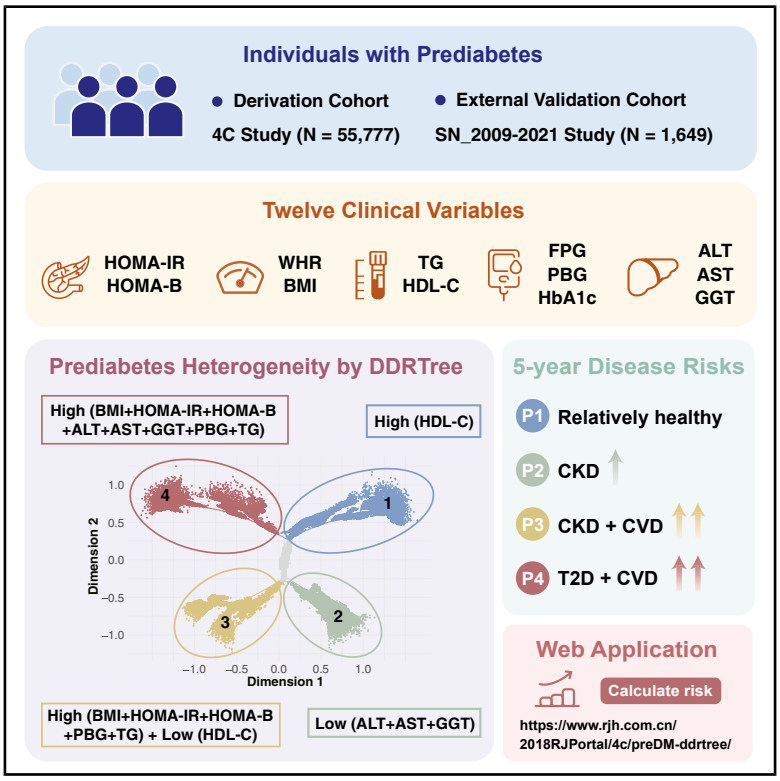
Cell Reports
Medicine

Article

Elucidating the heterogeneity of prediabetes through subphenotyping with a two-dimensional tree structure

Graphical abstract



Highlights

- Prediabetes heterogeneity is visualized via DDRTree based on 12 clinical variables
- Phenotypic heterogeneity influences subsequent risks of T2DM, CKD, and CVD
- CVD risk peaks leftward, with subtypes differentially distributed along dimension 2
- A user-friendly online tool is provided for personalized risk prediction

Authors

Hong Lin, Yilan Ding, Xiaojing Jia, ..., Yufang Bi, Jieli Lu, for the 4C Study Group

Correspondence

wqingw61@163.com (W.W.),
byf10784@rjh.com.cn (Y.B.),
ljl11319@rjh.com.cn (J.L.)

In brief

Lin et al. visualize prediabetes heterogeneity in 55,777 individuals from the 4C cohort using 12 clinical variables. They identify four distinct phenotypes with varying risks of developing type 2 diabetes mellitus, chronic kidney disease, and cardiovascular disease; validate their findings; and provide a user-friendly online tool for personalized risk prediction.

Lin et al., 2025, Cell Reports Medicine 6, 102212
July 15, 2025 © 2025 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.xcrm.2025.102212

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Elucidating the heterogeneity of prediabetes through subphenotyping with a two-dimensional tree structure

Hong Lin,[1,2,28] Yilan Ding,[1,2,28] Xiaojing Jia,[1,2,28] Xuejiang Gu,[3,28] Shuangyuan Wang,[1,2,28] Mian Li,[1,2] Yu Xu,[1,2] Min Xu,[1,2] Yiming Mu,[4] Lulu Chen,[5] Tianshu Zeng,[5] Lixin Shi,[6] Qing Su,[7] Yuhong Chen,[1,2] Xuefeng Yu,[8] Li Yan,[9] Guijun Qin,[10] Qin Wan,[11] Gang Chen,[12] Xulei Tang,[13] Zhengnan Gao,[14] Feixia Shen,[3] Ruying Hu,[15]

(Author list continued on next page)

[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 First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China
[4]Chinese People's Liberation Army General Hospital, Beijing, China
[5]Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
[6]Affiliated Hospital of Guiyang Medical College, Guiyang, China
[7]Xinhua Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China
[8]Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
[9]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China
[10]The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China
[11]The Affiliated Hospital of Southwest Medical University, Luzhou, China
[12]Fujian Provincial Hospital, Fujian Medical University, Fuzhou, China
[13]The First Hospital of Lanzhou University, Lanzhou, China
[14]Dalian Municipal Central Hospital, Dalian, China
[15]Zhejiang Provincial Center for Disease Control and Prevention, Hangzhou, China
[16]The First Affiliated Hospital of Guangxi Medical University, Nanning, China
[17]Qilu Hospital of Shandong University, Jinan, China
[18]Jiangxi Provincial People's Hospital Affiliated to Nanchang University, Nanchang, China
[19]The Second Affiliated Hospital of Harbin Medical University, Harbin, China
[20]The First Hospital of Jilin University, Changchun, China

(Affiliations continued on next page)

SUMMARY

Prediabetes, an intermediate stage of developing diabetes, exhibits considerable phenotypic heterogeneity. Here, we apply the Discriminative Dimensionality Reduction Tree (DDRTree) algorithm to explore prediabetes heterogeneity in 55,777 participants from the China Cardiometabolic Disease and Cancer Cohort (4C) study. Based on 12 clinically available variables, we identify four distinct phenotypes and observe differential risks of type 2 diabetes mellitus (T2DM), chronic kidney disease (CKD), and cardiovascular disease (CVD). Phenotype 4, characterized by hyperglycemia, insulin resistance, obesity, elevated triglycerides, and liver enzymes, has the highest T2DM risk, while phenotype 3, predominantly driven by obesity, insulin resistance, hyperglycemia, and dyslipidemia, has the highest CKD risk. Phenotypes 3 and 4 show higher CVD risk, with distinct distributions of CVD subtypes. These findings are validated in the external cohort SN_2009–2021, and a user-friendly online tool is provided for individual risk prediction. Overall, our study elucidates the intricate dynamics of prediabetes progression, aiding in personalized management for prediabetes care.

INTRODUCTION

Prediabetes represents an intermediate stage of glucose dysregulation that may precede type 2 diabetes mellitus (T2DM).[1] It affects approximately 720 million individuals worldwide in 2021 and will affect an estimated 1 billion people by 2045.[2] In China, the nationwide prospective China Cardiometabolic Disease and Cancer Cohort (4C) study reported a cumulative incidence

This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Zuojie Luo,[16] Yingfen Qin,[16] Li Chen,[17] Xinguo Hou,[17] Yanan Huo,[18] Qiang Li,[19] Guixia Wang,[20] Yinfei Zhang,[21] Chao Liu,[22] Youmin Wang,[23] Shengli Wu,[24] Tao Yang,[25] Huacong Deng,[26] Feiyue Huang,[2] Xingkun Xu,[2] Huapeng Wei,[2] Jie Zheng,[1,2] Tiange Wang,[1,2] Zhiyun Zhao,[1,2] Jiajun Zhao,[27] Guang Ning,[1,2] Weiqing Wang,[1,2,*] Yufang Bi,[1,2,*] Jieli Lu,[1,2,29,*] and for the 4C Study Group

[21]Central Hospital of Shanghai Jiading District, Shanghai, China
[22]Jiangsu Province Hospital on Integration of Chinese and Western Medicine, Nanjing, China
[23]The First Affiliated Hospital of Anhui Medical University, Hefei, China
[24]Karamay Municipal People's Hospital, Xinjiang, China
[25]The First Affiliated Hospital of Nanjing Medical University, Nanjing, China
[26]The First Affiliated Hospital of Chongqing Medical University, Chongqing, China
[27]Shandong Provincial Hospital affiliated to Shandong University, Jinan, China
[28]These authors contributed equally
[29]Lead contact
*Correspondence: wqingw61@163.com (W.W.), byf10784@rjh.com.cn (Y.B.), ljl11319@rjh.com.cn (J.L.)
https://doi.org/10.1016/j.xcrm.2025.102212

of 21.3% for progression from prediabetes, defined by combined impaired fasting tolerance, impaired glucose tolerance, and elevated hemoglobin A1c (HbA1c), to diabetes over up to 5 years of follow-up.[3] In another randomized clinical trials of diabetes prevention, the cumulative diabetes incidence reached 95.9% among participants with prediabetes in the control group during a 30-year follow-up period in the Chinese DaQing study.[4] Given the substantial high-risk population, effective screening and management strategies for individuals with prediabetes are critical to reducing the economic burden of T2DM and its complications.

