BMJ Open Diabetes Research & Care

Multilevel clustering approach driven by continuous glucose monitoring data for further classification of type 2 diabetes

Rui Tao,[1] Xia Yu,[1] Jingyi Lu,[2] Yun Shen,[2] Wei Lu,[2] Wei Zhu,[2] Yuqian Bao ,[2] Hongru Li,[1] Jian Zhou [2]

To cite: Tao R, Yu X, Lu J, *et al*. Multilevel clustering approach driven by continuous glucose monitoring data for further classification of type 2 diabetes. *BMJ Open Diab Res Care* 2021;**9**:e001869. doi:10.1136/bmjdrc-2020-001869

► Supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjdrc-2020-001869).

RT, XY and JL contributed equally.

Received 28 August 2020
Revised 7 January 2021
Accepted 3 February 2021

[1]College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning, China
[2]Department of Endocrinology and Metabolism, Shanghai Jiao Tong University Affilated Sixth People's Hospital, Shanghai Clinical Center for Diabetes, Shanghai, China

Correspondence to
Professor Jian Zhou;
zhoujian@sjtu.edu.cn

ABSTRACT

Introduction Mining knowledge from continuous glucose monitoring (CGM) data to classify highly heterogeneous patients with type 2 diabetes according to their characteristics remains unaddressed. A refined clustering method that retrieves hidden information from CGM data could provide a viable method to identify patients with different degrees of dysglycemia and clinical phenotypes.

Research design and methods From Shanghai Jiao Tong University Affiliated Sixth People's Hospital, we selected 908 patients with type 2 diabetes (18–83 years) who wore blinded CGM sensors (iPro2, Medtronic, California, USA). Participants were clustered based on CGM data during a 24-hour period by our method. The first level extracted the knowledge-based and statistics-based features to describe CGM signals from multiple perspectives. The Fisher score and variables cluster analysis were applied to fuse features into low dimensions at the second level. The third level divided subjects into subgroups with different clinical phenotypes. The four subgroups of patients were determined by clinical phenotypes.

Results Four subgroups of patients with type 2 diabetes with significantly different statistical features and clinical phenotypes were identified by our method. In particular, individuals in cluster 1 were characterized by the lowest glucose level factor and glucose fluctuation factor, and the highest negative glucose factor and C peptide index. By contrast, cluster 2 had the highest glucose level factor and the lowest C peptide index. Cluster 4 was characterized by the greatest degree of glucose fluctuation factor, was the most insulin-sensitive, and had the lowest insulin resistance. Cluster 3 ranked in the middle concerning the CGM-derived metrics and clinical phenotypes compared with those of the other three groups.

Conclusion A novel multilevel clustering approach for knowledge mining from CGM data in type 2 diabetes is presented. The results demonstrate that subgroups are adequately distinguished with notable statistical and clinical differences.

Significance of this study

What is already known about this subject?

- A refined clustering method that retrieves hidden information from continuous glucose monitoring (CGM) data could provide a viable method to identify patients with different features of diabetes.
- Type 2 diabetes mellitus is highly heterogeneous and traditional classification method may be too crude to meet the needs of personalized medicine.

What are the new findings?

- Our study shows that it is feasible to implement the data-driven method of artificial intelligence in diabetes research.
- It is a novel research that uses multilevel clustering method for mining knowledge from original high-dimensional CGM data to further classify patients with type 2 diabetes, which is difficult for clinicians.
- In particular, the results showed that the four novel subgroups of patients with type 2 diabetes had distinct clinical phenotypes.

How might these results change the focus of research or clinical practice?

- By the new classification of type 2 diabetes, personalized strategies of diabetes management may be developed, which may lead to better health outcomes in diabetes.
- Artificial intelligence could be an effective tool for precision medicine in diabetes.

INTRODUCTION

Diabetes and its complications can substantially impair the human body. Its treatment has always been a long-standing dilemma worldwide.[1 2] Type 2 diabetes not only comprises 90% of all diabetes, but also has considerable heterogeneity in genes and clinical phenotypes.[3 4] It is generally believed that further classification of type 2 diabetes will provide more personalized and precise treatments for patients.[3]

Cluster analysis, a data-mining technique that helps reveal hidden structures, has been widely used in the medical field.[5 6] Previous works using cluster analyses to distinguish patients with type 2 diabetes have largely concentrated on medical check-up data and

genetic data.[3 4 7 8] For instance, five subgroups of adult-onset diabetes were obtained by six check-up variables by Ahlqvist *et al*.[3] Further analyses have proven that the risk of diabetes complications among these five subgroups was different. On the other hand, continuous glucose monitoring (CGM) is a method of continuously measuring glucose levels over a number of days. Previous studies have shown that the use of CGM improves glucose control in patients with diabetes.[9–11] There are several ways in which CGM functions. The glucose values obtained from CGM can be blinded to the user (blinded/retrospective CGM) or viewed in real time (real-time CGM). Importantly, the wealth of information on glucose profile generated by CGM provides an opportunity to discover latent structures within type 2 diabetes.

To the best of our knowledge, empirical research on classification of patients with type 2 diabetes from CGM data is limited.[12 13] By dividing the complete CGM curve into segments, Hall and colleagues[12] found three different fluctuation modes of CGM, and the frequencies of the three modes in individuals without diabetes, patients with pre-diabetes and patients with diabetes were finally found to be different. Kahkoska and colleagues[13] used eight CGM features to identify new subgroups of type 1 diabetes. Three subgroups showed significant differences in glycated hemoglobin A1c (HbA1c).[13] The problem of using CGM indices to stratify patients with type 2 diabetes among different categories needs to be discussed further.

As a result, examining patients with type 2 diabetes by CGM data and clustering these patients into specific groups may provide useful information for decision-making. In this study, we proposed a new data-driven method to further classify patients with type 2 diabetes into different subgroups from the original CGM data. We attempt to identify these distinct subgroups and to look at the clinical phenotypes of this clustering.

METHODS

Study population

Patients who were admitted to the Department of Endocrinology and Metabolism, Shanghai Jiao Tong University Affiliated Sixth People's Hospital from January 2018 to the end of December 2018 were consecutively enrolled. They were diagnosed with type 2 diabetes according to the 2010 American Diabetes Association standards.[14] This study was a secondary analysis of the CGM and clinical data. For cluster analysis of patients with type 2, a blinded CGM system (iPro2, Medtronic, California, USA) was used in this study. The sensor of the CGM system was inserted on day 0 and removed after 72 hours. The CGM data were calibrated by the fingerstick blood glucose readings no more than every 12 hours. Each participant wore the CGM system for 3 days, and complete data for each participant were extracted on the second day (0:00–24:00). Patients met the following study inclusion criteria: hospitalized patients, age ≥18 years, presence of type 2 diabetes, stable glucose-lowering regimen over the previous 3 months and no data missing in CGM on day 2 and clinical phenotypes data. Exclusion criteria included diabetic ketoacidosis; hyperglycemic hyperosmolar state or severe and recurrent events within the previous 3 months; and history of malignancy or mental disorders. Finally, we succeeded in obtaining complete data for 908 participants. All patients provided written informed consent.

