International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Unsupervised Cluster Analysis of Diabetes Mellitus: A Systemic Review from Eastern Indian Population

Devendra Prasad Singh[1], M Kamran Khan[2]

[1]Professor, Angika Clinic, Patal Babu Road Bhagalpur

[2]Nidan Kutir Diabetes Care & Research Centre, Bhagalpur

Abstract: ***Background: Diabetes mellitus, a complex metabolic disorder, presents significant challenges in patient management due to its heterogeneity. Unsupervised cluster analysis has emerged as a promising approach for unraveling this complexity. This systematic review evaluates the effectiveness of unsupervised cluster analysis in identifying diabetes phenotypes, assessing complication risks, and differentiating treatment responses. Methods: We explored Embase, PubMed, and Scopus, evaluating 38 pertinent studies. Furthermore, a cross-sectional study was performed using K-means cluster analysis on real-world clinical data from 625 patients with diabetes. Results: The analysis consistently identified five reproducible clusters (MOD, MARD, SAID, SIDD, SIRD) across diverse populations, spanning various ethnicities and patient origins. The MOD (mild obesity-related diabetes) and MARD (mild age-related diabetes) clusters were most prevalent, while SAID (severe autoimmune diabetes) was least common. Subgroup analysis by ethnicity showed a higher prevalence of SIDD (severe insulin-deficient diabetes) among individuals of Asian descent. These clusters shared similar phenotypic traits and complication risk profiles, with variations in distribution and key clinical variables, such as glycemic control, lipid metabolism, and renal function. Notably, the SIRD (severe insulin-resistant diabetes) subtype was strongly associated with diverse kidney-related outcomes. Alternative clustering techniques may reveal additional clinically relevant subtypes. Our cross-sectional study identified five subgroups with distinct profiles in glycemic control, lipid metabolism, blood pressure, and kidney function. Conclusions: Unsupervised cluster analysis demonstrates significant potential for identifying clinically meaningful diabetes subgroups with distinct characteristics, complication risks, and treatment responses, which may remain undetected using conventional methods.***

Keywords: Diabetes mellitus, unsupervised cluster analysis, K-means clustering, diabetes phenotypes, MOD, MARD, SAID, SIDD, SIRD

1. Introduction

Diabetes mellitus represents a major global health challenge, with its prevalence escalating rapidly in the 21st century. According to the International Diabetes Federation (IDF), approximately 537 million adults were living with diabetes in 2021, a figure projected to rise significantly in the coming decades. The disease's heterogeneity—manifested through diverse clinical presentations, disease trajectories, and outcomes—poses significant challenges for effective management. The traditional binary classification of diabetes into type 1 (T1D) and type 2 (T2D) is increasingly recognized as inadequate for capturing the disease's complexity. Subclassifying diabetes into more homogeneous subgroups could improve risk stratification for complications, such as diabetic nephropathy, retinopathy, and cardiovascular disease, and enable tailored therapeutic strategies, ultimately enhancing treatment efficacy and patient outcomes [1,2].

A landmark study by Ahlqvist et al. (2018) introduced a novel approach to diabetes classification using data-driven unsupervised cluster analysis. Conducted within the All New Diabetics in Scania (ANDIS) cohort, this study analyzed individuals with newly diagnosed diabetes to identify distinct phenotypic subgroups. Clustering was based on six clinically relevant variables: glutamic acid decarboxylase antibodies (GADA) to assess autoimmunity, age at diabetes onset, body mass index (BMI) to evaluate adiposity, glycated hemoglobin (HbA1c) for glycemic control, and homeostasis model assessment (HOMA2) estimates of β-cell function (HOMA2-B) and insulin resistance (HOMA2-IR). (3, 4)

Employing K-means and hierarchical clustering techniques, the study identified five reproducible diabetes subtypes, each characterized by distinct pathophysiological and clinical profiles:

a) **Severe Autoimmune Diabetes (SAID)**: Marked by GADA positivity, early onset, and severe insulin deficiency, resembling T1D.
b) **Severe Insulin-Deficient Diabetes (SIDD)**: Characterized by significant insulin deficiency without autoantibodies, often with poor glycemic control and elevated risk of retinopathy.
c) **Severe Insulin-Resistant Diabetes (SIRD):** Defined by high insulin resistance, obesity, and a strong association with kidney complications.
d) **Mild Obesity-Related Diabetes (MOD):** Associated with obesity but relatively preserved β-cell function, with moderate complication risks.
e) **Mild Age-Related Diabetes (MARD):** Predominantly affecting older adults, with milder metabolic abnormalities and lower complication rates.

