Received: 10 August 2024 | Revised: 27 September 2024 | Accepted: 6 October 2024

DOI: 10.1111/dom.16022

ORIGINAL ARTICLE

Subphenotypes of adult-onset diabetes: Data-driven clustering in the population-based KORA cohort

Qiuling Dong MSc[1,2,3] | Yue Xi MSc[2,3] | Stefan Brandmaier PhD[1,2] | Markéta Fuchs MSc[1,2] | Marie-Theres Huemer PhD[2] | Melanie Waldenberger PhD[1,2] | Jiefei Niu MSc[1,2] | Christian Herder PhD[4,5,6] | Wolfgang Rathmann MD[4,7] | Michael Roden MD[4,5,6] | Wolfgang Koenig MD[8,9,10] | Gidon J. Bönhof PhD[4,5] | Christian Gieger PhD[1,2,11] | Barbara Thorand PhD[2,4,12] | Annette Peters PhD[2,4,8,12] | Susanne Rospleszcz PhD[2,8,12,13] | Harald Grallert PhD[1,2,11]

[1]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany

[2]Institute of Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany

[3]Pettenkofer School of Public Health, Faculty of Medicine, LMU Munich, Munich, Germany

[4]German Center for Diabetes Research (DZD), Düsseldorf, Germany

[5]Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[6]Department of Endocrinology and Diabetology, Medical Faculty and University Hospital Dusseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[7]Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[8]German Research Center for Cardiovascular Disease (DZHK), Partner site Munich Heart Alliance, Munich, Germany

[9]Deutsches Herzzentrum München, Technische Universität München, Munich, Germany

[10]Institute of Epidemiology and Medical Biometry, University of Ulm, Ulm, Germany

[11]German Center for Diabetes Research (DZD), Neuherberg, Germany

[12]Chair of Epidemiology, Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), Faculty of Medicine, Ludwig-Maximilians-University München, Munich, Germany

[13]Department of Diagnostic and Interventional Radiology, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

Correspondence
Qiuling Dong and Harald Grallert, Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, 85764 Neuherberg, Germany.
Email: qiuling.dong@helmholtz-munich.de and harald.grallert@helmholtz-munich.de

Funding information
The German Center for Diabetes Research (DZD); German Federal Ministry of Education and Research; Munich Center of Health Sciences; Ludwig-Maximilians-Universität München; Helmholtz Munich—German Research Center for Environmental Health;

Abstract

Aims: A data-driven cluster analysis in a cohort of European individuals with type 2 diabetes (T2D) has previously identified four subgroups based on clinical characteristics. In the current study, we performed a comprehensive statistical assessment to (1) replicate the above-mentioned original clusters; (2) derive de novo T2D subphenotypes in the Kooperative Gesundheitsforschung in der Region Augsburg (KORA) cohort and (3) describe underlying genetic risk and diabetes complications.

Methods: We used data from $n = 301$ individuals with T2D from KORA FF4 study (Southern Germany). Original cluster replication was assessed forcing $k = 4$ clusters

Susanne Rospleszcz and Harald Grallert contributed equally to this study.

China Scholarship Council, Grant/Award Number: 202008310176; State of Bavaria; German Federal Ministry of Health; Ministry of Culture and Science of the State North Rhine-Westphalia

using three different hyperparameter combinations. De novo clusters were derived by open *k*-means analysis. Stability of de novo clusters was assessed by assignment congruence over different variable sets and Jaccard indices. Distribution of polygenic risk scores and diabetes complications in the respective clusters were described as an indication of underlying heterogeneity.

Results: Original clusters did not replicate well, indicated by substantially different assignment frequencies and cluster characteristics between the original and current sample. De novo clustering using $k = 3$ clusters and including high sensitivity C-reactive protein in the variable set showed high stability (all Jaccard indices >0.75). The three de novo clusters ($n = 96$, $n = 172$, $n = 33$, respectively) adequately captured heterogeneity within the sample and showed different distributions of polygenic risk scores and diabetes complications, that is, cluster 1 was characterized by insulin resistance with high neuropathy prevalence, cluster 2 was defined as age-related diabetes and cluster 3 showed highest risk of genetic and obesity-related diabetes.

Conclusion: T2D subphenotyping based on its sample's own clinical characteristics leads to stable categorization and adequately reflects T2D heterogeneity.

KEYWORDS
clustering, cohort study, database research, diabetes complications, type 2 diabetes

1 | INTRODUCTION

Diabetes is a rapidly growing global health concern.[1,2] The underlying causes of pancreatic beta-cell dysfunction are heterogeneous, and individual trajectories of hyperglycaemia and subsequent diabetes complications vary widely.[3,4] Therefore, classifications of type 2 diabetes (T2D) that predict the risk of complications and provide options for a tailored treatment have been actively studied.[5–8]

Traditionally, diabetes is mainly classified into type 1 (T1D) and T2D, primarily determined by the presence (T1D) or absence (T2D) of autoantibodies. A novel approach to identify subphenotypes of diabetes was the hallmark study by Ahlqvist et al.[9] They used six diabetes-related variables including age at diagnosis, body mass index (BMI), haemoglobin A1c (HbA1c), homeostasis model assessment (HOMA) estimates of beta-cell function (HOMA2-B) and insulin resistance (HOMA2-IR) and glutamic acid decarboxylase antibodies (GADA) to categorize individuals with diabetes into five clusters. Thereby, four clusters mainly represent T2D subphenotypes and one cluster with severe autoimmune diabetes (SAID) mainly corresponds to the T1D subphenotype. The four T2D subphenotypes were labelled based on their distinctive features as severe insulin-deficient diabetes (SIDD), severe insulin resistant diabetes (SIRD), mild obesity-related diabetes (MOD) and mild age-related diabetes (MARD) and exhibited different risks of disease progression and diabetes complications. These clusters have been replicated in diverse ethnic groups such as British,[10] German,[11,12] American and Chinese,[13,14] Mexican,[15] Icelandic,[16] Japanese[17] and Asian Indian cohorts.[18] Recently, subphenotypes were characterized in more detail from a molecular perspective, including potential underlying genetic determinants[19,20] and cluster-specific signatures of metabolomics and proteomics.[21,22] There appear to be differences in biomarkers of inflammation between diabetes subphenotypes, which is in line with the involvement of inflammatory mechanisms, most often assessed by C-reactive protein (CRP), in the progression of diabetes.[12,23] Taken together, the current state of evidence suggests that diabetes subphenotyping, including deep molecular phenotyping, holds the potential to offer key insights into the underlying pathophysiology of glucose dysregulation and the onset of comorbidities among individuals with T2D, while it further enables the advancement of personalized treatment of diabetes.