Clinical phenotypes

In our study, each patient underwent physical examination that included measurements of height, weight, and blood pressure. Body mass index (BMI) was calculated as weight (kilograms) divided by squared height (meters). One day before the CGM monitoring period, a venous blood sample was drawn at 06:00 after a 10-hour overnight fast. Fasting C peptide (FCP), HbA1c, high-density lipoprotein cholesterol and low-density lipoprotein cholesterol were determined as previously reported.[15] In addition to FCP, C peptide at 120 min (CP2h) during a standard meal following overnight fast was assayed. The increments of C peptide in plasma levels at 120 min (ΔCP2h=CP2h−FCP) showed the ratio of C peptide response. Homeostasis model assessment (HOMA) indices were added to analyze each cluster according to the HOMA V.2.2.3 calculator.[16]

Multilevel clustering method

The multilevel clustering system was designed as a four-step data-mining cluster model. The first level, features extraction, aimed to extract knowledge-based and statistical-based features from CGM traces. Then Fisher score and variables cluster analyses were applied in the features selection at the second level, aiming at further dimension reduction and obtaining better clustering effect. The three factors from features clustering were put into the third level. The third level, subject clustering, clustered all subjects into different subgroups. The detailed multilevel cluster method follows an iterative procedure, as shown in online supplemental materials section 1.

Features extraction

The 908 CGM traces were first processed to extract the set of well-established indices already considered in the recent research by Weiping *et al*.[17] As a result, mean sensor glucose (MSG), SD of the sensor glucose (SDSG), coefficient of variation (CV), mean amplitude of glycemic excursions (MAGE), and the largest amplitude of glycemic excursions (LAGE) were calculated into the statistical-based features pool. The knowledge-based features pool had the following features: mean postprandial sensor glucose (MPSG), which means the average sensor value of postprandial glucose 3 hours after three meals; J index and M value; indices based on the permanence in the target glucose range, that is, percentages of values within the target range (TIR; 3.9–10 mmol/L) and percentages of values out of the target range (TOR;

Table 1 Clinical characteristics of participants with type 2 diabetes

Features	Values
Number of participants	908
Gender (male/female)	565/343
Age (years)	61 (53–67)
Diabetes duration (years)	12 (6–17)
BMI (kg/m^2)	24.89±3.62
MSG (mmol/L)	8.99 (7.79–10.45)
MPSG (mmol/L)	10.05 (8.36–11.83)
TOR (%)	0.34 (0.15–0.55)
TIR (%)	0.66 (0.45–0.85)
SDSG (mmol/L)	2.27 (1.60–3.04)
CV (%)	24.79 (18.80–31.92)
LAGE (mmol/L)	9.20 (6.70–11.90)
M value	36.97 (32.41–42.23)
MAGE (mmol/L)	5.58 (3.99–7.60)
J index	41.56 (30.06–58.45)
HBGI	6.38 (3.26–11.10)
LI	2.50 (1.48–3.79)
GRADE	373.09±87.06
ADRR	218.11 (193.01–242.54)

Normally distributed variables are presented as mean±SD, and non-normally distributed data are expressed as median with IQR. ADRR, average daily risk range; BMI, body mass index; CV, coefficient of variation; GRADE, glycemic risk assessment diabetes equation; HBGI, high blood glucose index; LAGE, largest amplitude of glycemic excursions; LI, liability index; MAGE, mean amplitude of glycemic excursions; MPSG, mean postprandial sensor glucose; MSG, mean sensor glucose; SDSG, SD of the sensor glucose; TIR, percentage of values within the target range (3.9–10 mmol/L); TOR, percentage of values out of the target range (<3.9 mmol/L or >10 mmol/L).

<3.9 mmol/L or >10 mmol/L); and indices based on glucose risk, that is, low and high blood glucose indices (LBGI, HBGI) and glycemic risk assessment diabetes equation (GRADE) score. In addition, average daily risk range (ADRR) and liability index (LI) were calculated as the feature indices according to the definitions. The ideal blood glucose level in the M value function was set as 5.8 mmol/L.

Through this operation, we reduce the 288-dimensional data of the CGM to a lower dimension of 15. Before the CGM traces were fed into the multilevel clustering, Z-score normalization was used to solve the problem of amplitude scaling and eliminate the offset effect.

Features selection

In the second step, an effective features selection method based on Fisher score[18] and feature clustering were applied. These steps help to find the pattern similarities of CGM features from feature extraction and reducing features for clustering (online supplemental materials section 2). First, the CGM features that have more discrimination information based on Fisher score were selected. Then the agglomerative hierarchical clustering method was adopted for feature fusion. Irrelevant and redundant features were removed. The average of elements in every cluster was calculated and used as a representative value of each cluster in further analysis. Finally, a simpler feature matrix of CGM data was generated.

Subject clustering

In the third step, we applied K-means++ algorithm to the simple features' matrix from feature selection in order to obtain the subtypes of patients with type 2 diabetes.[19] As an improvement version of K-means, K-means++ is more suitable for large amounts of patient data and its results were more meaningful and easier to interpret.[20–22] Moreover, the stability and efficiency of the K-means++ method were better than the traditional K-means method by improving the choice of clustering center.[20 21]

The K-means++ solution with Euclidean distance[23] was designed to identify potential groups among type 2 diabetes based on the simple features' matrix. We choose the optimal cluster number on the basis of the elbow method.[24] The K-means++ clustering was performed using the cluster function (runs=100) in the scikit-learn package in PYTHON V.3.7. Cluster stability was assessed by resampling the data set 2000 times by the Jaccard