Prediabetes is not a uniform state but rather a spectrum with considerable variation in metabolic profiles, risk factors, and potential outcomes.[5,6] Individuals with prediabetes exhibit significant differences in insulin sensitivity, insulin secretion, blood glucose levels, and lipid metabolism.[7,8] These factors influence metabolic status, pathophysiological mechanisms, and the risk of progressing from prediabetes to T2DM, as well as developing complications such as cardiovascular disease (CVD) and chronic kidney disease (CKD). Wagner et al. used *k*-means clustering in a European population to identify six prediabetes subtypes, revealing pathophysiological heterogeneity and a high-risk group for complications. We previously applied the *k*-means clustering method to stratify the Chinese prediabetic population into six subgroups, each exhibiting distinct risks for diabetes, CKD, and CVD progression. However, this approach does not provide individualized risk predictions.[6,9] Understanding this heterogeneity is crucial for developing personalized intervention and enabling more effective management to prevent T2DM and reduce the incidence of CVD and CKD events in patients with prediabetes.

The unprecedented increase in big data has brought considerable difficulties in data visualization and subsequent analysis. Conventional dimension reduction approaches like principal component analysis are challenged by explicitly recovering the intrinsic structure from the data and the discriminative feature representation.[10–12] Discriminative Dimensionality Reduction Tree (DDRTree), a computational method for visualizing and analyzing high-dimensional biological data, integrates principles from dimensionality reduction and decision tree algorithms to provide a more interpretable representation of complex data.[13] Besides, DDRTree has advantages in inferring the ordering as well as reconstructing the temporal progression and bifurcation structure of any data type.[10] Interestingly, recent studies have applied the DDRTree algorithm to visualize heterogeneity in type 2 diabetes[14,15] and type 1 diabetes,[16] revealing mixed traits converging at the tree trunk and distinct phenotypes at the distal branches. This method allows a clear two-dimensional visualization of complex, continuous phenotypic variations within individuals. Considering the central discrepancies in the pathophysiology of prediabetes, applying this method could provide significant insights into heterogeneity of prediabetes.

Hence, in this study, we aimed to elucidate (1) how the tree structure discriminates metabolic parameters among individuals with prediabetes and (2) how the heterogeneity of metabolic phenotypes predicts different complications among participants with prediabetes from large-scale prospective cohorts.

RESULTS

Baseline characteristics of the 4C study population were described in Table 1. The 12 variables used to determine prediabetes heterogeneity were (mean [standard deviation (SD)] or median [interquartile range (IQR)]) waist-to-hip ratio (WHR), 0.88 (SD 0.06); body mass index (BMI), 24.42 (SD 3.28) kg/m^2; fasting glucose (FPG), 5.57 (SD 0.52) mmol/L; post-load plasma glucose (PBG), 7.12 (SD 1.68) mmol/L; HbA1c, 5.81% (SD 0.33); homeostasis model assessment of insulin resistance (HOMA-IR), 1.67 (IQR 1.18–2.34); homeostasis model assessment of β cell function (HOMA-B), 67.42 (IQR 47.86–94.51); triglyceride (TG), 1.31 (IQR 0.94–1.87) mmol/L; high-density lipoprotein cholesterol (HDL-C), 1.35 (SD 0.35) mmol/L; alanine transaminase (ALT), 14.00 (IQR 10.80–20.00) U/L; aspartate aminotransferase (AST), 21.78 (SD 7.39) U/L; and glutamyl transferase (GGT), 20.00 (IQR 14.00–30.00) U/L.

Phenotypic characteristics and gradients across the tree structure

To get a visual representation of the phenotypic characteristics, we overlaid the data from 12 independent clinical variables on the tree, as shown in Figure 1A. We employed two metrics to evaluate the distribution patterns. Figure 1B and Table S1 showed the regression of each variable against the tree's two dimensions. WHR, BMI, HOMA-IR, and HOMA-B showed the

CellPress
OPEN ACCESS

Table 1. Baseline characteristics of the study population

Variables	4C study (*N* = 55,777)	SN_2009–2021 study (*N* = 1,649)
Males	18,762 (33.64%)	607 (36.8%)
Age (year)	56.26 ± 8.50	58.31 ± 9.07
WHR	0.88 ± 0.06	0.89 ± 0.06
BMI (kg/m^2)	24.42 ± 3.28	24.90 ± 3.39
FPG (mmol/L)	5.57 ± 0.52	4.99 ± 0.53
PBG (mmol/L)	7.12 ± 1.68	6.98 ± 1.64
HbA1c (%)	5.81 ± 0.33	5.98 ± 0.25
HOMA-IR	1.67 (1.18–2.34)	1.45 (0.93–2.14)
HOMA-B	67.42 (47.86–94.51)	88.64 (58.00–138.00)
TG (mmol/L)	1.31 (0.94–1.87)	1.32 (0.95–1.89)
HDL-C (mmol/L)	1.35 ± 0.35	1.38 ± 0.31
ALT (U/L)	14.00 (10.80–20.00)	17.00 (13.00–25.00)
AST (U/L)	21.78 ± 7.39	22.11 ± 6.79
GGT (U/L)	20.00 (14.00–30.00)	21.61 (15.93–31.87)

Data are expressed as mean ± SD, number (percentage), or median (IQR). ALT, alanine transaminase; AST, aspartate aminotransferase; WHR, waist-to-hip ratio; BMI, body mass index; FPG, fasting glucose; GGT, glutamyl transferase; HDL-C, high-density lipoprotein cholesterol; HOMA-B, homeostasis model assessment of β cell function; HOMA-IR, homeostasis model assessment of insulin resistance; PBG, post-load plasma glucose; TG, triglyceride. See also Figures S2 and S3.

consistent gradient pattern across both dimensions of the tree, with the highest values in the lower-left branches and the lowest values in the upper-right branch. Liver enzyme levels, including ALT, AST, and GGT, showed significant gradients in both dimensions across the tree, reaching the lowest in the lower-right branch and highest in the upper-left branches. Glycemic traits (FPG, PBG, and HbA1c) uniformly exhibited decreasing gradients along dimension 1, gathering higher concentrations in the left part. Along dimension 2, FPG and PBG showed slight increasing trends, whereas HbA1c displayed a reverse trend. TG and HDL-C showed completely opposite gradients in the two dimensions: HDL-C reached its highest values in the upper and right part, while TG was similar to glycemic traits with higher levels in the left part of the tree. The global Moran's I index was plotted to depict the intensity of spatial correlation on the tree (Figure 1C), and the local Moran's I (Figure S1) showed how strongly individuals within the tree are correlated with their neighbors. Among the 12 variables analyzed, ALT (Moran's I r = 0.72), AST (r = 0.72), and HOMA-IR (r = 0.64) had the strongest impact on the tree coordinates, followed by BMI and HOMA-B, with HDL-C, GGT, TG, and WHR having a comparatively minor impact. Glycemic traits, including FPG, PBG, and HbA1c, exhibited minimal variation across the tree structure. Given the inherent definition of prediabetes, the variation in baseline blood glucose levels in this population is relatively modest, thereby limiting its discriminatory potential.

In light of the clustering outcomes, we attributed four phenotypes to each branch within the tree for subsequent descriptions (Figure 1D). Individuals classified as phenotype 1 were the metabolically healthiest, with the lowest obesity indices, insulin resistance, and glycemic and lipid trait levels. Meanwhile, phenotype 2 had higher obesity indices, insulin resistance, and glycemic and lipid trait levels but lowest liver enzyme levels. Both phenotypes 3 and 4 included patients with unhealthy metabolic profiles, characterized by high TG, glycemic levels, obesity, and insulin resistance overall. The difference lies in that phenotype 3 included patients with highest HbA1c and lowest HDL-C levels, while phenotype 4 included patients with elevated liver enzymes.

Heterogeneity of prediabetes in relation to clinical outcomes

During up to 5 years of follow-up, we observed 3,615 cases of T2DM, 1,020 cases of CKD, and 933 cases of CVD (582 cases of stroke, 142 cases of myocardial infarction [MI], and 70 cases of heart failure). We then visualized the heterogeneity of prediabetes progression and clinical outcomes in the four phenotypes (Figures 2 and 3). The risks of T2DM, CKD, and CVD exhibited distinct spatial distributions across the DDRTree structure. T2DM risks peaked in the upper-left quadrant (phenotype 4) of the tree with hazard ratios (HRs) (95% confidence interval [95% CI]) at 0.62 (0.60–0.65) in dimension 1 and 1.14 (1.09–1.20) in dimension 2 (Figures 2A and 2C). Higher probabilities of T2DM were parallel to higher baseline glycemic traits, liver enzyme levels, insulin resistance, and obesity. Individuals classified as phenotype 3 had the highest risks for CKD with HR (95% CI) reaching 0.84 (0.78–0.91) in dimension 1 and 0.69 (0.62–0.76) in dimension 2 (Figures 2B and 2C). Incident CKD risks were driven by the combination of obesity, insulin resistance, glycemic traits, and dyslipidemia.