In the current study, we aimed to perform a comprehensive statistical assessment of T2D subphenotyping in the Cooperative Health Research in the Region of Augsburg (KORA) FF4 cohort (Southern Germany). Our aims were threefold: (1) to investigate to which extent the original clusters from Ahlqvist et al.[9] could be replicated in the KORA sample; (2) to derive novel T2D subphenotypes based on data-driven clustering, also accounting for inflammation and (3) to investigate heterogeneity between the de novo derived subphenotypes by describing the distribution of genetically predicted risk as captured by a polygenic risk score (PRS), diabetes-related complications and parental history of diabetes. An overview of the study design is shown in Figure 1.

2 | METHODS

2.1 | Study population and clinical data

KORA comprises several deeply phenotyped population-based epidemiological surveys.[24] The current analysis is based on data from the KORA-FF4 study, conducted between 2013 and 2014. Details about

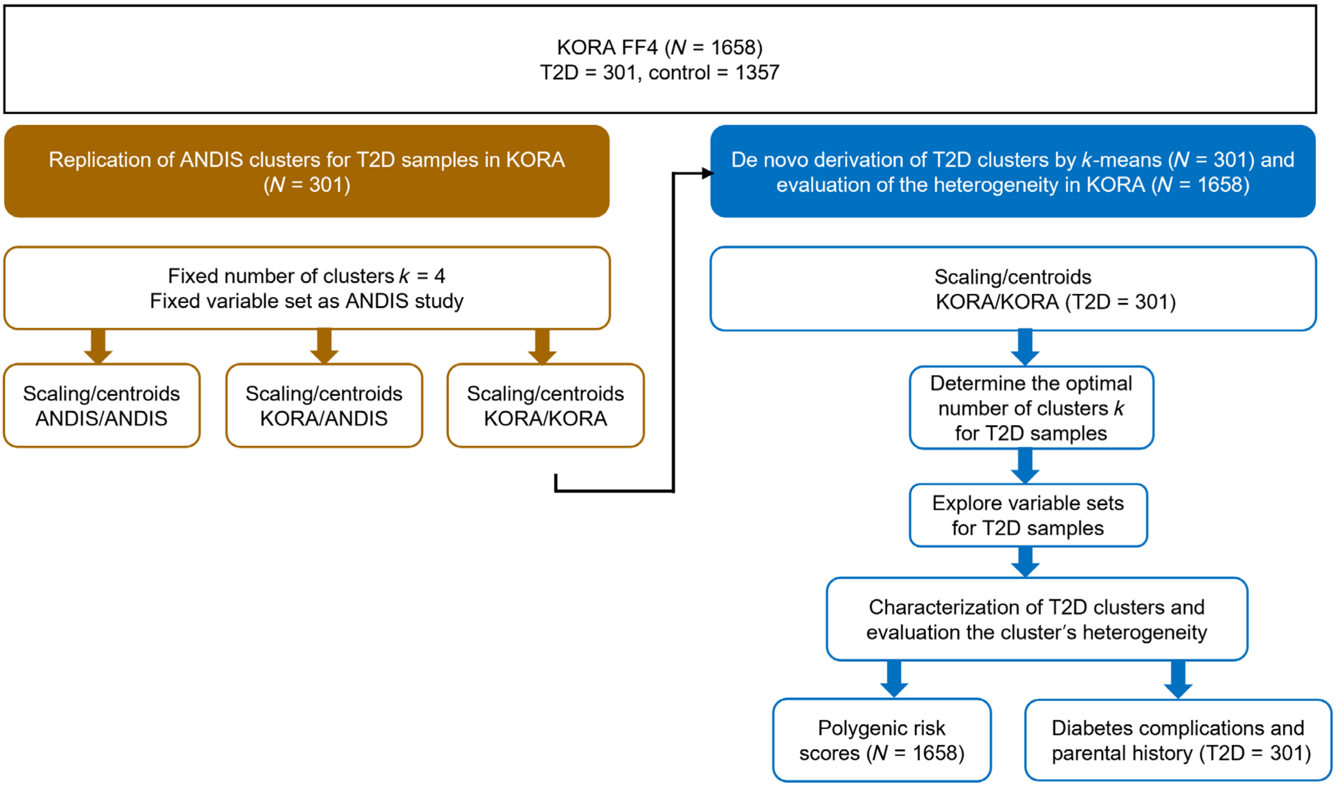
KORA FF4 (N = 1658)
T2D = 301, control = 1357
Replication of ANDIS clusters for T2D samples in KORA (N = 301)
Fixed number of clusters k = 4
Fixed variable set as ANDIS study
Scaling/centroids ANDIS/ANDIS
Scaling/centroids KORA/ANDIS
Scaling/centroids KORA/KORA
De novo derivation of T2D clusters by k-means (N = 301) and evaluation of the heterogeneity in KORA (N = 1658)
Scaling/centroids KORA/KORA (T2D = 301)
Determine the optimal number of clusters k for T2D samples
Explore variable sets for T2D samples
Characterization of T2D clusters and evaluation the cluster's heterogeneity
Polygenic risk scores (N = 1658)
Diabetes complications and parental history (T2D = 301)

FIGURE 1 Study design. The left part in orange corresponds to aim (1) whereas the right part in blue corresponds to aims (2) and (3). The fixed variable set contained the basic variables: Age, body mass index, haemoglobin A1c, homeostasis model assessment (HOMA) estimates of beta-cell function (HOMA2-B) and insulin resistance (HOMA2-IR), corresponding to the original ANDIS study. Variable sets in KORA contained the basic variables plus one additional variable, respectively: High sensitivity C-reactive protein, triglycerides, HDL-cholesterol or systolic blood pressure. ANDIS, Swedish All New Diabetics in Scania cohort; KORA, 'Cooperative Health Research in the Region of Augsburg' cohort; T2D, type 2 diabetes; PRS, polygenic risk score.

TABLE 1 Characteristics of the KORA FF4 participants for men and women.

	Men ($N = 178$)	Women ($N = 123$)	p
Age at examination (years) mean (SD)	69.6 (10.0)	69.4 (10.2)	0.83
BMI (kg/m^2) mean (SD)	30.3 (4.9)	32.2 (5.7)	0.003
HbA1c (mmol/mol) mean (SD)	46.9 (11.7)	47.9 (10.9)	0.48
HOMA2-B % mean (SD)	72.3 (38.4)	71.5 (31.8)	0.85
HOMA2-IR (SD)	2.1 (1.2)	2.0 (1.0)	0.52
hsCRP (mg/L) mean (SD)	2.7 (3.3)	4.5 (6.0)	0.001
TG (mmol/L) mean (SD)	1.9 (1.1)	1.6 (0.8)	0.025
HDL-C (mmol/L) mean (SD)	1.4 (0.4)	1.6 (0.4)	<0.001
SBP (mmHg) mean (SD)	130.4 (17.8)	122.3 (19.9)	<0.001
Fasting glucose (mmol/L) mean (SD)	7.6 (2.0)	7.5 (1.9)	0.543
Use of metformin	84 (47.2)	57 (46.3)	0.94
Any oral antidiabetic medication or insulin treatment	96 (53.9)	64 (52.0)	0.80
Known diabetes (%)	123 (69.1%)	84 (68.3%)	0.982

Note: Mean and standard deviation (SD) are provided for quantitative variables and differences were evaluated by student's t test. Count and percentage are provided for categorical variables and differences were evaluated by chi square test.