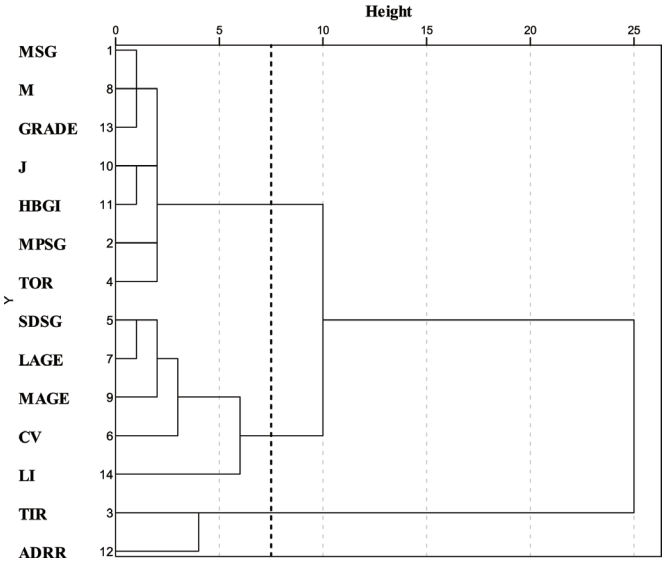


Figure 1 Dendrogram of hierarchical cluster analysis of continuous glucose monitoring variables. Height indicates the distance of correlation method between substructures. ADRR, average daily risk range; CV, coefficient of variation; GRADE, glycemic risk assessment diabetes equation score; HBGI, high blood glucose indices; J, J index; LAGE, largest amplitude of glycemic excursions; LI, liability index; M, M value; MAGE, mean amplitude of glycemic excursions; MPSG, mean postprandial sensor glucose; MSG, mean sensor glucose; SDSG, SD of the sensor glucose; TIR, percentages of values within the target range (3.9–10 mmol/L); TOR, percentages of values out of the target range (<3.9 mmol/L or >10 mmol/L).

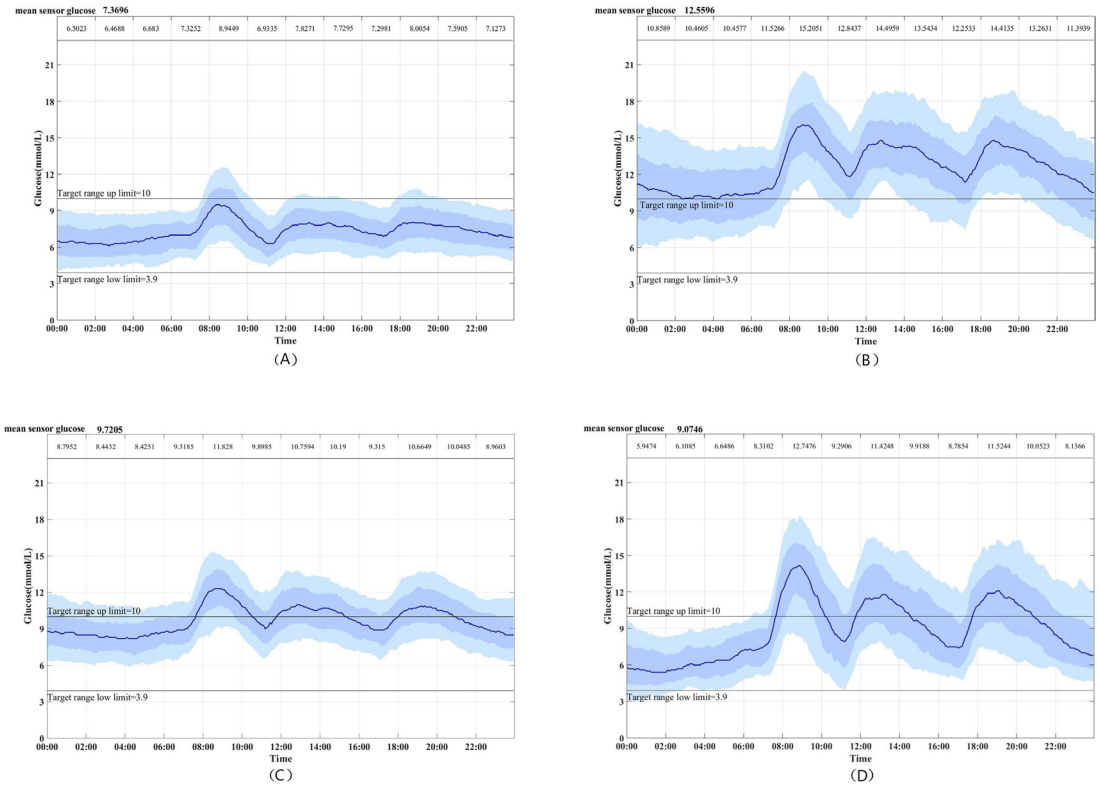


Figure 2 Continuous glucose monitoring curve of each subgroup: (A) cluster 1, (B) cluster 2, (C) cluster 3 and (D) cluster 4. The solid line is the median. The dark blue bar means the 25th–75th percentiles and the light blue bar means the 10th–90th percentiles. The mean sensor glucose is given at the top left corner.

bootstrap method. A stable cluster generally yields Jaccard similarity index of greater than 0.75.[25]

Statistical analysis

Data were analyzed with PYTHON V.3.7 and PASW Statistics V.25 (SPSS, Chicago, Illinois). To determine the clinical significance of the clustering results, this paper compared the four clusters with the CGM feature matrix and clinical phenotypes. Kruskal-Wallis test was used for non-normally distributed data to determine differences between the four subgroups. A two-tailed p value of <0.05 was considered statistically significant.

RESULTS

Participant characteristics

The clinical characteristics are shown in table 1. A total of 908 patients (565 men and 343 women) had a median age of 61 (53–67) years and BMI of 24.89±3.62. Two-sample Kolmogorov-Smirnov test indicated that there was no significant difference between sex and all clinical phenotype indices. Therefore, the next step in the subjects' cluster should not consider sex (online supplemental table S3).

Features cluster

The LBGI was excluded after counting Fisher score. Then the CGM variables were classified into three clusters (figure 1). The first cluster, which was composed of MSG, M value, GRADE, J index, MPSG, HBGI, and TOR, had a high correlation with MSG and HBGI. The second cluster, which was composed of SDSG, CV, LAGE, MAGE, and LI, had the highest correlation with CV. The third cluster was composed of ADRR and TIR. It had a negative correlation with the first and second clusters (online supplemental table S4). Therefore, this study defined the first cluster as the glycemic level factor, the second cluster as the glycemic fluctuation factor and the third cluster as the negative glucose factor.