Overall cardiovascular outcomes increased in the left parts of the tree (phenotypes 3 and 4) with HR (95% CI) reaching 0.91 (0.84–0.98) in dimension 1 and 0.99 (0.90–1.09) in dimension 2 (Figure 3), corresponding to the regions with unhealthy metabolic profiles. A detailed assessment was then conducted in view of CVD subtypes, including MI, stroke, and heart failure, overlaying their probabilities on the tree (Figure 3A). Stroke, accounting for 57% of overall CVD, showed the similar risk distribution, with phenotypes 3 and 4 reaching higher risks (0.83 (0.75–0.92) in dimension 1 and 0.98 (0.86–1.12) in dimension 2). However, the significance of other CVD subtypes, namely MI (highest risks in phenotype 3) and heart failure (highest risks in phenotype 4), exhibited different distributions against dimension 2, with 0.74 (0.56–0.98) and 1.10 (0.74–1.62), respectively (Figure 3B). The strength of global spatial correlation, as indicated by Moran's I (Figures 2D and 3C), showed a strong and significant spatial correlation ($p < 0.001$) of each outcome across the tree.

External validation of prediabetes heterogeneity

For external validation, we used the data of the SN_2009–2021 cohort (Figure 4). We obtained T2DM, CVD, and CKD outcomes in SN_2009–2021 (Figure 5). As expected, the probability of T2DM overlaid on the tree revealed consistent highest risks in phenotype 4 observed in the 4C study, with HR (95% CI) reaching 0.34 (0.22–0.54) in dimension 1. The probabilities of CVD and CKD followed similar distributions to the 4C cohort, with the lowest risks all concentrating in phenotype 1. Those with higher risks had a tendency toward the lower and left part of the tree for the two outcomes reaching highest in phenotype 3 or 4.

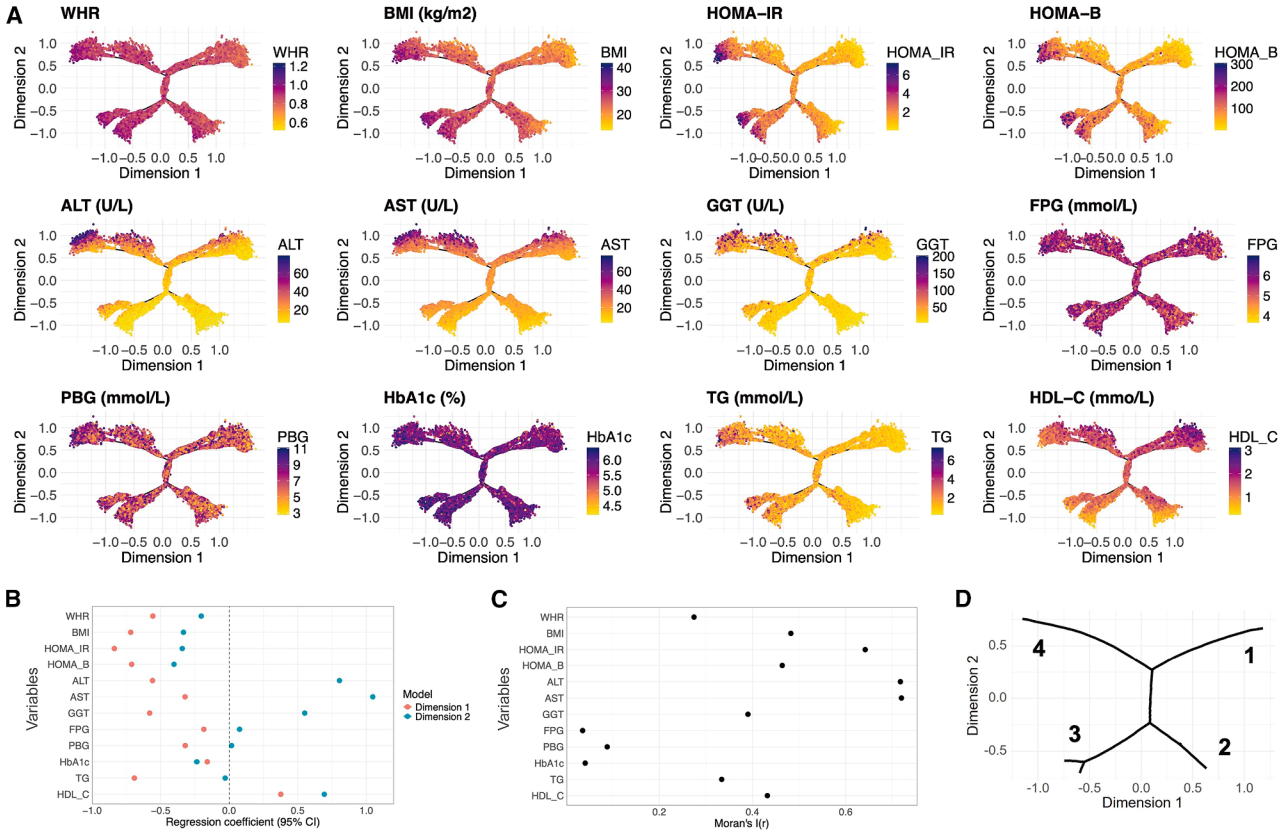


Figure 1. A visual representation of the phenotypic characteristics of 55,777 participants with prediabetes

(A) DDRTree was used to reduce the 12 clinical variables (BMI, WHR, HOMA-IR, HOMA-B, ALT, AST, GGT, FPG, PBG, HbA1c, TG, and HDL-C, residualized for age and gender into a non-linear tree structure [$N = 55,777$]). The values of clinical variables are overlayed on the tree structure to visualize the distribution of each variable over the reduced tree structure. Each point in the figure represents one individual.

(B) Linear regression ($N = 55,777$) estimates (with 95% CI) between the DDRTree dimensions and the 12 clinical variables, showing the association between variables and dimensions.

(C) Spatial autocorrelation ($N = 55,777$ for each spatial correlation analysis) of the 12 variables. The Moran's I statistic is shown on the x axis, with higher values representing variables that are more strongly autocorrelated; all values were at $p < 0.0001$.

(D) Attribution of numbers to four phenotypes based on four branches (from 1 to 4).

See also Table S1 and Figure S1.

However, it was slightly compromised by lacking a sufficient number of participants in the SN_2009–2021 cohort (Table 1).

Development of a risk visualization tool for prediabetes progression

To assist clinicians and patients in demonstrating and visualizing the risks of progression to diabetes and associated complications in individuals with prediabetes, we developed an application (https://www.rjh.com.cn/2018RJPortal/4c/preDM-ddrtree/). Individuals can be placed within the prediabetes continuum (tree structure) according to their risk profiles, with predictions made for T2DM, CKD, and CVD over a 5-year period. These risk predictions are based on age, gender, and twelve continuous phenotypic measures used to define the tree structure. Using DDRTree dimensions with age and gender, the estimated receiver operating characteristic area under the curve (ROC AUC) values of T2DM, CKD, and CVD were 0.632, 0.706, and 0.716, respectively. For CVD subtypes, the model performance ranged from ROC AUC 0.698 to 0.705. External validation demonstrated ROC AUC values between 0.700 and 0.814, with good calibration across the models.

DISCUSSION

In this study, we employed the DDRTree algorithm to elucidate the heterogeneity of prediabetes in a nationwide, population-based prospective Chinese cohort and validated it in another independent cohort. The resulting tree structure demonstrated four distinct metabolic branches based on 12 clinically relevant continuous phenotypic variables, which can effectively represent the differences in metabolic profiles and distinguish the risk of 5-year disease outcomes including incident T2DM, CKD, CVD, and even subtypes of CVD (MI, stroke, and heart failure). This study applies the DDRTree model to stratify prediabetic populations based on phenotypic profiles and