Abbreviations: BMI, body mass index; HbA1C, haemoglobin A1c; HDL-C, high-density lipoprotein cholesterol; hsCRP, high sensitivity C-reactive protein; HOMA2-B, homeostasis model assessment estimates of beta-cell function; HOMA2-IR, homeostasis model assessment estimates of insulin resistance; Known diabetes, the diabetes diagnosis was known prior to the study; TG, triglycerides; SBP, systolic blood pressure.

the study sample and the assessment of clinical data are presented in Supplementary Material.

For the current analysis, only participants with T2D were included for cluster analysis. Participants with T1D ($n = 6$) were excluded from all analyses. Moreover, participants with missing values for clustering variables (described below) were excluded ($n = 20$). Finally, the cluster analysis comprised $N = 301$ individuals with T2D (Table 1). For the assessment of genetically T2D risk, a PRS was calculated for all individuals with T2D ($N = 301$) and without T2D ($N = 1357$) as the control group (Table S1).

2.2 | Genotyping and polygenetic risk score

Genetically predicted T2D risk was calculated by an established PRS, as described in Supplementary Material.

2.3 | Statistical analysis

Statistical analysis was conducted using R version 4.1.1. A two-sided *p* value <0.05 was considered statistically significant. A detailed description of (1) the replication of original clusters, including different combinations of scaling and centroid hyperparameters, (2) de novo cluster derivation in the KORA study and (3) assessment of differences between clusters with respect to PRS, parental history of diabetes and diabetes complications is presented in Supplementary Material.

3 | RESULTS

3.1 | Study sample

The final sample included 301 individuals with T2D, thereof 94 (31.2%) with newly detected diabetes by oral glucose tolerance test (oGTT). Comparison between women and men showed higher BMI, hsCRP and HDL-C values in women and higher TG and SBP levels in men, whereas medication intake (metformin and any other oral antidiabetic medication or insulin treatment) was similar (Table 1). Fasting glucose and HbA1c values over time are presented in Figure S1.

3.2 | Replication of the four ANDIS T2D clusters

3.2.1 | Assignment by using ANDIS scaling and ANDIS centroids

First, clinical variables of the KORA participants were scaled based on ANDIS's scaling parameters, and each participant was assigned to a single cluster based on the Euclidean distance to the ANDIS centroids.[25] The characteristics of four clusters are shown in Table S2 and Figure 2A. The SIDD cluster in KORA was characterized by a relatively younger age, lower insulin secretion (HOMA2-B) and highest HbA1c; the SIRD cluster had the highest level of insulin resistance (HOMA2-IR) and insulin secretion (HOMA2-B); the MOD cluster had a high BMI but younger age and the MARD cluster showed low insulin resistance, low BMI and older age. The relative cluster sizes in KORA were not comparable to those found in the ANDIS study. SIDD made up only 2% of the T2D cases in KORA compared to 17.5% in ANDIS. More than 80% of participants in KORA were assigned to the MARD cluster, compared to only around 40% in ANDIS.

3.2.2 | Assignment by using KORA scaling and ANDIS centroids

Second, the clinical variables of KORA participants were scaled based on own scaling parameters derived from the KORA sample and then assigned to a single cluster based on the Euclidean distance to the ANDIS centroids.[21] The characteristics of four clusters are shown in Table S2 and Figure S2A. The SIDD cluster in KORA was characterized by a relatively younger age, lower insulin secretion (HOMA2-B) and poorer glycaemic control (higher HbA1c); the SIRD cluster had the highest level of insulin resistance (HOMA2-IR) and insulin secretion (HOMA2-B); the MOD cluster had a high BMI and individuals were younger and the MARD cluster had low insulin resistance and low BMI, but an older age. All these variables followed the same trend in KORA and ANDIS. The relative cluster sizes in KORA were comparable to those found in the ANDIS study, for example, most participants were allocated to MARD for both KORA (46.8%) and ANDIS (39.1%), and 15.3% of individuals in KORA were assigned to SIDD which was similar to the ANDIS study (17.5%).

We then investigated the transfer of individuals when using ANDIS centroids, with either ANDIS scaling or KORA scaling. Sixty-five percent of participants were assigned to the same clusters (Figure S2B). Compared to ANDIS scaling, clusters were more evenly distributed when using KORA scaling. Most strikingly, a substantial part of the MARD cluster when using ANDIS scaling was allocated to the SIDD, SIRD and MOD clusters using KORA scaling.

3.2.3 | Assignment by using KORA scaling and KORA centroids

Third, clusters were derived based on hyperparameters from KORA data alone, using *k*-means clustering on the same variable set (age, BMI, HbA1c, HOMA2-B and HOMA2-IR) forcing the same number of clusters ($k = 4$) as in the ANDIS cohort. As shown in Figure S3, cluster 1 was characterized by low insulin secretion (low HOMA2-B), high BMI and poor metabolic control (high HbA1c); thus, we labelled cluster 1 as SIDD. Cluster 2 had insulin resistance as evidenced by a high HOMA2-IR which could be compared to SIRD. Cluster 3 featured