Subjects cluster

By elbow method, the cluster number of subjects at four would be the most appropriate value (online supplemental materials section 3). Each subgroup has different characteristics on CGM traces throughout follow-up (figure 2). We compared the differences between the clustering factors and selected CGM parameters of the four subgroups (figure 3). Cluster 1, which included 338 (37.22%) patients, was characterized by the lowest glucose level factor and glucose fluctuation factor and highest negative glucose factor, and was labeled as low-level and low-fluctuation diabetes (LLLFD). Cluster 2, which included 174 (19.16%) patients and was labeled as high-level and high-fluctuation diabetes (HLHFD), was different from cluster 1; specifically, it had the highest glucose level factor, relatively high glucose fluctuation factor and lowest negative glucose factor. Cluster 3, labeled as the moderate-level and moderate-fluctuation diabetes (MLMFD), included 227 (25.00%) patients. It

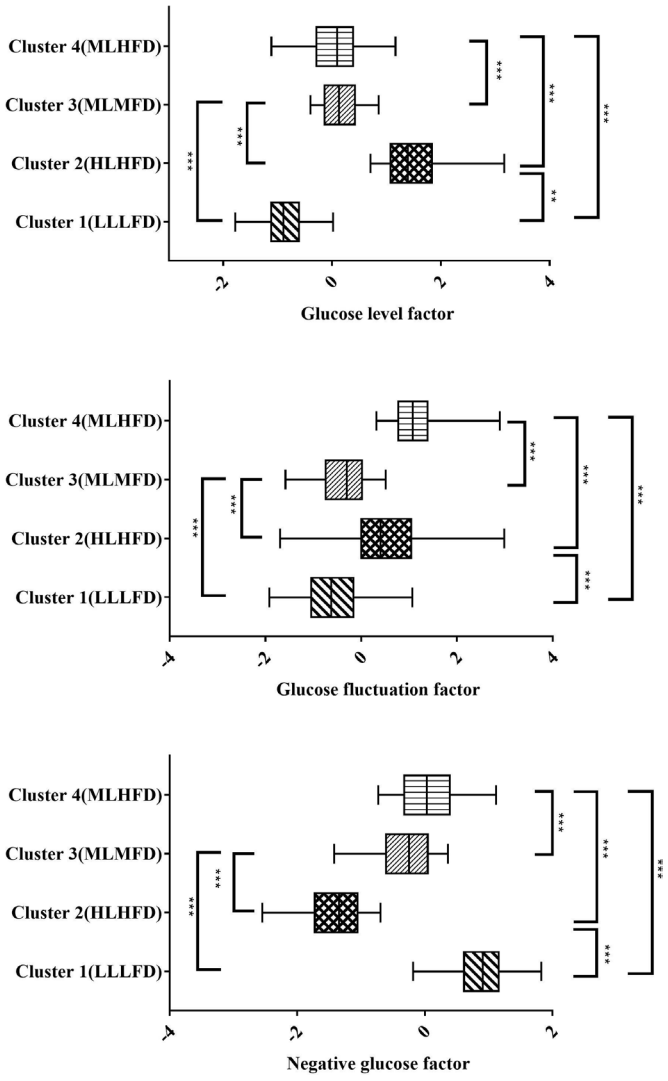


Figure 3 Box plot of cluster factors in patients. Adjusted p value was used. **Adjusted p values were under 0.01; ***adjusted p values were under 0.001. HLHFD, high-level and high-fluctuation diabetes; LLLFD, low-level and low-fluctuation diabetes; MLHFD, moderate-level and high-fluctuation diabetes; MLMFD, moderate-level and moderate-fluctuation diabetes.

was characterized by reasonably low glucose fluctuation factor and high glucose level factor, but otherwise was similar to cluster 1. Overall, 169 (18.61%) patients in cluster 4 (labeled as moderate-level and high-fluctuation diabetes, MLHFD) showed a higher glucose fluctuation factor than those in the other three groups.

To be effective clinically, the study compared the clusters regarding clinical phenotypes and antidiabetes agents, as shown in table 2. Cluster 1 (LLLFD) was characterized by the lowest HbA1c (7.2%, 6.5%–8.3%) and highest beta-cell capacity as assessed by C peptide and HOMA-2%β. In contrast, cluster 2 (HLHFD) had the highest HbA1c (10.2%, 9.1%–11.3%) and the most severely impaired beta-cell function. Cluster 3 (MLMFD) had a relatively low HbA1c (8.6%, 7.7%–10.0%) and preserved C peptide indices. Additionally, cluster 3 (MLMFD) and cluster 4 (MLHFD) were comparable (p=0.795) in terms of HbA1c. Cluster 4 showed relatively low levels of beta-cell function indices and a modest level of HbA1c (8.9, 7.5–10.1). However, individuals in cluster 4 (MLHFD) were most sensitive to insulin as reflected by the highest HOMA-2%Sensitivity (31.90, 20.50–53.55) and the lowest HOMA-2%Insulin resistance (3.10, 1.90–4.90). Concerning the treatment received, cluster 1 had the lowest percentage of insulin use (54.4%), while cluster 2 had the highest propensity (93.1%) to receive insulin. Moreover, cluster 4, which is the most insulin-sensitive, had the lowest use of metformin (30.2%) and α-glucosidase inhibitors.

DISCUSSION

This study used a data-driven clustering approach to extract implicit information from extensive high-dimensional CGM data. Distinct from previous efforts,[12 13] this is a method that could not only provide us with those knowledge-based and statistics-based CGM indices, but also offer features selection to improve the clustering results of subjects. It may provide new insights into the clustering of type 2 diabetes and more tailored diabetes management. The results demonstrate that this multi-level clustering approach could identify patients with type 2 diabetes with diverse clinical phenotypes, such as beta-cell function. Some medical check-ups like C peptide testing is not routinely measured, and glucose pattern represents the most obvious phenotypes of diabetes and may reveal the inherent nature of the disease. Furthermore, the method can be applied to CGM sensors for the convenience of patients with type 2 diabetes at home.

The robustness of the clustering was confirmed by the distinct characteristics of clinical phenotypes, particularly for insulin secretion/sensitivity indices, among the four subgroups. In short, cluster 1 (LLLFD) and cluster 2 (HLHFD) were related to the best and the worst beta-cell function, respectively. Intriguingly, while beta-cell function was modestly impaired, patients in cluster 4 (MLHFD), which was characterized by unstable glucose levels, had the highest level of insulin sensitivity, highlighting the role of insulin sensitivity in glucose homeostasis. Consistent with our study, a positive relationship between insulin sensitivity and glycemic variability was reported in 366 insulin-treated hospitalized patients with type 2 diabetes.[26] Indeed, from the clinical perspective, those who are more insulin-sensitive may be more prone to glycemic excursions after dose adjustment. Given the distinct clinical features among the four clusters, different strategies of glucose-lowering treatment may be considered. For instance, for cluster 2, with the most impaired beta-cell function, insulin therapy should be initiated at the early stage of the disease to achieve glycemic targets. For patients in cluster 4, antidiabetic agents that alleviate glucose fluctuations, such as glucagon-like peptide 1 receptor agonists, should be preferentially considered.