CellPress
OPEN ACCESS

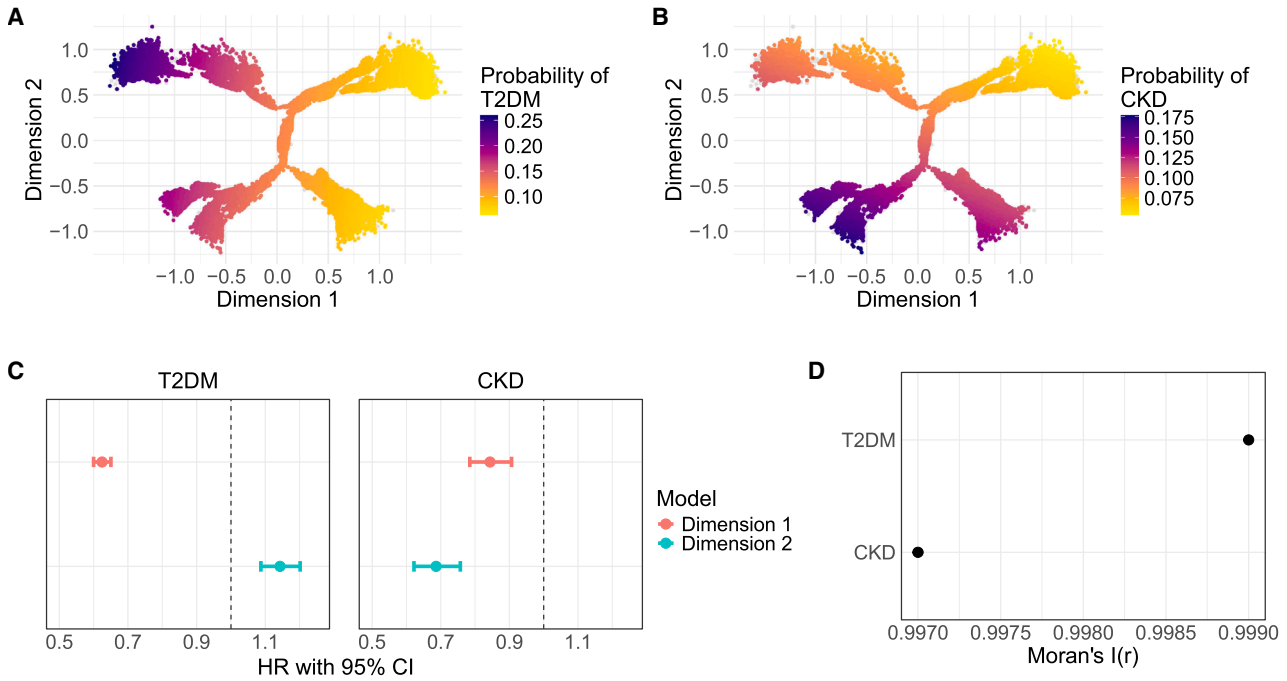


Figure 2. Visualizing the heterogeneity in prediabetes progression in participants with prediabetes

(A) Predicted probability of incident T2DM (N = 3,615) over a 5-year period (N = 47,692).

(B) Probability of incident CKD (N = 1,020) over a 5-year period (N = 37,409). For all outcomes (A and B), probabilities were generated from Cox hazard risk models constructed with DDRTree dimensions.

(C) HRs (95% CIs) of DDRTree dimensions for each outcome from Cox hazard risk models.

(D) Spatial autocorrelation of incident T2DM and CKD. Moran's I statistic is shown on the x axis, with higher values representing clinical variables that are more strongly autocorrelated; all values were at $p < 0.0001$. All predictions are from the models with DDRTree dimensions.

See also Figure S1.

emphasizes the importance of personalized management strategies in prediabetes care.

In the face of challenges including obesity, unhealthy dietary habits, and an aging demographic, approximately 35.2% of Chinese adults were estimated to be prediabetic.[17] Prediabetes is a multifaceted disease, where glucose measures alone inadequately capture the risk variations among individuals. Including variables representing body fat, glycemic measures, insulin resistance, blood lipids, and liver enzymes promotes the early identification of precision subgroup for prediabetes.

Intriguingly, beyond well-known T2DM risk factors such as glucose, body fat, and insulin resistance,[18] elevated levels of fasting triglycerides and liver enzymes were also observed in phenotype 4 conferring a significant increase in future diabetes risk. This aligns with previous studies suggesting that elevated or even high-normal levels of liver enzymes could be indicative of an elevated diabetes risk.[19,20] Phenotype 3, characterized primarily by highest obesity indices, insulin resistance, and glycemic levels but lowest HDL-C levels, presented the highest risk for future CKD. Abnormalities in HDL-C led to a subsequent reduction in its antioxidant and anti-inflammatory activities, which serve as major contributors to the pathogenesis of CKD-associated dysfunctions.[21] The mechanisms responsible for this renal damage were diverse and mainly involved hemodynamic alterations, oxidative stress, podocyte viability, and activation of the renin-angiotensin-aldosterone system.[22–24] These factors might also explain why phenotype 3 had a much greater probability of incident CKD than phenotype 4, despite both phenotypes displaying unhealthy metabolic profiles.

Different CVD subtypes, including stroke, MI, and heart failure, exhibited distinct risk distributions across the tree diagram. MI presented a higher risk in phenotype 3, while heart failure showed a higher risk in phenotype 4. These results mainly highlighted the effects of obesity indices and insulin resistance, especially HbA1c and HDL-C levels, on the onset of MI.[25–27] In addition to these parameters, the distribution of heart failure emphasized cardiohepatic interactions, which correspond to the gradient pattern of liver enzymes.[28] Overall, we have provided a clear visual representation that distinguishes the risk among CVD subtypes, thereby facilitating more precise prediction.

Compared with traditional methods of prediabetes classification, such as clustering approaches,[6,9] the DDRTree methodology offers a more detailed and comprehensive representation of the complexity and continuity in prediabetes progression. In previous research, Wagner et al. and our team utilized the k-means clustering method to classify prediabetes into distinct groups, thereby providing an accurate depiction of the risks for future diabetes and complications for each group.[6,9] The k-means algorithm, known for its simplicity and suitability for large datasets, provides clear

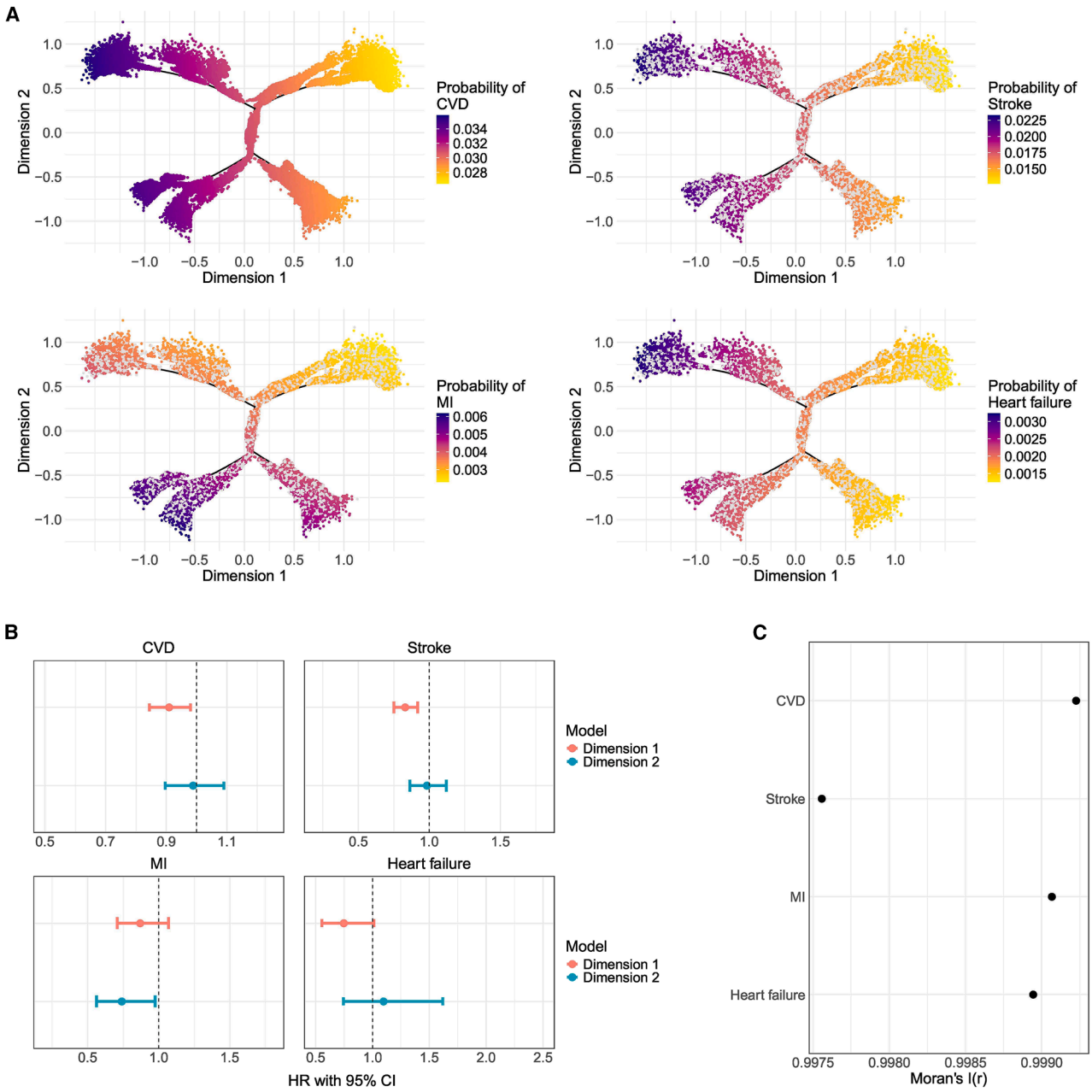


Figure 3. Visualizing the heterogeneity in CVD outcomes in participants with prediabetes

(A) Predicted probability of total CVD, stroke, MI, and heart failure over a 5-year period (cases/*N* = 933/47,818, 582/44,066, 142/46,493, and 70/46,270). For all outcomes, probabilities were generated from Cox hazard risk models constructed with DDRTree dimensions.