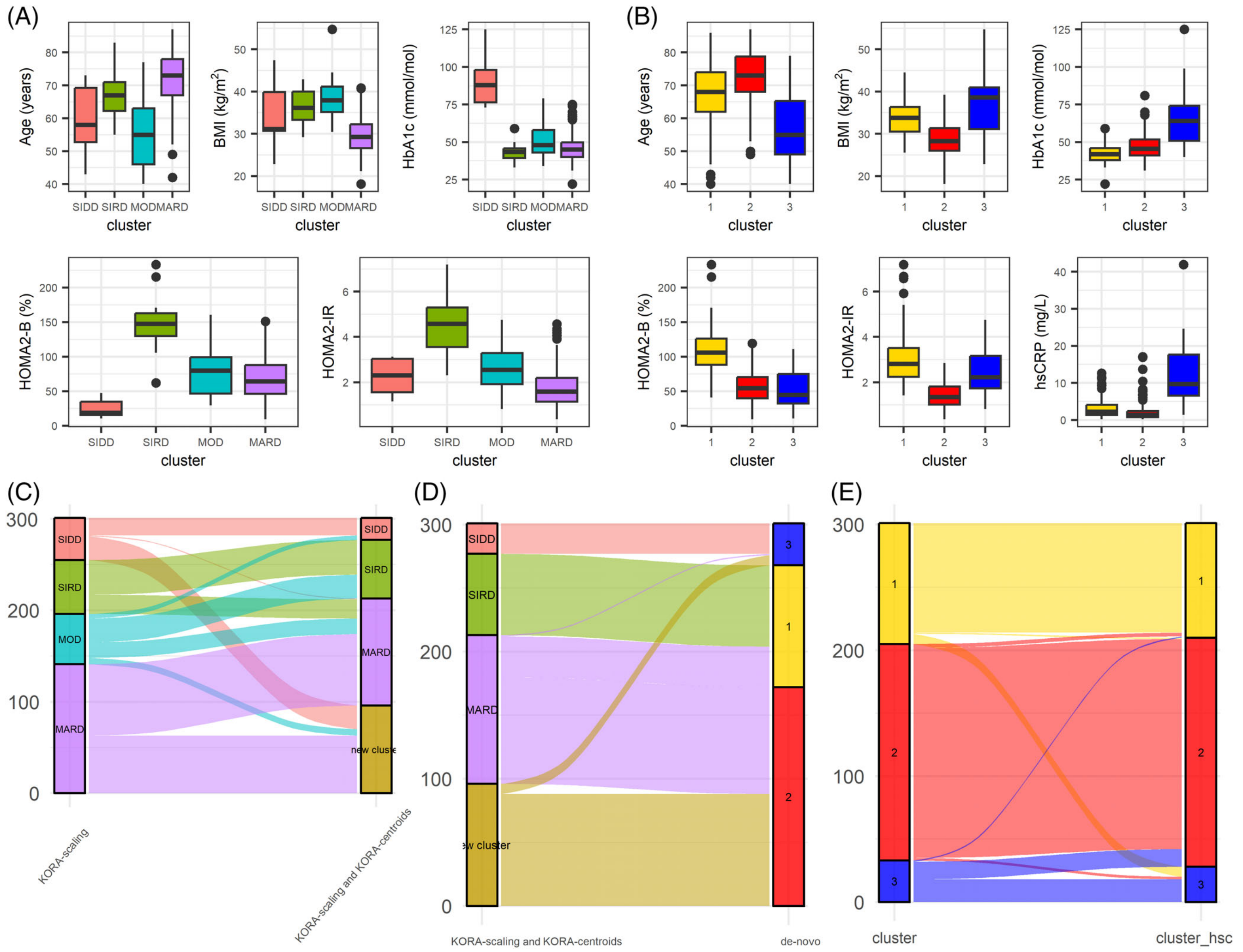

FIGURE 2 Distributions of age at examination, body mass index (BMI), haemoglobin A1c (HbA1c), homeostasis model assessment (HOMA) estimates of beta-cell function (HOMA2-B) and insulin resistance (HOMA2-IR) in the KORA FF4 cohort for each cluster (A) using ANDIS scaling and ANDIS centroids or (B) with additional high sensitivity C-reactive protein (hsCRP) derived from de novo *k*-means with $k = 3$. The upper and lower bounds of boxes represent the first and third quartiles, box centres represent the median values and circles represent outliers. (C) Sankey diagram displaying the transfer of individuals between the clusters identified using KORA scaling and ANDIS centroids (left side) and the clusters identified using KORA scaling and KORA centroids (right side), (D) transition of individuals between the clusters originally replicated using KORA scaling and KORA centroids (left side, corresponding to the right side of Figure 2C) and de novo derivation (right side) and (E) transition of individuals between the de novo clusters identified using basic variables (left side) and with the additional variable hsCRP (right side). MARD, mild age-related diabetes; MOD, mild obesity-related diabetes; SIDD, severe insulin-deficient diabetes; SIRD, severe insulin resistant diabetes.

elderly individuals with relatively mild metabolic irregularities which is similar to MARD in the ANDIS study. Cluster 4 represented a novel distinct subphenotype with the overall most modest metabolic impairments and low BMI and was thereby distinct from the ANDIS cluster MOD.

We also generated the Sankey diagram to visualize and compare the cluster assignment based on the second approach (using KORA scaling and ANDIS centroids) and the third approach (using KORA scaling and KORA centroids). We observed consistent cluster assignments for only 45% of the individuals between the second and third approach (Figure 2C). Taken together, these results suggest that the original ANDIS clusters do not fully reflect the characteristics of the KORA sample.

3.3 | De novo cluster derivation in KORA

3.3.1 | Determination of *k* and cluster derivation

Both silhouette width and the elbow plot methods agreed that $k = 3$ rather than $k = 4$ was the optimal number of clusters for the KORA data (Figure S4). Subsequently, *k*-means was used on the basic variable set to categorize KORA participants into three clusters representing three T2D subphenotypes. Clinical characteristics according to each subphenotype are shown in (Figure S5A). Cluster 1 ($n = 96$, 31.9%) was characterized by hyperinsulinemia and insulin resistance (most similar to the SIRD cluster in the ANDIS cohort); participants in cluster 2 ($n = 172$, 57.1%) had older age, low BMI and low insulin resistance which could be compared to MARD in the ANDIS cohort; and cluster 3 ($n = 33$, 11.0%) showed insulin deficiency (low HOMA2-B), high BMI and poor glycaemic control (high HbA1c), which is a distinct cluster from those present in the ANDIS cohort. We then compared participant transitions from the original cluster replication using KORA centroids and KORA scaling (third approach as described above) with the de novo derived clusters (Figure 2D). Individuals previously allocated to the MARD subphenotype were reallocated to the new cluster 1 and cluster 2. Individuals previously allocated to the novel distinct subphenotype were mainly reallocated not only to the new cluster 2, but also to the new cluster 3 (distinct). Individuals previously allocated to the SIRD subphenotype were reallocated to the new cluster 3.

3.3.2 | Different variable sets and final clusters

We assessed the stability of cluster assignments when using different sets of variables for clustering: basic variables (age at examination, BMI, HbA1c, HOMA2-B and HOMA2-IR) plus hsCRP, TG, HDL-C or SBP, respectively. In general, the addition of these variables did not substantially influence the distribution of the basic variables between clusters and did not lead to substantial transition of participants between clusters (Figure 2B,D, Figures S5, S6 and S7). In detail, 90%, 93%, 90% and 98% of participants were allocated to the same cluster when using basic variables compared to when adding hsCRP, TG, HDL-C or SBP, respectively. To account for the role of systemic inflammation in diabetes differentiation, we defined the clusters derived from the variable set of age, BMI, HbA1c, HOMA2-B and HOMA2-IR plus hsCRP as the final subphenotypes, presented in Figure 2B, Tables S3 and S4. Cluster 1 included 91 participants (30.2%) and was characterized by insulin resistance (high HOMA2-IR) and hyperinsulinemia, with a high proportion of newly diagnosed diabetes cases (most similar to the SIRD cluster in ANDIS). Cluster 2 included 182 individuals (60.5%) and was characterized by high age, low BMI and low insulin resistance (most similar to the MARD cluster in ANDIS). Cluster 3 included 28 participants (9.3%) and was characterized by a high BMI, poor glycaemic control, high level of subclinical inflammation (high hsCRP) and relative insulin deficiency, broadly resembling a typical patient seen in clinical practice (most similar to SIDD/MOD cluster in ANDIS).