Table 2 Clinical characteristics of the various clusters of type 2 diabetes

Clinical phenotypes	Clusters				P value
	LLLFD (n=338)	HLHFD (n=174)	MLMFD (n=227)	MLHFD (n=169)	
Basic information					
Age (years)	62 (54–68)	61 (51–67)	60 (53–67)	62 (54–67)	0.166
BMI (kg/m^2)	24.23 (22.49–26.63)	24.94 (22.65–27.26)	24.80 (22.82–27.31)	24.22 (22.56–26.36)	0.150
Duration (years)	12 (6–17)	14 (7–20)	11 (6–18)	12 (5–17)	0.154
HOMA indices					
HOMA-2%β	172.45 (118.08–239.40)	73.20* (50.95–106.95)	115.70*† (82.70–155.50)	110.90*† (71.75–172.00)	<0.001
HOMA-2%Sensitivity	23.40 (16.60–32.83)	26.25 (16.58–38.93)	21.80 (14.90–33.60)	31.90*†‡ (20.50–53.55)	<0.001
HOMA-2%Insulin resistance	4.30 (3.00–6.00)	3.80 (2.58–6.03)	4.60 (3.00–6.70)	3.10*†‡ (1.90–4.90)	<0.001
Clinical measures					
HbA1c (%)	7.2 (6.4–8.3)	10.2* (9.1–11.3)	8.6*† (7.7–10.0)	8.9*† (7.5–10.1)	<0.001
FCP (ng/mL)	1.84 (1.30–2.53)	1.43* (0.97–2.14)	1.73† (1.21–2.62)	1.26*‡ (0.76–1.98)	<0.001
CP2h (ng/mL)	4.47 (2.73–6.56)	2.43* (1.73–3.88)	4.31† (2.35–6.17)	3.29*‡ (1.89–5.38)	<0.001
ΔCP2h (ng/mL)	2.65 (1.08–4.38)	1.35* (0.55–2.17)	2.21† (0.88–3.83)	1.89† (0.93–3.41)	<0.001
SBP (mm Hg)	130 (120–140)	130 (120–140)	130 (120–140)	130 (120–140)	0.620
DBP (mm Hg)	78 (70–82)	80 (70–84)	80 (70–84)	78 (70–82)	0.793
Triglyceride (mmol/L)	1.32 (0.92–1.88)	1.55* (1.12–2.80)	1.67* (1.13–2.50)	1.27†‡ (0.97–1.94)	<0.001
HDL cholesterol (mmol/L)	1.00 (0.84–1.23)	0.96 (0.82–1.15)	0.97 (0.80–1.17)	1.06†‡ (0.90–1.25)	0.008
LDL cholesterol (mmol/L)	2.51 (1.96–3.22)	2.65 (2.06–3.23)	2.55 (1.95–3.30)	2.75 (1.98–3.35)	0.286
Antidiabetic agents, n (%)					
Metformin	153 (45.3)	61 (35.1)	97 (42.7)	51*‡ (30.2)	0.004
Sulfonylurea	83 (24.6)	22* (12.6)	63† (27.8)	25*‡ (13.0)	<0.001
Thiazolidinediones	18 (5.3)	5 (2.9)	15 (6.6)	8 (4.7)	0.400
Glinides	20 (5.9)	5 (2.9)	14 (6.2)	11 (6.5)	0.400
DPP-4 inhibitors	32 (9.5)	9 (5.2)	30† (13.2)	11 (6.5)	0.025
AGI	132 (39.1)	48* (27.6)	86 (37.9)	29*‡ (17.2)	<0.001
Insulin	184 (54.4)	162* (93.1)	157*† (69.2)	143*†‡ (84.6)	<0.001

All data are expressed as median with IQR.
For clusters, p values are from Kruskal-Wallis test for continuous variables and from χ^2 for categorical variables.
Duration refers to duration of diabetes.
*Significant difference in unpaired (p<0.05), Dunn-Bonferroni test for post-hoc comparisons, compared with cluster 1.
†Significant difference in unpaired (p<0.05), Dunn-Bonferroni test for post-hoc comparisons, compared with cluster 2.
‡Significant difference in unpaired (p<0.05), Dunn-Bonferroni test for post-hoc comparisons, compared with cluster 3.
AGI, α-glucosidase inhibitor; BMI, body mass index; CP2h, C peptide levels at 2 hours; ΔCP2h, increments of C peptide in plasma levels 120 min; DBP, diastolic blood pressure; DPP-4, Dipeptidyl peptiduse 4; FCP, C peptide variables as fasting C peptide; HbA1c, glycated hemoglobin A1c; HDL, high-density lipoprotein; HLHFD, high-level and high-fluctuation diabetes; HOMA, Homeostasis model assessment; LDL, low-density lipoprotein; LLLFD, low-level and low-fluctuation diabetes; MLHFD, moderate-level and high-fluctuation diabetes; MLMFD, moderate-level and moderate-fluctuation diabetes; SBP, systolic blood pressure.

Another aspect of clinical significance is the presumed differential risk for chronic diabetic complications across the four clusters. Cluster 2 (HLHFD) and cluster 4 (MLHFD) are more associated with hyperglycemia because their glucose fluctuated wildly.[27] In addition, the high degree of glycemic fluctuation factor noted in these two clusters may confer additional risk for microvascular and macrovascular complications.[28] Cluster 3 (MLMFD), with relatively higher insulin resistance, may have a high risk of kidney disease due to the correlation between insulin resistance and glomerular hypertension and hyperfiltration.[29] It may help doctors to choose precise treatment more quickly according to cluster phenotypes.

The limitations of this study are worth noting. First, the abnormal samples in the database were not removed by our approach. This may reduce the stability of clustering results. In future studies, the method should be improved to exclude outlier subjects. Second, our study only included hospitalized patients with type 2 diabetes in China. The generalizability of the study results is uncertain and needs to be tested in other populations. Third, it is notable that some factors could affect the stability of glucose profile over time within individuals, such as diet, exercise, stress and treatment regimen. In addition, only 1-day CGM data were used in our study, while two previous studies suggested that 14 days of CGM may be needed for reliable estimation of overall glucose control.[30 31] Therefore, the results of CGM clustering should be interpreted with caution. Finally, whether the new method of clustering translates into improved glycemic control and subsequent diabetes-related outcomes remains unknown. Further studies are warranted to validate the clinical significance of this method.