(B) HRs (95% CIs) of DDRTree dimensions for each outcome from Cox hazard risk models.

(C) Spatial autocorrelation of CVD outcomes. Moran's I statistic is shown on the *x* axis, with higher values representing variables that are more strongly autocorrelated; all values were at $p < 0.0001$, all predictions are from the models with DDRTree dimensions.

See also Figure S1.

clustering results and is particularly effective when data structures are relatively straightforward with well-defined boundaries. However, when dealing with continuous phenotypic and complex hierarchical representation of disease heterogeneity, the DDRTree algorithm was chosen over conventional methods like *k*-means because it preserves the continuous and nonlinear structure of phenotypic heterogeneity, allowing for a more interpretable trajectory-based representation of prediabetes subtypes. Unlike *k*-means, which forces hard clustering, DDRTree organizes individuals along a continuum, capturing transitional states between

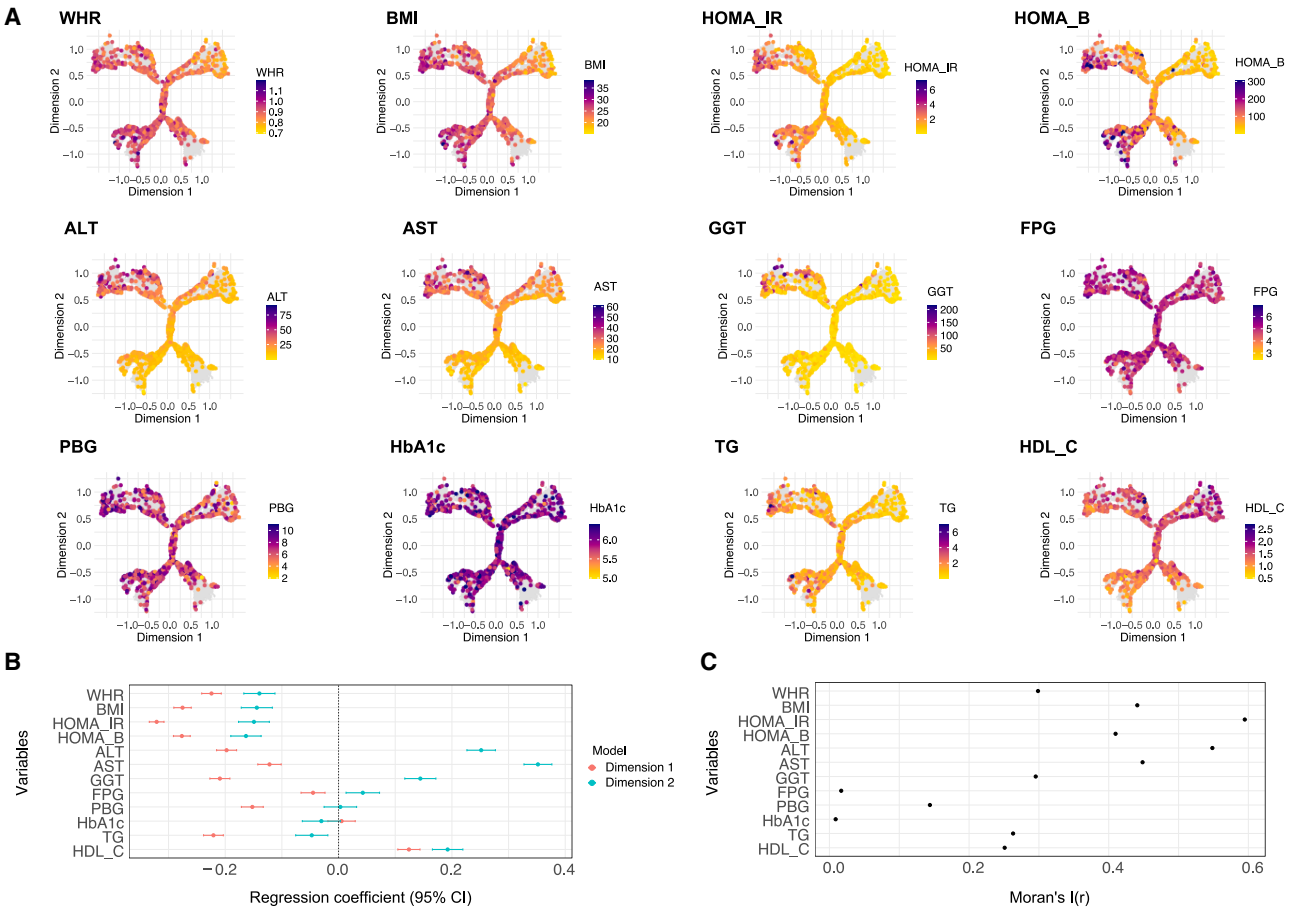


Figure 4. A visual representation of the phenotypic characteristics of (*N* = 1,649) participants with prediabetes at baseline in SN_2009–2021 study

(A) The distribution of 12 clinical variables (WHR, BMI, HOMA-IR, HOMA-B, ALT, AST, GGT, FPG, PBG, HbA1c, TG, and HDL-C) over the reduced tree structure. A mapping function was used to position individuals in SN_2009–2021 study (*N* = 1,649) onto the 4C (reference) tree. Each dot represents the position of one individual from the SN_2009–2021 study. The magenta color of the point indicates higher values of the clinical variable, and the yellow color indicates lower values.

(B) Linear regression (*N* = 1,649) estimates (with 95% CI) between the DDRTree dimensions and the 12 clinical variables, showing the association between variables and dimensions.

(C) Spatial autocorrelation (*N* = 1,649 for each spatial correlation analysis) of the 12 variables. The Moran's I statistic is shown on the *x* axis, with higher values representing variables that are more strongly autocorrelated.

subtypes. Compared to other clustering methods, such as random forests, DDRTree provides a low-dimensional, biologically interpretable structure that facilitates the visualization of metabolic progression. It provides a more understandable two-dimensional visualized model, which allows prediction of prediabetes progression at the individual level. Rather than defining individuals within each subgroup as homogeneity, combining continuous phenotypic measures to predict specific outcomes for individual patients probably has greater clinical utility.[15] Moreover, we have developed a user-friendly, web-based tool that provides risk scores, offering clinicians and patients an accessible means to guide personalized management strategies for prediabetic population.

Our study has several strengths. We applied the DDRTree algorithm to prediabetes in a nationwide population-based prospective cohort study. The large sample size enabled us to construct a robust two-dimensional tree structure, which excels in exploring disease progression and understanding heterogeneity, making it a valuable tool for capturing the intricate dynamics of disease evolution. This analysis provided a comprehensive continuous description of prediabetes heterogeneity, contributing to a more personalized prediabetes care. Particularly for individuals at high risk of diabetes and complications, especially specific CVD subtypes (stroke, MI, and heart failure), the DDRTree-based prevention strategy might bring benefits.

In conclusion, in the large-scale nationwide prospective cohort of prediabetes, we captured the intricate dynamics of prediabetes progression via the DDRTree algorithm, emphasizing the need for personalized approaches in prediabetes care and highlighting the complex interplay of phenotypic factors in disease risk stratification.