The assessment of cluster stability showed that Jaccard indices of all final clusters were above 0.75, indicating reasonably high cluster stability for the final variable set (Table S5). Of note, with additional variables TG, HDL-C or SBP, stability slightly decreased for all clusters and cluster 3 even showed Jaccard indices below 0.75 (Table S5). Besides, the majority of individuals (95%) were assigned to the same cluster as in the initial data analysis, and both men and women showed the same trend on the clinical variable distribution (Figure S8), suggesting a lack of substantial sex-specific effects.

3.4 | Cluster differences in genetic risk, diabetes-related complications and parental history

3.4.1 | Polygenic risk score

The overall distribution of the PRS in the KORA FF4 sample is given in Figure 3A. Participants with T2D had significantly higher PRS values ($p < 0.001$) compared to those without T2D and were

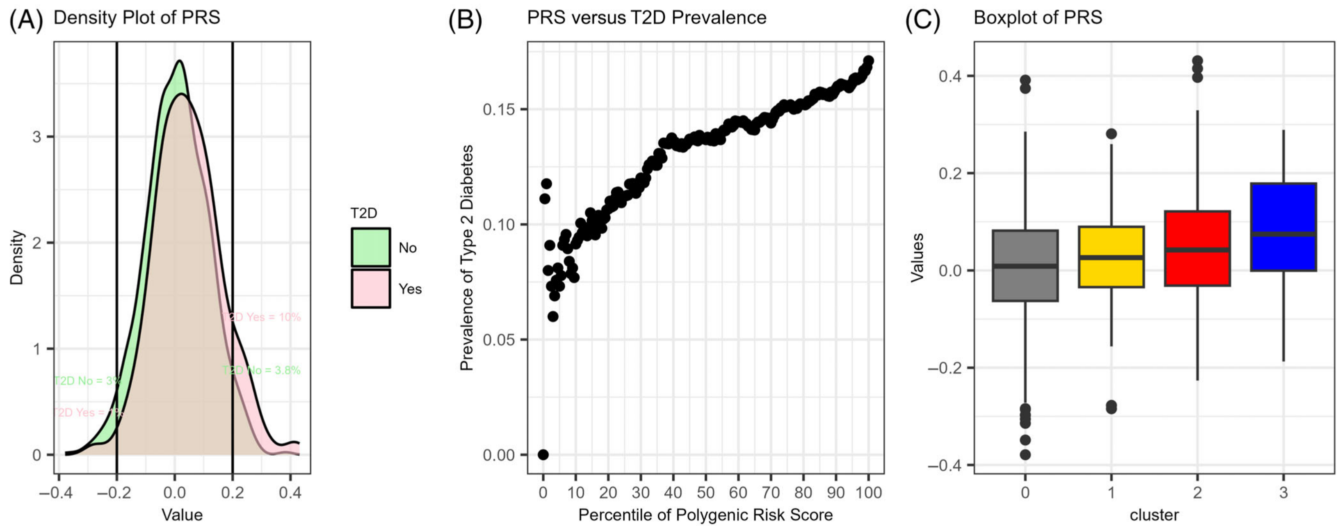

FIGURE 3 (A) Density plot shows the polygenic risk score (PRS) distribution in the KORA FF4 sample without (light green) and with (light red) type 2 diabetes (T2D). Data beyond the two vertical lines indicate extreme values of the PRS distribution, and the corresponding numbers reflect the proportion of individuals without (light green) and with (light red) T2D who showed extreme PRS values. (B) Percentile of increasing PRS (*x*-axis) versus the prevalence of T2D (*y*-axis). (C) Distributions of PRS in control group (marked as 0) and three clusters representing T2D subphenotypes.

overrepresented in the highest quantiles of the distribution (Figure 3A,B). When comparing the distribution of PRS in the respective clusters to individuals without diabetes (Figure 3C), the PRS in cluster 2 and cluster 3 was significantly different to the control group (both $p < 0.001$, respectively) but the PRS in cluster 1 was not different to the control group. An additional t test confirmed that cluster 3 had a significantly higher PRS than cluster 1 ($p = 0.034$), whereas there was no significant difference between cluster 1 and cluster 2.

3.4.2 | Diabetes-related complications and parental history of diabetes

We evaluated the prevalence of diabetes-related complications and parental history of T2D in the three clusters. As shown in Figure S9 and Table S7, in general, individuals in clusters 1 and 2 had a substantially higher prevalence of myocardial infarction, stroke and chronic kidney disease (CKD). Individuals in cluster 3 had a more frequently positive parental history of diabetes.

Moreover, compared to cluster 1, cluster 2 had a lower frequency of neuropathy ($p = 0.043$) but a higher prevalence of stroke (not significant) and CKD ($p = 0.030$).

4 | DISCUSSION

The T2D subphenotype classification scheme proposed by Ahlqvist et al.[9] has been replicated in different populations and has proven to be a useful tool to further characterize potential pathophysiological pathways and diabetes progression. Our study aimed at a comprehensive assessment of original cluster replication, including a systematic illustration of participant transitions between replicated clusters, de novo cluster derivation, including the assessment of cluster stability, and underlying genetic risk and complication distribution. We found that the original clusters only partially reflected the characteristics of individuals with T2D in the KORA sample, whereas de novo derived clusters showed excellent stability and captured the underlying heterogeneity between the T2D subphenotypes. Our results therefore underscore the importance of subphenotyping by illustrating the importance of individual study characteristics, and we contribute another potential T2D subphenotype to the existing panel.

Our results align with recent findings, which indicated that 11 of 18 studies either delineated distinct subphenotypes or failed to identify all ANDIS subphenotypes.[26] Part of the lack of replicability of the original clusters may be attributed to differences in the study setup and participants' characteristics. For example, we used age at examination for clustering, since age of diabetes onset for most T2D participants was not available. Therefore, the average age used in the KORA sample was significantly higher compared to the ANDIS cohort (Table S1), especially in the de novo cluster 2 (comparable to MARD). Moreover, individuals in KORA had better glycaemic control and less insulin resistance compared to the ANDIS sample (Table S1), indicating that KORA potentially included a larger proportion of T2D cases with less severe disease. Furthermore, our HOMA models were based on insulin instead of C-peptide, which might have led to differences in estimates. Some studies[27,28] suggested that C-peptide better reflected insulin secretion, while another study[29] suggested that both of them performed similar in evaluating beta cell function.