In summary, a multilevel clustering method driven by CGM data was designed to divide patients with type 2 diabetes into subgroups with distinct clinical phenotypes. This method has implications for big data from CGM systems and for digital precision in diabetes research. Furthermore, the results of the study show the promise of developing personalized medicine and intelligent healthcare for type 2 diabetes.

Acknowledgements We would like to thank all the involved clinicians, nurses, and technicians in Shanghai Clinical Center for Diabetes for dedicating their time and skills to the completion of this study.

Contributors RT, XY and JZ designed the study. RT was responsible for data analysis and overall design of the algorithm. XY was responsible for design and analysis of the algorithm. JZ and JL were responsible for clinical phenotypes analysis. JL, YS, WL, WZ and YB assisted in data collection and offered suggestion for the study. HL and JZ funded the study. All authors read and approved the final manuscript.

Funding This work was funded by the National Key R&D Program of China (2018YFC2001004), the Shanghai Municipal Education Commission-Gaofeng Clinical Medicine Grant Support (20161430) and the National Natural Science Foundation of China (61903071, 61973067).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The study was approved by the Ethics Committee of Shanghai Jiao Tong University Affiliated Sixth People's Hospital (2020-KY-024) and complied with the principles of the Helsinki Declaration.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iDs
Yuqian Bao http://orcid.org/0000-0002-4754-3470
Jian Zhou http://orcid.org/0000-0002-1534-2279

REFERENCES

1 American Diabetes Association. 2. Classification and Diagnosis of Diabetes: *Standards of Medical Care in Diabetes-2019*. *Diabetes Care* 2019;42:S13–28.
2 Davies MJ, D'Alessio DA, Fradkin J, *et al*. Management of hyperglycaemia in type 2 diabetes, 2018. A consensus report by the American diabetes association (ADA) and the European association for the study of diabetes (EASD). *Diabetologia* 2018;61:2461–98.
3 Ahlqvist E, Storm P, Käräjämäki 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;6:361–9.
4 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:e001506.
5 Ilmarinen P, Tuomisto LE, Niemelä O, *et al*. Cluster analysis on longitudinal data of patients with adult-onset asthma. *J Allergy Clin Immunol Pract* 2017;5:967–78.
6 Hyun S, Kaewprag P, Cooper C, *et al*. Exploration of critical care data by using unsupervised machine learning. *Comput Methods Programs Biomed* 2020;194:105507.
7 Zou X, Zhou X, Zhu Z, *et al*. Novel subgroups of patients with adult-onset diabetes in Chinese and US populations. *Lancet Diabetes Endocrinol* 2019;7:9–11.
8 Udler MS, Kim J, von Grotthuss M, *et al*. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: a soft clustering analysis. *PLoS Med* 2018;15:e1002654.
9 Battelino T, Conget I, Olsen B, *et al*. The use and efficacy of continuous glucose monitoring in type 1 diabetes treated with insulin pump therapy: a randomised controlled trial. *Diabetologia* 2012;55:3155–62.
10 Bolinder J, Antuna R, Geelhoed-Duijvestijn P, *et al*. Novel glucose-sensing technology and hypoglycaemia in type 1 diabetes: a multicentre, non-masked, randomised controlled trial. *Lancet* 2016;388:2254–63.
11 Haak T, Hanaire H, Ajjan R, *et al*. Use of Flash Glucose-Sensing Technology for 12 months as a Replacement for Blood Glucose Monitoring in Insulin-treated Type 2 Diabetes. *Diabetes Ther* 2017;8:573–86.
12 Hall H, Perelman D, Breschi A, *et al*. Glucotypes reveal new patterns of glucose dysregulation. *PLoS Biol* 2018;16:e2005143.
13 Kahkoska AR, Adair LA, Aiello AE, *et al*. Identification of clinically relevant dysglycemia phenotypes based on continuous glucose monitoring data from youth with type 1 diabetes and elevated hemoglobin A1c. *Pediatr Diabetes* 2019;20:556–66.
14 American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* 2014;37 Suppl 1:S81–90.

15 Shen Y, Si Y, Lu J, *et al*. Association between 1,5-Anhydroglucitol and acute C peptide response to arginine among patients with type 2 diabetes. *J Diabetes Res* 2020;2020:1–7.
16 Levy JC, Matthews DR, Hermans MP. Correct homeostasis model assessment (HOMA) evaluation uses the computer program. *Diabetes Care* 1998;21:2191–2.
17 Weiping J, Jian Z, Yuqian B. *Continuous glucose monitoring*. Shanghai science and Technology Press, 2019: 55–71.
18 Gu Q, Li Z, Han J. Generalized Fisher score for feature selection. *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, Barcelona, Spain, 2011:266–73.
19 Arthur D. k-means++: the advantages of careful seeding. In: proc. Of the 18th ACM-SIAM on Discrete Algorithms 2007:1027–35.
20 Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min Knowl Discov* 1998;2:283–304.
21 Kapoor A, Singhal A. A comparative study of k-means, K-Means++ and fuzzy C-Means clustering algorithms. *IEEE 3rd international conference on computational intelligence & communication technology (CICT)*, 2017:1–6.
22 Liao M, Li Y, Kianifard F, *et al*. Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. *BMC Nephrol* 2016;17:25.
23 Elmore KL, Richman MB. Euclidean distance as a similarity metric for principal component analysis. *Mon Weather Rev* 2001;129:540–9.
24 Bholowalia P, Kumar A. EBK-means: a clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications* 2014;105:17–24.
25 Hennig C. Cluster-wise assessment of cluster stability. *Comput Stat Data Anal* 2007;52:258–71.
26 Huang Y, Heng C, Wei J, *et al*. Influencing factors of glycemic variability in hospitalized type 2 diabetes patients with insulin therapy: a Strobe-compliant article. *Medicine* 2017;96:e8021.
27 DiMeglio LA, Acerini CL, Codner E, *et al*. ISPAD clinical practice consensus guidelines 2018: glycemic control targets and glucose monitoring for children, adolescents, and young adults with diabetes. *Pediatr Diabetes* 2018;19 Suppl 27:105–14.
28 Monnier L, Colette C, Owens DR. Glycemic variability: the third component of the dysglycemia in diabetes. is it important? how to measure it? *J Diabetes Sci Technol* 2008;2:1094–100.
29 Spoto B, Pisano A, Zoccali C. Insulin resistance in chronic kidney disease: a systematic review. *Am J Physiol Renal Physiol* 2016;311:F1087–108.
30 Riddlesworth TD, Beck RW, Gal RL, *et al*. Optimal sampling duration for continuous glucose monitoring to determine long-term glycemic control. *Diabetes Technol Ther* 2018;20:314–6.
31 Xing D, Kollman C, Beck RW, *et al*. Optimal sampling intervals to assess long-term glycemic control using continuous glucose monitoring. *Diabetes Technol Ther* 2011;13:351–8.