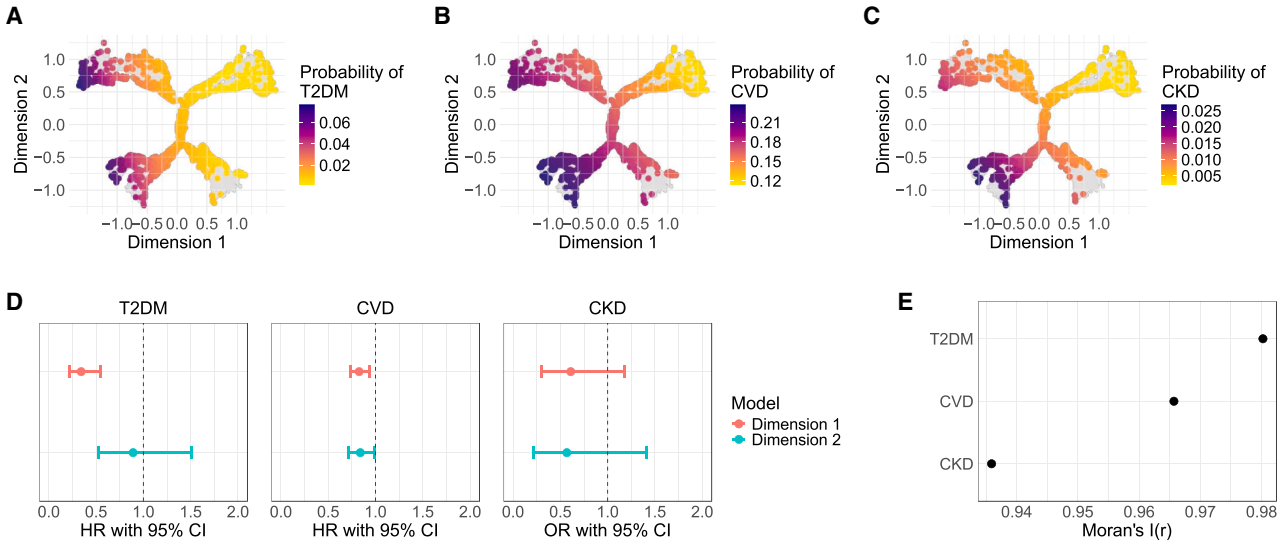


Figure 5. Visualizing the heterogeneity in disease progression in SN_2009–2021 study

A mapping function was used to position individuals in SN_2009–2021 study (N = 1,649) onto the 4C (reference) tree. Each dot represents the position of one individual from the SN_2009–2021 study.

(A) Predicted probability of incident T2DM (N = 40) over a 10-year period (N = 1,631).

(B) Predicted probability of incident CVD (N = 394) over a 10-year period (N = 1,646).

(C) Predicted probability of CKD (cases/N = 14/1,262). The probability of CKD was generated from logistic regression model constructed with DDRTree dimensions. For T2DM (A) and CVD outcomes (B), probabilities were generated from Cox proportional hazard models constructed with DDRTree dimensions.

(D) ORs/HRs (95% CI) of DDRTree dimensions for each outcome.

(E) Spatial autocorrelation of three outcomes. Moran's I statistic is shown on the x axis, with higher values representing variables that are more strongly autocorrelated; all value was at $p < 0.0001$. All predictions are from the models with DDRTree dimensions.

Limitations of the study

Our study has several potential limitations. First, we only recruited Chinese individuals aged 40 years and old, limiting the generalizability to other ethnic and age groups. Second, an average follow-up time of 3.8 years seems relatively short, especially for a population with prediabetes. Due to the relatively short period of follow-up duration, the statistical power for estimating individual complication risks might also be slightly compromised. Nevertheless, we observed 3,615 T2DM, 1,020 CKD, and 933 CVD events. Future work should integrate data from large-scale, multi-cohort studies to enhance long-term predictions and refine early complication risk models. Third, inflammatory markers may improve the discriminatory power of the tree structure; however, they were not measured in this cohort. To enhance the generalizability of our model, we focused on utilizing clinically available indicators. The lack of newer biomarkers (e.g., inflammatory cytokines, gut microbiota composition, and adipokines) limits our ability to achieve a more comprehensive understanding of metabolic heterogeneity and its relationship with various complications. Fourth, although we treated individual traits as independent features in the modeling process, we acknowledge that interactions among these traits may exist and contribute jointly to disease risk. While our clustering approach may have captured some of these combined effects indirectly, future work incorporating explicit modeling of trait interactions may further enhance the interpretability and predictive power of the model. Finally, beyond the phenotypic heterogeneity represented by the tree, further exploration is needed to understand other forms of heterogeneity in the genetic etiology of prediabetes, including genetic factors or any family history of diabetes. The lack of genetic data in this study limits such analyses, which remains an important direction for future research.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Jieli Lu (ljl11319@rjh.com.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Data generated and/or analyzed in this study, excluding identifying personal information, are available from the lead contact, Jieli Lu (ljl11319@rjh.com.cn) with reasonable request to protect research participant privacy. This paper does not report the original code. Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by Noncommunicable Chronic Diseases-National Science and Technology Major Project (grant nos. 2023ZD0508402, 2023ZD0508401, 2023ZD0508901, and 2023ZD0508902), the Ministry of Science and Technology (China) of China (grant nos. 2022YFC2505202 and 2023YFC2506704), National Natural Science Foundation of China (grant

nos. 82370810, 91857205, 82088102, 82170819, 82372347, and 82200998), Science and Technology Committee of Shanghai (grant nos. 23Y11908400, 23JS1400900, and 24QA2708000), Clinical Research Project of Shanghai Municipal Health Commission (grant nos. 20224Y0087 and 20214Y0002), and Innovative Research Team of High-Level Local Universities in Shanghai.

AUTHOR CONTRIBUTIONS

H.L., Y.D., X.J., X.G., and S.W. performed the statistical analysis. H.L., Y.D., X. J., X.G., and S.W. drafted the manuscript. J.L., G.N., Y.B., and W.W. conceived and designed the study. G.N. supervised the study. M.L., Y.X., M. X., Y.M., L.C., T.Z., L.S., Q.S., Y.C., X.Y., L.Y., G.Q., Q.W., G.C., X.T., Z.G., F.S., R.H., Z.Y., Z.L., Y.Q., L.C., X.H., Y.H., Q.L., G.W., Y.Z., C.L., Y.W., S. W., T.Y., H.D., F.H., X.X., H.W., J. Zheng, T.W., Z.Z., and J. Zhao contributed to acquisition, analysis, or interpretation of data. All authors revised the report and approved the final version before submission. J.L., G.N., Y.B., and W.W. are the guarantors of this work and, as such, had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

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

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cohort and variable definitions
  - 4C study
  - SN_2009–2021 cohort
- METHOD DETAILS
  - Definitions of prediabetes
  - Outcome assessment
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Data dimensionality reduction with DDRTree
  - Risk estimates of prediabetes-related outcome
  - Quantitative the distribution pattern across the tree
  - Constructing mapping function for external validation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2025.102212.

Received: December 3, 2024
Revised: April 9, 2025
Accepted: June 3, 2025
Published: July 2, 2025

REFERENCES

1. Echouffo-Tcheugui, J.B., Perreault, L., Ji, L., and Dagogo-Jack, S. (2023). Diagnosis and Management of Prediabetes: A Review. JAMA *329*, 1206–1216. https://doi.org/10.1001/jama.2023.4063.
2. Magliano, D.J., and Boyko, E.J.; IDF Diabetes Atlas Committee (2021). IDF Diabetes Atlas, 10th ed. (International Diabetes Federation).
3. 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 A(1c) on Risk of Diabetes and Complications in Chinese Adults. Diabetes Care *42*, 1539–1548. https://doi.org/10.2337/dc18-1390.
4. Li, G., Zhang, P., Wang, J., Gregg, E.W., Yang, W., Gong, Q., Li, H., Li, H., Jiang, Y., An, Y., et al. (2008). 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 *371*, 1783–1789. https://doi.org/10.1016/s0140-6736(08)60766-7.
5. 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. https://doi.org/10.1016/s0140-6736(12)60283-9.
6. 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. https://doi.org/10.1038/s41591-020-1116-9.
7. Tabák, A.G., Jokela, M., Akbaraly, T.N., Brunner, E.J., Kivimäki, M., and Witte, D.R. (2009). Trajectories of glycaemia, insulin sensitivity, and insulin secretion before diagnosis of type 2 diabetes: an analysis from the Whitehall II study. Lancet *373*, 2215–2221. https://doi.org/10.1016/s0140-6736(09)60619-x.
8. Magkos, F., Lee, M.H., Lim, M., Cook, A.R., Chhay, V., Loh, T.P., Chia, K. S., Baig, S., Ang, I.Y.H., Tay, J.Y.Y., et al. (2022). Dynamic assessment of insulin secretion and insulin resistance in Asians with prediabetes. Metab. Clin. Exp. *128*, 154957. https://doi.org/10.1016/j.metabol.2021.154957.
9. Zheng, R., Xu, Y., Li, M., Gao, Z., Wang, G., Hou, X., Chen, L., Huo, Y., Qin, G., Yan, L., et al. (2023). Data-driven subgroups of prediabetes and the associations with outcomes in Chinese adults. Cell Rep. Med. *4*, 100958. https://doi.org/10.1016/j.xcrm.2023.100958.
10. Mao, Q., Wang, L., Goodison, S., and Sun, Y. (2015). Dimensionality Reduction Via Graph Structure Learning. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Association for Computing Machinery).
11. Pantanowitz, L., Pearce, T., Abukhiran, I., Hanna, M., Wheeler, S., Soong, T.R., Tafti, A.P., Pantanowitz, J., Lu, M.Y., Mahmood, F., et al. (2025). Non-generative Artificial Intelligence in Medicine: Advancements and Applications in Supervised and Unsupervised Machine Learning. Mod. Pathol. *38*, 100680. https://doi.org/10.1016/j.modpat.2024.100680.
12. Huertas-García, Á., Martín, A., Huertas-Tato, J., and Camacho, D. (2023). Exploring Dimensionality Reduction Techniques in Multilingual Transformers. Cognit. Comput. *15*, 590–612. https://doi.org/10.1007/s12559-022-10066-8.
13. Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods *14*, 979–982. https://doi.org/10.1038/nmeth.4402.
14. Nair, A.T.N., Wesolowska-Andersen, A., Brorsson, C., Rajendrakumar, A. L., Hapca, S., Gan, S., Dawed, A.Y., Donnelly, L.A., McCrimmon, R., Doney, A.S.F., et al. (2022). Heterogeneity in phenotype, disease progression and drug response in type 2 diabetes. Nat. Med. *28*, 982–988. https://doi.org/10.1038/s41591-022-01790-7.
15. 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. https://doi.org/10.1016/s2213-8587(19)30087-7.
16. Fagherazzi, G., Aguayo, G.A., Zhang, L., Hanaire, H., Picard, S., Sablone, L., Vergès, B., Hamamouche, N., Detournay, B., Joubert, M., et al. (2024). Heterogeneity of glycaemic phenotypes in type 1 diabetes. Diabetologia (Berl.) *67*, 1567–1581. https://doi.org/10.1007/s00125-024-06179-4.
17. 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. https://doi.org/10.1136/bmj.m997.
18. Zheng, Y., Ley, S.H., and Hu, F.B. (2018). Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat. Rev. Endocrinol. *14*, 88–98. https://doi.org/10.1038/nrendo.2017.151.