Employing different scaling parameters generated a big difference in cluster allocation, and different studies applied different approaches.[21,25] The incongruence of cluster assignment, together with the identification of a novel, distinct subphenotype not present in ANDIS when using KORA centroids, shows that the original clusters do not capture the characteristics of the KORA sample as well. We consider this finding important for personalized prevention. While the ANDIS cohort captured crucial subphentoypes, these clusters might take different shapes or not fully reflect the underlying sample in other cohorts with different characteristics. Contributions from multiple studies are therefore needed to expand and refine the current panel of T2D subphenotypes.

Determination of the optimal number of clusters k based on silhouette and elbow plot showed that in the KORA sample, $k = 3$ was the best number of clusters, which is consistent with the Danish DD2 study.[25] A head-to-head comparison between the clusters from KORA and DD2 revealed major similarities (Table S8). Consistent with the research from Safai et al.,[30] which did not identify an evident MOD-like cluster in their de novo cluster analysis (when using $k = 5$, including SAID), the clusters with the highest BMI also exhibited significantly higher insulin resistance. Besides the clinical characteristics used for clustering, multiple other factors are associated with T2D. We thus assessed cluster stability across different variable sets, additionally including hsCRP, HDL-C, TG or SBP, respectively. We found that these additional variables did not contribute much to the reallocation of individuals, as more than 90% individuals were still assigned to the same cluster, indicating high cluster stability and robustness towards different variable sets. One could thus hypothesize that the original variables already capture a major part of T2D heterogeneity and are adequate to identify clinically meaningful T2D subphenotypes. Other studies[18,30,31] also applied analytical approaches for a wider range of clusters or included different variables than the ANDIS study but did not systematically evaluate how participants were reallocated when using different clustering variables.

CRP is regulated by proinflammatory cytokines derived from adipose tissue.[32,33] In individuals with T2D, CRP levels are chronically elevated.[34] In the current analysis, we included hsCRP for clustering to account for the role of subclinical inflammation and assess potential differences according to subphenotypes. The de novo derived cluster 3 could not be mapped to one of the original ANDIS clusters and was characterized by high BMI, high hsCRP and relatively low HOMA2-B. Increased CRP levels have been linked to excess body weight since adipose tissue produces tumour necrosis factor α (TNF-α) and interleukin-6 (IL-6), which are pivotal factors for CRP stimulation.[32,33] We could thus hypothesize that cluster 3 represents a T2D subphenotype with chronic, obesity-induced subclinical inflammation. The PRS and the prevalence of self-reported parental history of diabetes were both the highest in cluster 3. So, cluster 3 could represent a T2D subphenotype with higher genetically induced risk for both diabetes and obesity, resulting in chronic subclinical inflammation (Table S6 and Figure S10). We note that the use of a PRS to define subgroups of diabetes is still questionable and would render the algorithm less readily applicable in clinical practice and other studies, which is why we only use it descriptively. Since non-genetic risk factors might have even stronger unfavourable impacts in individuals with genetic predisposition, the group in cluster 3 would particularly benefit from rigorous weight control, either through lifestyle modifications or drug treatment. Moreover, these individuals should be monitored for potential other causes of inflammation, such as infections or wounds.

The analysis of diabetes complications showed that in cluster 2, there was a higher proportion of CKD cases and a relatively higher percentage of stroke (not significant) compared to cluster 3. This could be due to the higher average age in cluster 2, since it is well-established that age is a major risk factor for metabolic complications in T2D.[35,36] Because risk in cluster 2 is mainly conferred by aging processes, and age is a non-modifiable factor, for this cluster in particular, close monitoring of comorbidities and strict, potentially medication-based, control of, for example, blood pressure and renal function is advisable. Cluster 1 was characterized by hyperinsulinemia and a comparatively higher prevalence of neuropathy compared to cluster 2. Insulin dysregulation can contribute to neuropathic changes in sensory neurons, and the peripheral nervous system is one of several organ systems that are profoundly affected in diabetes.[37] Interestingly, HbA1c levels in cluster 1 were comparatively low, so it would be crucial to investigate the use of glucose-lowering therapy in this cluster to evaluate their role in the prevention of neuropathy in this subphenotype. Medication therapy in this cluster was comparatively low, likely due to the high proportion of newly diagnosed diabetes cases, so this would be an obvious target to tackle insulin resistance in these individuals. Moreover, lifestyle interventions would be beneficial, including dietary changes by reducing calorie intake and limiting high glycaemic index carbohydrates and regular physical activity which enhances calorie burning and increases insulin sensitivity in muscle tissue.[38,39] Evidence indicates that an increased level of hsCRP is linked with diabetes-related complications,[40,41] but cluster 3 with the highest hsCRP levels was not characterized by a high load of complications. This may be due to the younger age of individuals in cluster 3 (Figure 2B), since given the potential pathway discussed above about a genetic predisposition to obesity-induced inflammation, it would be possible that diabetes complications in cluster 3 have not yet developed.

We acknowledge the limitations of our current study. The sample size was relatively small compared to other population-based studies, and although unsupervised clustering does not have strict sample size requirements, the small number of individuals with diabetes-related complications and family history information impedes the interpretation of shared disease characteristics. While the clusters represent a true underlying structure in the data from a statistical perspective, this structure could also have emerged due to other shared characteristics of the respective individuals, for example, environmental factors, and do not necessarily represent shared pathophysiology. Moreover, our results regarding diabetes complications need to be interpreted with caution, since complications were self-reported, and the sample size was small. We were unable to model medication effects, since medication could not be included as a variable in the clustering procedure, and participants' individual medication regimes could not be disentangled. Moreover, our participants were exclusively of white European ethnicity, which limits the generalizability to other populations.

In conclusion, to exploit the full advantages of T2D subphenotyping, a potential mismatch between reported T2D clusters and the individual study characteristics has to be taken into account. Since adapting the clustering algorithm might not always be possible, further efforts should be undertaken to identify further subtypes from different well-characterized studies, in order to expand and refine the current panel of T2D subphenotypes.