A Multilevel Clustering Approach Driven by Continuous Glucose Monitoring Data for Further Classification of Type 2 Diabetes

Rui Tao[#1], Xia Yu[#1], Jingyi Lu[#2], Yun Shen[2], Wei Lu[2], Wei Zhu[2], Yuqian Bao[2],

Hongru Li[1], Jian Zhou[*2]

[1] College of Information Sciences and Engineering, Northeastern University, Shenyang, 110819, China

[2] Department of Endocrinology and Metabolism, Shanghai Clinical Center for Diabetes, Shanghai Diabetes Institute, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai 200233, China

Keywords: Type 2 Diabetes, multilevel cluster analysis, knowledge mining, CGM data, clinical phenotypes

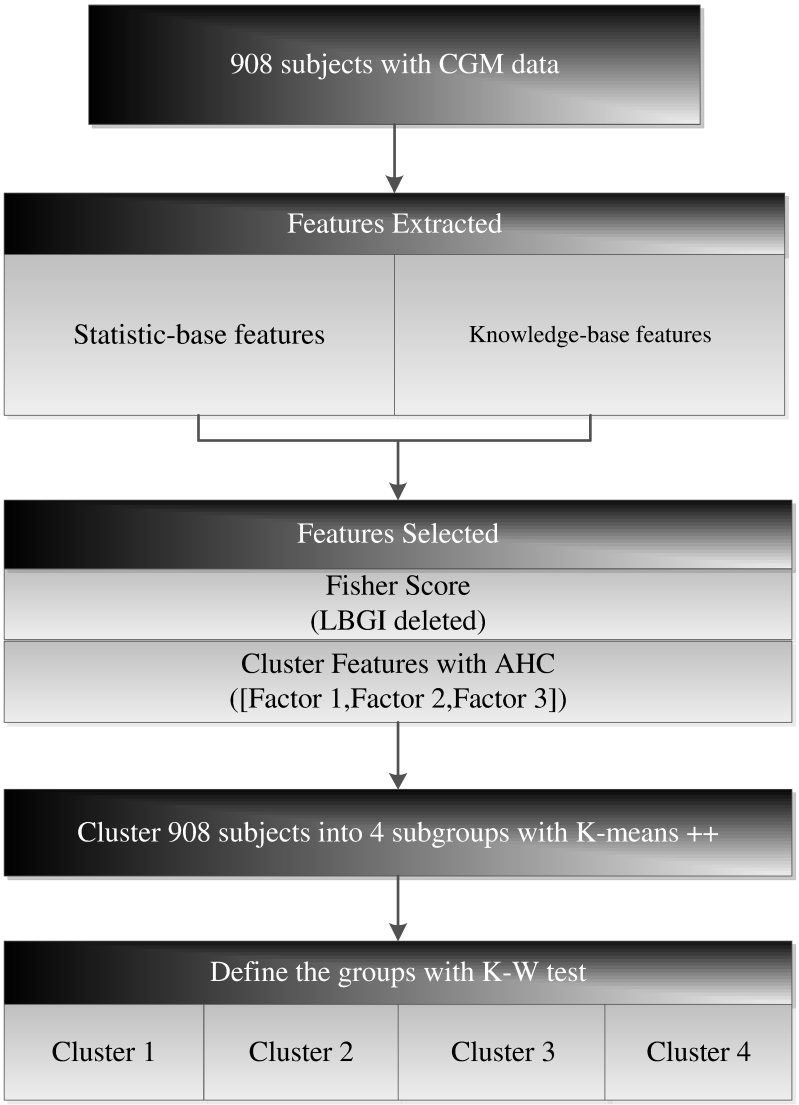
# These three authors contributed equally to the paper.

* Corresponding author.

E-mail address: zhoujian@sjtu.edu.cn (Jian Zhou)

1. The overview of multilevel clustering of CGM data:

The multilevel clustering method as a data-mining model contains four steps. The first step extract features from CGM data. Statistic-base features like mean values (mean sensor glucose, MSG), standard deviation of the values (SD of the sensor glucose, SDBG) and coefficient of variation (CV) were calculated. In addition, some knowledge-base features which be used in clinical also extracted. Globally, a 15-dimensional group of feature variables extracted from time-series CGM data was used to characterize the patient for future analysis. Then we designed a step for selecting features in the feature group. In this step, an effective feature selection method based on Fisher score and agglomerative hierarchical clustering was applied to find the pattern similarities of CGM features and reducing features. Through this step, the original 15-dimensional features could be further reduced to 3-dimensional features with clinical significance. In the third step, we used the 3-dimensional features in the second step to characterize the participants. And subgroups of the participants were found by K-means++ method. At the last step, four subgroups were identified and their clinical phenotypes were compared. The overall flow chart is shown in **Supplementary Figure S1**. Furthermore, the details of the method are shown in **Supplementary Table S1**.



Supplementary Figures S1: The flow chart of the cluster method.
Abbreviations: LBGI-low blood glucose indices, K-W test-Kruskal Wails test, AHC-agglomerative hierarchical clustering

Supplementary Table S1: The iterative procedure of Multilevel cluster for CGM data

Algorithm Multilevel cluster for CGM data (MLC-CGM)
Input: CGM datasets ***Output:*** the groups' labels of each subject ***Step 1: Feature Extraction*** ***Step 1.1:*** Calculate ***knowledge-based features*** ***Step 1.2:*** Calculate ***statistical-base features*** ***Step 1.3:*** Initialize the set of features ***S*** ***Step 2: Feature Selection*** ***Step 2.1:*** Build a simple clustering using ***S*** as input ***Step 2.2:*** For every feature f_i in ***S***: ***Step 2.2.1:*** Compute the Fisher Score (***FS***). ***Step 2.2.2:*** Find $f^* = f_i$ if ***FS>threshold*** such that adding it to ***S set=S*** ***Step 2.3:*** Update the ***S* set*** ***Step 2.4:*** Cluster features in ***S* set*** ***Step 3: Subjects Cluster*** ***Step 3.1:*** Choose the optimal clustering number ***n*** by elbow method ***Step 3.2:*** Cluster patients into n subgroups among ***S* set*** ***Step 4:*** Definite the subgroups using ***Kruskal-Wallis test***