19. Lorenzo, C., Hanley, A.J., Rewers, M.J., and Haffner, S.M. (2013). The association of alanine aminotransferase within the normal and mildly elevated range with lipoproteins and apolipoproteins: the Insulin Resistance Atherosclerosis Study. Diabetologia (Berl.) *56*, 746–757. https://doi.org/10.1007/s00125-012-2826-4.

20. Schneider, A.L.C., Lazo, M., Ndumele, C.E., Pankow, J.S., Coresh, J., Clark, J.M., and Selvin, E. (2013). Liver enzymes, race, gender and diabetes risk: the Atherosclerosis Risk in Communities (ARIC) Study. Diabet. Med. *30*, 926–933. https://doi.org/10.1111/dme.12187.

21. Vaziri, N.D. (2016). HDL abnormalities in nephrotic syndrome and chronic kidney disease. Nat. Rev. Nephrol. *12*, 37–47. https://doi.org/10.1038/nrneph.2015.180.

22. Jiang, Z., Wang, Y., Zhao, X., Cui, H., Han, M., Ren, X., Gang, X., and Wang, G. (2023). Obesity and chronic kidney disease. Am. J. Physiol. Endocrinol. Metab. *324*, E24–e41. https://doi.org/10.1152/ajpendo.00179.2022.

23. Lastra, G., Dhuper, S., Johnson, M.S., and Sowers, J.R. (2010). Salt, aldosterone, and insulin resistance: impact on the cardiovascular system. Nat. Rev. Cardiol. *7*, 577–584. https://doi.org/10.1038/nrcardio.2010.123.

24. 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. https://doi.org/10.1038/nrneph.2016.145.

25. Prospective Studies Collaboration; Lewington, S., Whitlock, G., Clarke, R., Sherliker, P., Emberson, J., Halsey, J., Qizilbash, N., Peto, R., and Collins, R. (2007). Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths. Lancet *370*, 1829–1839. https://doi.org/10.1016/s0140-6736(07)61778-4.

26. Di Angelantonio, E., Sarwar, N., Perry, P., Kaptoge, S., Ray, K.K., Thompson, A., Wood, A.M., Lewington, S., Sattar, N., et al.; Emerging Risk Factors Collaboration (2009). Major lipids, apolipoproteins, and risk of vascular disease. JAMA *302*, 1993–2000. https://doi.org/10.1001/jama.2009.1619.

27. Rawshani, A., Rawshani, A., Franzén, S., Sattar, N., Eliasson, B., Svensson, A.M., Zethelius, B., Miftaraj, M., McGuire, D.K., Rosengren, A., and Gudbjörnsdottir, S. (2018). Risk Factors, Mortality, and Cardiovascular Outcomes in Patients with Type 2 Diabetes. N. Engl. J. Med. *379*, 633–644. https://doi.org/10.1056/NEJMoa1800256.

28. Samsky, M.D., Patel, C.B., DeWald, T.A., Smith, A.D., Felker, G.M., Rogers, J.G., and Hernandez, A.F. (2013). Cardiohepatic interactions in heart failure: an overview and clinical implications. J. Am. Coll. Cardiol. *61*, 2397–2405. https://doi.org/10.1016/j.jacc.2013.03.042.

29. Wang, T., Lu, J., Shi, L., Chen, G., Xu, M., Xu, Y., Su, Q., Mu, Y., Chen, L., Hu, R., et al. (2020). 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. https://doi.org/10.1016/s2213-8587(19)30425-5.

30. Pickering, T.G., Hall, J.E., Appel, L.J., Falkner, B.E., Graves, J., Hill, M.N., Jones, D.W., Kurtz, T., Sheps, S.G., and Roccella, E.J. (2005). Recommendations for blood pressure measurement in humans and experimental animals: part 1: blood pressure measurement in humans: a statement for professionals from the Subcommittee of Professional and Public Education of the American Heart Association Council on High Blood Pressure Research. Circulation (New York, N. Y.) *111*, 697–716. https://doi.org/10.1161/01.CIR.0000154900.76284.F6.

31. Matthews, D.R., Hosker, J.P., Rudenski, A.S., Naylor, B.A., Treacher, D.F., and Turner, R.C. (1985). Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia (Berl.) *28*, 412–419. https://doi.org/10.1007/bf00280883.

32. Lin, L., Zhang, J., Jiang, L., Du, R., Hu, C., Lu, J., Wang, T., Li, M., Zhao, Z., Xu, Y., et al. (2020). Transition of metabolic phenotypes and risk of subclinical atherosclerosis according to BMI: a prospective study. Diabetologia (Berl.) *63*, 1312–1323. https://doi.org/10.1007/s00125-020-05116-5.

33. Wang, T., Lu, J., Xu, Y., Li, M., Sun, J., Zhang, J., Xu, B., Xu, M., Chen, Y., Bi, Y., et al. (2013). Circulating prolactin associates with diabetes and impaired glucose regulation: a population-based study. Diabetes Care *36*, 1974–1980. https://doi.org/10.2337/dc12-1893.

34. American Diabetes Association (2010). Diagnosis and classification of diabetes mellitus. Diabetes Care *33*, S62–S69. https://doi.org/10.2337/dc10-S062.

35. Andrassy, K.M. (2013). Comments on 'KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Kidney Int. *84*, 622–623. https://doi.org/10.1038/ki.2013.243.

36. 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. https://doi.org/10.1186/s12882-022-02979-1.

37. Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.A., and Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. Nat. Methods *14*, 309–315. https://doi.org/10.1038/nmeth.4150.

38. Ord, J.K., and Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geogr. Anal. *27*, 286–306. https://doi.org/10.1111/j.1538-4632.1995.tb00912.x.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Serum samples of cases and controls	4C study	Serum samples of cases and controls
Software and algorithms		
R version 4.3.3	R-Project	https://www.r-project.org/
DDRTree algorithm:R package monocle version 2.34.0	Bioconductor	https://www.bioconductor.org/packages/release/bioc/html/monocle.html
Spatial autocorrelation analysis:R package spdep version 1.3.5	R CRAN	https://mirrors.tuna.tsinghua.edu.cn/CRAN/

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cohort and variable definitions

We used two community-dwelling prospective cohort datasets for this analysis: (1) 4C study —a nationwide, multicenter, prospective, population-based study among Chinese adults; (2) SN_2009–2021 —a prospective cohort established in Songnan District, Shanghai, China.

4C study

The 4C study was designed to examine the relationship between glycemic parameters and clinical outcomes, including diabetes, CVD, cancer, and mortality.[3,9,29] The study design of the 4C Study has been described in detail previously.[3] From 2011 to 2012, a total of 193,846 adults aged over 40 years old were recruited from 20 communities located at different geographical regions across mainland China. A comprehensive set of questionnaires, clinical measurements, and laboratory examinations were carried out at the baseline visit. Body weight, height, waist circumference, and hip circumference were measured according to a standard protocol by trained study nurses.[30] Three blood pressure measurements obtained with participants in a seated position after five minutes of quiet rest using an automated electronic device (OMRON Model HEM-752 FUZZY, Dalian, China), were averaged for analysis.