AUTHOR CONTRIBUTIONS

Conceptualization: Qiuling Dong, Yue Xi, Harald Grallert and Susanne Rospleszcz. *Methodology*: Qiuling Dong and Yue Xi. *Software*: Qiuling Dong and Yue Xi. *Validation*: Qiuling Dong and Yue Xi. *Formal analysis*: Qiuling Dong. *Investigation*: Qiuling Dong and Yue Xi. *Resources*: Susanne Rospleszcz, Stefan Brandmaier, Barbara Thorand, Marie-Theres Huemer, Melanie Waldenberger, Christian Herder, Wolfgang Rathmann, Wolfgang Koenig, Gidon J. Bönhof, Christian Gieger, Annette Peters and Harald Grallert. *Data curation*: Christian Herder, Wolfgang Rathmann, Wolfgang Koenig, Gidon J. Bönhof and Christian Gieger. *Writing—original draft preparation*: Qiuling Dong. *Writing—review and editing*: Qiuling Dong, Yue Xi, Susanne Rospleszcz, Stefan Brandmaier, Barbara Thorand, Marie-Theres Huemer, Melanie Waldenberger, Jiefei Niu, Christian Herder, Wolfgang Rathmann, Wolfgang Koenig, Gidon J. Bönhof, Christian Gieger, Annette Peters and Harald Grallert. *Visualization*: Qiuling Dong. *Supervision*: Annette Peters, Susanne Rospleszcz and Harald Grallert. *Project administration*: Harald Grallert and Annette Peters. *Funding acquisition*: Harald Grallert and Annette Peters. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGEMENTS

We express our appreciation to all KORA study participants for donating their blood and time. We thank the field staff in Augsburg conducting the KORA studies. We would like to thank the China Scholarship Council (CSC) for the financial support (No. 202008310176).

FUNDING INFORMATION

The KORA study was initiated and financed by Helmholtz Munich—German Research Center for Environmental Health, which is financed by the German Federal Ministry of Education and Research and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. The German Diabetes Center is funded by the German Federal Ministry of Health (Berlin, Germany) and the Ministry of Culture and Science of the State North Rhine-Westphalia (Düsseldorf, Germany) and receives additional funding from the German Federal Ministry of Education and

Research (BMBF) through the German Center for Diabetes Research (DZD e.V.).

CONFLICT OF INTEREST STATEMENT

M.R. reports receipt of consulting fees by AstraZeneca, Boehringer Ingelheim, Echosens, Eli Lilly, Madrigal, NovoNordisk and institutional research grants from Boehringer Ingelheim, Novartis Pharma, Novo-Nordisk and Nutriticia/Danone outside of the topic of this publication. W.R. reports the receipt of consulting fees for attending educational sessions or advisory boards run by AstraZeneca, Boehringer Ingelheim and NovoNordisk and institutional research grants from NovoNordisk outside of the topic of this publication. The other authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/dom.16022.

DATA AVAILABILITY STATEMENT

The KORA FF4 datasets are not publicly available but can be accessed upon application through the KORA-PASST (Project application self-service tool, https://www.helmholtz-munich.de/epi/research/cohorts/kora-cohort/data-use-and-access-via-korapasst/index.html).

INFORMED CONSENT STATEMENT

Written informed consent has been obtained from the study participants.

ORCID

Qiuling Dong https://orcid.org/0000-0002-3369-4120

REFERENCES

1. Collaboration NCDRF. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. *Lancet*. 2016;387(10027):1513-1530. doi:10.1016/S0140-6736(16)00618-8
2. World Health Organization. April 5 2023. Diabetes https://www.who.int/news-room/fact-sheets/detail/diabetes. accessed at 2024. 01.10
3. Davidson MB. Diagnosing diabetes with glucose criteria: worshiping a false god. *Diabetes Care*. 2011;34(2):524-526. doi:10.2337/dc10-1689
4. American Diabetes Association Professional Practice C. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2022. *Diabetes Care*. 2022;45(suppl 1):S17-S38. doi:10.2337/dc22-S002
5. Schwartz SS, Epstein S, Corkey BE, Grant SF, Gavin JR 3rd, Aguilar RB. The time is right for a new classification system for diabetes: rationale and implications of the beta-cell-centric classification schema. *Diabetes Care*. 2016;39(2):179-186. doi:10.2337/dc15-1585
6. Bancks MP, Casanova R, Gregg EW, Bertoni AG. Epidemiology of diabetes phenotypes and prevalent cardiovascular risk factors and diabetes complications in the National Health and Nutrition Examination Survey 2003–2014. *Diabetes Res Clin Pract*. 2019;158:107915. doi:10.1016/j.diabres.2019.107915
7. Thorens B, Rodriguez A, Cruciani-Guglielmacci C, Wigger L, Ibberson M, Magnan C. Use of preclinical models to identify markers of type 2 diabetes susceptibility and novel regulators of insulin secretion – a step towards precision medicine. *Mol Metab*. 2019;27S (Suppl):S147-S154. doi:10.1016/j.molmet.2019.06.008
8. Herder C, Roden M. A novel diabetes typology: towards precision diabetology from pathogenesis to treatment. *Diabetologia*. 2022;65(11):1770-1781. doi:10.1007/s00125-021-05625-x
9. Ahlqvist E, Storm P, Karajamaki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol*. 2018;6(5):361-369. doi:10.1016/S2213-8587(18)30051-2
10. 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):442-451. doi:10.1016/S2213-8587(19)30087-7
11. Zaharia OP, Strassburger K, Strom A, et al. Risk of diabetes-associated diseases in subgroups of patients with recent-onset diabetes: a 5-year follow-up study. *Lancet Diabetes Endocrinol*. 2019;7(9):684-694. doi:10.1016/S2213-8587(19)30187-1
12. Herder C, Maalmi H, Strassburger K, et al. Differences in biomarkers of inflammation between novel subgroups of recent-onset diabetes. *Diabetes*. 2021;70(5):1198-1208. doi:10.2337/db20-1054
13. Zou X, Zhou X, Zhu Z, Ji L. Novel subgroups of patients with adult-onset diabetes in Chinese and US populations. *Lancet Diabetes Endocrinol*. 2019;7(1):9-11. doi:10.1016/S2213-8587(18)30316-4
14. Li X, Yang S, Cao C, et al. Validation of the Swedish diabetes re-grouping scheme in adult-onset diabetes in China. *J Clin Endocrinol Metab*. 2020;105(10):e3519-e3528. doi:10.1210/clinem/dgaa524
15. Bello-Chavolla OY, Bahena-Lopez JP, Vargas-Vazquez A, et al. Clinical characterization of data-driven diabetes subgroups in Mexicans using a reproducible machine learning approach. *BMJ Open Diabetes Res Care*. 2020;8(1):e001550. doi:10.1136/bmjdrc-2020-001550
16. Gudmundsdottir V, Zaghlool SB, Emilsson V, et al. Circulating protein signatures and causal candidates for type 2 diabetes. *Diabetes*. 2020;69(8):1843-1853. doi:10.2337/db19-1070
17. Tanabe H, Saito H, Kudo A, et al. Factors associated with risk of diabetic complications in novel cluster-based diabetes subgroups: a Japanese retrospective cohort study. *J Clin Med*. 2020;9(7):2083. doi:10.3390/jcm9072083
18. Anjana RM, Baskar V, Nair ATN, et al. Novel subgroups of type 2 diabetes and their association with microvascular outcomes in an Asian Indian population: a data-driven cluster analysis: the INSPIRED study. *BMJ Open Diabetes Res Care*. 2020;8(1):e001506.
19. Mansour Aly D, Dwivedi OP, Prasad RB, et al. Genome-wide association analyses highlight etiological differences underlying newly defined subtypes of diabetes. *Nat Genet*. 2021;53(11):1534-1542. doi:10.1038/s41588-021-00948-2
20. Zaharia OP, Strassburger K, Knebel B, et al. Role of patatin-like phospholipase domain-containing 3 gene for hepatic lipid content and insulin resistance in diabetes. *Diabetes Care*. 2020;43(9):2161-2168. doi:10.2337/dc20-0329
21. Zaghlool SB, Halama A, Stephan N, et al. Metabolic and proteomic signatures of type 2 diabetes subtypes in an Arab population. *Nat Commun*. 2022;13(1):7121. doi:10.1038/s41467-022-34754-z
22. Wesolowska-Andersen A, Brorsson CA, Bizzotto R, et al. Four groups of type 2 diabetes contribute to the etiological and clinical heterogeneity in newly diagnosed individuals: an IMI DIRECT study. *Cell Rep Med*. 2022;3(1):100477. doi:10.1016/j.xcrm.2021.100477
23. Stanimirovic J, Radovanovic J, Banjac K, et al. Role of C-reactive protein in diabetic inflammation. *Mediators Inflamm*. 2022;2022:3706508. doi:10.1155/2022/3706508