2. Fisher score

As a state-of-the-art feature selection, the main idea of the Fisher score is that the features with strong discrimination performance keep the distance within clusters as small as possible and keep the distance between clusters as large as possible. Compared to Mutual information (MI),[1] Relieff,[2] Laplacian Score[3] and Trace ratio criterion,[4] Fisher Score (FS) is one of the most widely used criteria for filter-based feature selection [5]. Here, the Fisher score of a single feature, used as the criterion for selecting the CGM indices can be expressed by the following formulas:

$$J_{\text{fisher}}(k) = S_B(k) / S_w(k)$$
$$S_B(k) = \sum_{i=1}^{c} \frac{n_i}{n} \left(m_i(k) - m(k) \right)^2 \quad (1)$$
$$S_w(k) = \sum_{i=1}^{c} \sum_{x \in w_i} \left(x(k) - m_i(k) \right)^2$$

where there are n samples in the data set that belong to C clusters, each cluster contains n_i samples respectively. $x(k)$, $m_i(k)$, $m(k)$ means the sample x, mean values of the i cluster, and all samples' mean values on the k features. The Fisher score under the threshold will be removed [6]. Here, we set the threshold value to 0.5. And the Fisher score of the CGM features are shown in **Supplementary Tables S2.**

Supplementary Table S2. The fisher score results in the feature selection

Feature	FS values	Features	FS values	Features	FS values
MSG	3.23	CV	0.92	LBGI	0.22
MPSG	2.53	LAGE	1.44	HBGI	3.75
TIR	3.44	M value	2.98	ADRR	2.08
TOR	3.44	MAGE	1.00	GRADE	2.37
SDSG	1.46	J-index	3.56	LI	0.57

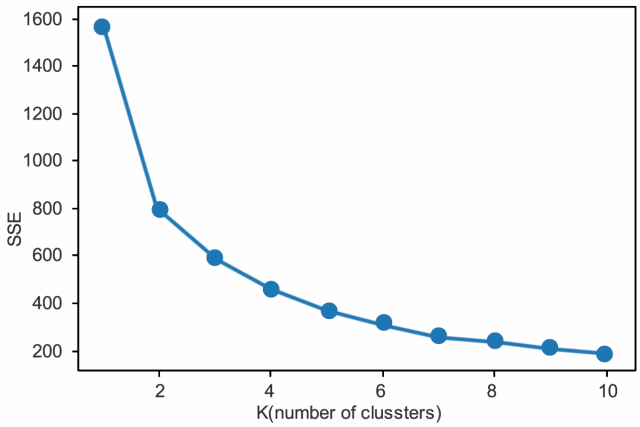
Abbreviations: MSG-mean sensor glucose; MPSG-mean postprandial sensor glucose; TIR-percentages of values within the target range ([3.9-10] mmol/L); TOR-percentages of values out of the target range (<3.9 or >10 mmol/L); SDSG-standard deviation of the sensor glucose, CV-coefficient of variation; LAGE-largest amplitude of glycemic excursions; MAGE-mean amplitude of glycemic excursions; HBGI,LBGI-high or low blood glucose indices; GRADE-glycemic risk assessment diabetes equation score; LI-liability index; ADRR-average daily risk range.

3. Elbow method

To properly choose the optimal number of the K-means++ on the basis of the elbow method [7 8], we divided the data into categories 1 to 11. Then the error sum of the square distance (SSE) between the particle of each cluster and the sample point in the cluster were calculated as:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - u_i\|^2$$
$$u_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \quad (2)$$

where u_i is the center of cluster i, x is the sample point, and k is the value of K-means++. Distortions of the patient's matrix at 4 will be the most appropriate values as shown in **Supplementary Figures S2.**



Supplementary Figure S2.the SSE line chart of the elbow method. The curve reached the break point at k=4. As k increases, SSE decreases slowly.

Supplementary Table S3. Two sample Kolmogorov-Smirnov test results

SEX	AGE	BMI	Duration	SBP	DBP
	0.000	0.001	0.109	0.067	0.158
SEX	HbA1c	HDL	LDL	FCP	CP2h
	0.003	0.000	0.025	0.145	0.175

A two-tailed P value of <0.010 was considered statistically significant. As shown in table, most adjusted P values above 0.010. These results verified the similar distributions at characteristics in men and women. So, sex could not be considered in the analysis.

Abbreviations: FCP -C peptide variables as fasting C peptide; CP2h-C peptide levels at 2 hours; ΔCP120min- the increments of C peptide in plasma levels 90 minutes; SBP-systolic blood pressure; DBP-diastolic blood pressure; BMI -body mass index; Duration-duration of diabetes; HDL-high density lipoprotein; LDL-low density lipoprotein.

Supplementary Table S4. Glycemic level factor (GLF), the glycemic fluctuation factor (GFF) and negative glycemic factor (NGF) correlation analysis

	MSG	HBGI	CV	TIR
GLF	0.988	0.993	0.120	-0.925
GFF	0.365	0.530	0.899	-0.553
NGF	-0.986	-0.980	-0.047*	0.935

A two-tailed P value of <0.010 was considered statistically significant. And the number with a * means its P value above 0.010. As shown in the table (Table S2), GLF had a higher correlation with MSG and HBGI, and GFF had the highest correlation with CV. NGF had the negative correlation with the indices unlike Factor 1 and Factor 2.

Abbreviations: MSG-mean sensor glucose; HBGI- high blood glucose indices; CV-coefficient variation; TIR-time in range **[3.9-10] mmol/L**

References:

1. Koller D, Sahami M. Toward optimal feature selection: Stanford InfoLab, 1996.
2. Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. *J Machine learning* 2003;53(1-2):23-69.
3. X He, D Cai, P Niyogi P. Laplacian score for feature selection. Advances in neural information processing systems; 2006.
4. Nie F, Xiang S, Jia Y. et.al.: Trace ratio criterion for feature selection. AAAI 2008; 2:671-676.
5. Tang JA, Salem, Liu H. Feature selection for classification: A review. *Data classification: Algorithms applications* 2014:37.
6. Zhou M. Hybrid feature selection method based on fisher score and genetic algorithm. *J Journal of Mathematical Sciences: Advances Applications* 2016;37(37):51-78.
7. Pham DT, Dimov SS, Nguyen CD. Selection of K in K-means clustering. *Journal of Mechanical Engineering Science* 2005;219(1):103-19.
8. Bholowalia PK, Arvind EBK-means: A clustering technique based on elbow method and k-means in WSN. *J International Journal of Computer Applications* 2014;105(9)