All participants underwent an OGTT, and plasma glucose was obtained at zero and two hours during the test. Plasma glucose concentrations were analyzed locally using a glucose oxidase or hexokinase method within two hours after blood sample collection under a stringent quality control program. Hemoglobin Capillary Collection System (Bio-Rad Laboratories, CA, USA) was used to collect finger capillary whole blood and shipped at 2°C-8°C to a certificated central laboratory at Ruijin Hospital. This clinical laboratory is certificated by the US National Glycohemoglobin Standardization Program and passed the College of American Pathologists (CAP)'s Laboratory Accreditation Program. HbA1c was determined using high-performance liquid chromatography (VARIANT II System, Bio-Rad Laboratories, CA, USA). Serum insulin, ALT, AST, GGT, HDL-C, and TG were measured at the central laboratory using an auto-analyzer (ARCHITECT ci16200 analyzer, Abbott Laboratories, Illinois, USA).

The eGFR was calculated using the chronic kidney disease epidemiology collaboration (CKD-EPI) equation. Insulin resistance was estimated using the HOMA-IR index with the following formula: fasting insulin (μIU/mL) × fasting glucose (mg/dL)/405. β-cell function was estimated using the HOMA-B index with the following formula: (360×fasting insulin [μIU/mL])/(fasting glucose [mg/dL]–63).[31]

During 2014–2016, all participants were invited to attend an in-person follow-up visit. Trained staff queried medical history using the same standard questionnaire as was used at baseline visit. Anthropometric values and blood pressure were measured, OGTT was performed, and blood samples were obtained using the same protocols as used during the baseline examination.

In the 4C study, 64,650 participants were diagnosed with prediabetes based on OGTT and HbA1c measurements. Among them, 7,416 participants with missing data in biochemical detection and physical measurement, and 1,457 participants with outlier values for the clustering variables were excluded. Finally, 55,777 individuals diagnosed with prediabetes were included in the analysis (Figure S2). To ensure a robust and unbiased selection of clustering variables, we employed a forward stepwise logistic regression model to identify the most important biomarkers independently associated with T2DM based on the collected anthropometric and biochemical measurements. Variables with statistical significance ($p < 0.05$)[9] were retained for clustering, including BMI and WHR as measures of adiposity; HOMA-IR and HOMA-B as indicators of insulin resistance and β-cell function; and a range of glycemic and metabolic markers (FPG, 2 h PG, HbA1c, TG, HDL-C, ALT, AST, and GGT).

The 4C study was 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.

SN_2009–2021 cohort

SN_2009-2021 study, a community-dwelling prospective cohort including 4,012 individuals aged ≥40 years old was established from Songnan region in Shanghai, China, from June to August 2009, and followed until November 2021.[32,33] All participants received the OGTT, and blood samples were collected at 0 and 2 h. FPG and PBG were measured using the glucose oxidase method on an autoanalyzer (ADVIA-1650 Chemistry System, Erlangen, Germany). Plasma and serum samples were collected and immediately stored in Eppendorf tubes at -80°C. 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) and HbA1c was determined by the method of automated high-performance liquid chromatography analyzer (Bio-Rad, Hercules, CA). Based on OGTT and HbA1c tests, we identified 1,649 adults with prediabetes and complete baseline information from this cohort at baseline (Figure S3). The median follow-up was 12.4 (12.3–12.4) years.

METHOD DETAILS

Definitions of prediabetes

Prediabetes was defined according to the American Diabetes Association 2010 criteria as follows,[34] among participants without diabetes, FPG of 5.6–6.9 mmol/L, PBG of 7.8–11.0 mmol/L, or HbA1c of 5.7%–6.4%.

Outcome assessment

Diabetes: In the 4C study, incident diabetes was defined as FPG ≥7.0 mmol/L, 2h post-load glucose ≥11.1 mmol/L, or HbA1c ≥ 6.5%, or physician-diagnosed diabetes.[29] In the SN_2009-2021 study, incident diabetes was defined as physician-diagnosed diabetes based on National Health Insurance System.

CKD: 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 ≥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.[35,36]

CVD: 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. 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.[3]

QUANTIFICATION AND STATISTICAL ANALYSIS

Baseline characteristics of the study participants were presented as mean ± SD or median (interquartile range) for continuous variables, and number (percentage) for categorical variables.

Data dimensionality reduction with DDRTree

Missing values or outliers exceeding 5 SD from the mean were excluded, followed by rank normalization of the data. We further adjusted each variable for age and gender using linear regression analysis. The resulting age-and gender-residualized matrix of variables was input into the DDRTree algorithm from Monocle R package to reduce the multi-dimensional residualized matrix into a two-dimensional non-linear tree structure.[37] A smoothed graph was constructed from the reduced data, with similar data points positioned adjacently to form a tree structure with branches. Individuals located in the tree trunk represented the central cluster of the data, characterized by mixed variables, while increasingly similar and distinctive phenotypic patterns were observed toward the distal branches. The output was a 2D tree structure and each point represented one individual, which allowed for the continuous visualization of phenotypic profile and disease risks along the tree. To facilitate interpretation, we assigned individuals to four phenotypes based on the DDRTree branches they located. However, as this was a soft clustering approach, membership across phenotypes existed on a continuum rather than representing discrete homogeneous groups.

Risk estimates of prediabetes-related outcome

To evaluate the individual risk of each prediabetes progression outcome over a 5-year follow-up period, we obtained predicted probabilities from Cox hazard risk models using the dimension 1 and 2 of the tree as independent variables and further overlaid it on the tree diagram to visualize the heterogeneity in prediabetes progression. The dimension 1 and 2 derived from the soft-clustering projection represent individuals' positions in the reduced phenotypic space. These dimensions are continuous numerical variables without a predefined clinical scale, reflecting underlying variation across multiple metabolic traits. Effect estimates (HRs or ORs) were calculated per one-unit increase in each dimension. The proportional hazards assumption for the Cox models was assessed using Schoenfeld residuals.

Quantitative the distribution pattern across the tree

Beyond the visual gradient, we used two metrics to assess the distribution pattern of the phenotypic profile and probabilities across the tree. First, we regressed each clinical variable against the tree dimensions and plotted the regression coefficients and 95% CIs.

For each outcome, we derived HRs with 95% CIs from the Cox models, indicating how the risk of prediabetes progression outcomes varied across the two-dimensional space (x axis and y axis). Second, we conducted spatial autocorrelation analysis using global and local Moran's I-statistics.[38] A positive Moran's I value suggests that similar values (variable/event probabilities) cluster together, whereas a negative value indicates that dissimilar values are located in close proximity.

Constructing mapping function for external validation

For external validation in SN_2009-2021 study, we constructed a "mapping function" to map individuals with prediabetes to the 4C reference tree. This function uses twelve prediabetes-related clinical variables (WHR, BMI, HOMA-IR, HOMA-B, ALT, AST, GGT, FPG, PBG, HbA1c, TG, and HDL-C), while accounting for age and gender, to assign each individual a position in the 4C tree. The mapping function comprises two components: (1) two generalized additive models with smooth terms fitted with cubic regression splines to predict the two DDRTree dimensions with twelve variables, age and gender as covariates, which had great performance with adjusted R^2 values of 0. 913 and 0.758, respectively Table S2); and (2) a distance-estimating algorithm to calculate the Euclidean distance between points in the two-dimensional space. In brief, we utilized individuals' clinical information (i.e., the original clustering variables) to obtain temporary 2D coordinates. Each individual was then assigned to the nearest node on the original DDRTree based on Euclidean distance. For an individual, when all twelve variables, as well as age and gender are provided, a personal position on the tree was determined by the trained spline model and its' distance to all 55,777 nodes in the 4C tree were calculated by the distance-estimating algorithm to assign this individual to the nearest node. All 1,649 with prediabetes were then mapped on the 4C tree by this process. Then, the variables and predicted probabilities of prediabetes-related outcomes from SN_2009-2021 cohort were overlaid onto this tree, which show similar distribution across the tree compared to the 4C cohort. Finally, Cox proportional hazard models were constructed using a similar approach as in the 4C study for T2DM and CVD outcomes, while a logistic regression model was employed to assess the risk and probability of CKD. The HR/OR values presented in the figures reflect the occurrence of the outcome event (0/1) across the distribution of the tree structure. Specifically, these are based on each individual's coordinates along dimension 1 or dimension 2, capturing how the risk of the outcome changes within this two-dimensional space.

All statistical analyses were performed using R software (version 4.3.3).