24. Holle R, Happich M, Lowel H, Wichmann HE; Group MKS. KORA – a research platform for population based health research. *Gesundheitswesen*. 2005;67(suppl 1):S19-S25. doi:10.1055/s-2005-858235
25. Christensen DH, Nicolaisen SK, Ahlqvist E, et al. Type 2 diabetes classification: a data-driven cluster study of the Danish Centre for Strategic Research in Type 2 Diabetes (DD2) cohort. *BMJ Open Diabetes Res Care*. 2022;10(2):e002731. doi:10.1136/bmjdrc-2021-002731
26. Varghese JS, Carrillo-Larco RM, Narayan KV. Achieving replicable subphenotypes of adult-onset diabetes. *Lancet Diabetes Endocrinol*. 2023;11(9):635-636. doi:10.1016/S2213-8587(23)00195-X
27. Song YS, Hwang YC, Ahn HY, Park CY. Comparison of the usefulness of the updated homeostasis model assessment (HOMA2) with the original HOMA1 in the prediction of type 2 diabetes mellitus in Koreans. *Diabetes Metab J*. 2016;40(4):318-325. doi:10.4093/dmj.2016.40.4.318
28. Caumo A, Perseghin G, Brunani A, Luzi L. New insights on the simultaneous assessment of insulin sensitivity and beta-cell function with the HOMA2 method. *Diabetes Care*. 2006;29(12):2733-2734. doi:10.2337/dc06-0070
29. Li X, Zhou ZG, Qi HY, Chen XY, Huang G. Replacement of insulin by fasting C-peptide in modified homeostasis model assessment to evaluate insulin resistance and islet beta cell function. *Zhong Nan Da Xue Xue Bao Yi Xue Ban*. 2004;29(4):419-423.
30. Safai N, Ali A, Rossing P, Ridderstrale M. Stratification of type 2 diabetes based on routine clinical markers. *Diabetes Res Clin Pract*. 2018; 141:275-283. doi:10.1016/j.diabres.2018.05.014
31. Slieker RC, Donnelly LA, Fitipaldi H, et al. Replication and cross-validation of type 2 diabetes subtypes based on clinical variables: an IMI-RHAPSODY study. *Diabetologia*. 2021;64(9):1982-1989. doi:10.1007/s00125-021-05490-8
32. Freeman DJ, Norrie J, Caslake MJ, et al. C-reactive protein is an independent predictor of risk for the development of diabetes in the west of Scotland coronary prevention study. *Diabetes*. 2002;51(5):1596-1600. doi:10.2337/diabetes.51.5.1596
33. Sproston NR, Ashworth JJ. Role of C-reactive protein at sites of inflammation and infection. *Front Immunol*. 2018;9:754. doi:10.3389/fimmu.2018.00754
34. Kanmani S, Kwon M, Shin MK, Kim MK. Association of C-reactive protein with risk of developing type 2 diabetes mellitus, and role of obesity and hypertension: a large population-based Korean cohort study. *Sci Rep*. 2019;9(1):4573. doi:10.1038/s41598-019-40987-8
35. Bhatti GK, Bhadada SK, Vijayvergiya R, Mastana SS, Bhatti JS. Metabolic syndrome and risk of major coronary events among the urban diabetic patients: North Indian Diabetes and Cardiovascular Disease Study—NIDCVD-2. *J Diabetes Complications*. 2016;30(1):72-78.
36. Shamshirgaran SM, Mamaghanian A, Aliasgarzadeh A, Aiminisani N, Iranparvar-Alamdari M, Ataie J. Age differences in diabetes-related complications and glycemic control. *BMC Endocr Disord*. 2017;17(1): 25. doi:10.1186/s12902-017-0175-5
37. Grote CW, Wright DE. A role for insulin in diabetic neuropathy. *Front Neurosci*. 2016;10:581. doi:10.3389/fnins.2016.00581
38. Racz O, Linkova M, Jakubowski K, Link R, Kuzmova D. Az inzulinkezeles elkezdesenek gyakorlati akadalyai 2-es tipusu cukorbetegekben – a "pszichologiai inzulinrezisztencia" lekuzdese (Barriers of the initiation of insulin treatment in type 2 diabetic patients – conquering the "psychological insulin resistance"). *Orv Hetil*. 2019;160(3):93-97. doi:10.1556/650.2019.31269
39. Yaribeygi H, Atkin SL, Simental-Mendia LE, Sahebkar A. Molecular mechanisms by which aerobic exercise induces insulin sensitivity. *J Cell Physiol*. 2019;234(8):12385-12392. doi:10.1002/jcp.28066
40. Mugabo Y, Li L, Renier G. The connection between C-reactive protein (CRP) and diabetic vasculopathy. Focus on preclinical findings. *Curr Diabetes Rev*. 2010;6(1):27-34. doi:10.2174/157339910790442628
41. Esser N, Paquot N, Scheen AJ. Inflammatory markers and cardiometabolic diseases. *Acta Clin Belg*. 2015;70(3):193-199. doi:10.1179/2295333715Y.0000000004

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Dong Q, Xi Y, Brandmaier S, et al. Subphenotypes of adult-onset diabetes: Data-driven clustering in the population-based KORA cohort. *Diabetes Obes Metab*. 2024;1-10. doi:10.1111/dom.16022