–49.
28. Harrell F. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis 2015.
29. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid Based Ment Health. 2019;22(4):153–60.
30. Hu S, Ji W, Zhang Y, Zhu W, Sun H, Sun Y. Risk factors for progression to type 2 diabetes in prediabetes: a systematic review and meta-analysis. BMC Public Health. 2025;25(1):1220.
31. Hubbard D, Colantonio LD, Tanner RM, Carson AP, Sakhuja S, Jaeger BC, et al. Prediabetes and risk for cardiovascular disease by hypertension status in Black adults: the Jackson Heart Study. Diabetes Care. 2019;42(12):2322–9.
32. Derakhshan A, Bagherzadeh-Khiabani F, Arshi B, Ramezankhani A, Azizi F, Hadaegh F. Different combinations of glucose tolerance and blood pressure

status and incident diabetes, hypertension, and chronic kidney disease. J Am Heart Assoc. 2016. https://doi.org/10.1161/JAHA.116.003917.
33. Jia G, Sowers JR. Hypertension in diabetes: an update of basic mechanisms and clinical disease. Hypertension. 2021;78(5):1197–205.
34. Fraser A, Harris R, Sattar N, Ebrahim S, Davey Smith G, Lawlor DA. Alanine aminotransferase, gamma-glutamyltransferase, and incident diabetes: the British women's heart and health study and meta-analysis. Diabetes Care. 2009;32(4):741–50.
35. Morinaga A, Iwanaga K, Maki K, Ueno T, Kawano K, Funakoshi S, et al. Association between serum ALT levels and incidence of new-onset diabetes in general population of Japanese: a longitudinal observational study (ISSA-CKD). BMJ Open. 2023;13(8):e074007.
36. Ochiai H, Shirasawa T, Yoshimoto T, Nagahama S, Watanabe A, Sakamoto K, et al. Elevated alanine aminotransferase and low aspartate aminotransferase/alanine aminotransferase ratio are associated with chronic kidney disease among middle-aged women: a cross-sectional study. BMC Nephrol. 2020;21(1):471.
37. Bonnet F, Ducluzeau P-H, Gastaldelli A, Laville M, Anderwald CH, Konrad T, et al. Liver enzymes are associated with hepatic insulin resistance, insulin secretion, and glucagon concentration in healthy men and women. Diabetes. 2011;60(6):1660–7.
38. Hanley AJG, Wagenknecht LE, Festa A, D'Agostino RB Jr., Haffner SM. Alanine aminotransferase and directly measured insulin sensitivity in a multiethnic cohort: the Insulin Resistance Atherosclerosis Study. Diabetes Care. 2007;30(7):1819–27.
39. Gehrke N, Schattenberg JM. Metabolic inflammation-a role for hepatic inflammatory pathways as drivers of comorbidities in nonalcoholic fatty liver disease? Gastroenterology. 2020;158(7):1929-47.e6.
40. Ziolkowska S, Binienda A, Jabłkowski M, Szemraj J, Czarny P. The interplay between insulin resistance, inflammation, oxidative stress, base excision repair and metabolic syndrome in nonalcoholic fatty liver disease. Int J Mol Sci. 2021;22(20):11128.
41. Coral DE, Smit F, Farzaneh A, Gieswinkel A, Tajes JF, Sparso T, et al. Subclassification of obesity for precision prediction of cardiometabolic diseases. Nat Med. 2025;31(2):534–43.
42. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans R Soc Lond A Math Phys Eng Sci. 2016;374(2065):20150202.
43. Hou CKJ, Behdinan K. Dimensionality reduction in surrogate modeling: a review of combined methods. Data Sci Eng. 2022;7(4):402–27.
44. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018.
45. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R j. 2016;8(1):289–317.
46. Kahl S, Gancheva S, Straßburger K, Herder C, Machann J, Katsuyama H, et al. Empagliflozin effectively lowers liver fat content in well-controlled type 2 diabetes: a randomized, double-blind, phase 4, placebo-controlled trial. Diabetes Care. 2020;43(2):298–305.
47. Bi Y, Li M, Liu Y, Li T, Lu J, Duan P, et al. Intensive blood-pressure control in patients with type 2 diabetes. N Engl J Med. 2025;392(12):1155–67.
48. Dennis JM, Shields BM, Henley WE, Jones AG, Hattersley AT. Clusters provide a better holistic view of type 2 diabetes than simple clinical features–authors' reply. Lancet Diabetes Endocrinol. 2019;7(9):669.
49. Ohukainen P, Kuusisto S, Kettunen J, Perola M, Järvelin MR, Mäkinen VP, et al. Data-driven multivariate population subgrouping via lipoprotein phenotypes versus apolipoprotein B in the risk assessment of coronary heart disease. Atherosclerosis. 2020;294:10–5.
50. Dennis JM, Shields BM, Henley WE, Jones AG, Hattersley AT. 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. 2019;7(6):442–51.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.