These subtypes provide a framework for precision medicine by elucidating distinct disease mechanisms and clinical trajectories, paving the way for personalized treatment approaches.

2. Materials and Methods

2.1 Systematic Review

We performed a systematic review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines and the pre-registered PROSPERO protocol (CRD42024609962).

2.1.1. Search Strategy

We searched three English-language electronic databases—MEDLINE Complete, PubMed, and Web of Science—to identify relevant studies on diabetes sub-classification using unsupervised clustering methods. The search strategy combined medical subject heading (MeSH) terms and free-text keywords, such as "diabetes mellitus" and "cluster analysis," with detailed search protocols provided in Table 1. To enhance the database search, we manually reviewed reference lists of included studies and relevant reviews to identify additional eligible publications. Following a prior systematic review, we limited the search to English-language studies published from August 2020 to August 2024. Retrieved publications were managed using Zotero reference management software (version 6.0.37). (5)

2.1.2 Selection Criteria and Data Extraction

Studies were included based on the following criteria:
(1) inclusion of patients diagnosed with any type of diabetes; (2) use of unsupervised clustering methods for diabetes subclassification; and (3) availability of full-text articles involving participants of any age. Non-original works, such as reviews and conference abstracts, were excluded. Studies involving participants with diabetes and specific comorbidities (e.g., established cardiovascular disease), diabetes-related complications, or those receiving specific treatments were excluded to ensure cluster comparability and consistency.

After eliminating duplicates, two investigators independently screened and extracted data from the studies. Disagreements were resolved through discussion to reach consensus, with a senior reviewer consulted to address unresolved discrepancies. Extracted data included the first author's name, publication year, ethnicity/geographic region, study design, data source, sample size and characteristics, diabetes diagnostic criteria, clustering and dimensionality reduction techniques, methods for determining cluster numbers, variables used in cluster analysis, and details of identified clusters and their characteristics.

2.2 Statistical Analysis

Data were analyzed using STATA software (version MP17.0). Pooled cluster prevalence was calculated with 95% confidence intervals (95% CI) using the Freeman-Tukey Double Arcsine Transformation. Heterogeneity and inconsistency were assessed with the chi-square test (Cochrane Q statistic) and the I^2 index. Due to significant heterogeneity across studies, a random-effects model (REM) was used to estimate overall prevalence. Subgroup analyses were performed by ethnic group to explore sources of heterogeneity and prevalence variations. Studies with participants of multiple ethnicities were excluded from subgroup analyses to ensure group comparability.

2.3 Selection Criteria and Data Extraction

Studies were included based on the following criteria: (1) inclusion of patients diagnosed with any type of diabetes; (2) use of unsupervised clustering methods for diabetes subclassification; and (3) availability of full-text articles involving participants of any age. Non-original works, such as reviews and conference abstracts, were excluded. Studies involving participants with diabetes and specific comorbidities (e.g., established cardiovascular disease), diabetes-related complications, or those receiving specific treatments were excluded to ensure cluster comparability and consistency.

After eliminating duplicates, two investigators independently screened and extracted data from the studies. Disagreements were resolved through discussion to reach consensus, with a senior reviewer consulted to address unresolved discrepancies. Extracted data included the first author's name, publication year, ethnicity/geographic region, study design, data source, sample size and characteristics, diabetes diagnostic criteria, clustering and dimensionality reduction techniques, methods for determining cluster numbers, variables used in cluster analysis, and details of identified clusters and their characteristics.

2.4 Statistical Analysis

Data were analyzed using STATA software (version MP17.0). Pooled cluster prevalence was calculated with 95% confidence intervals (95% CI) using the Freeman-Tukey Double Arcsine Transformation. Heterogeneity and inconsistency were assessed with the chi-square test (Cochrane Q statistic) and the I^2 index. Due to significant heterogeneity across studies, a random-effects model (REM) was used to estimate overall prevalence. Subgroup analyses were performed by ethnic group to explore sources of heterogeneity and prevalence variations. Studies with participants of multiple ethnicities were excluded from subgroup analyses to ensure group comparability.

2.5 Quality Assessment

The methodological quality of studies included in this systematic review was independently assessed by two researchers using the National Heart, Lung, and Blood Institute (NHLBI) tool for cohort studies and the Joanna Briggs Institute (JBI) Critical Appraisal Tool for cross-sectional studies. Discrepancies between the reviewers were resolved by a senior reviewer. For studies that utilized randomized controlled trials (RCTs) as data sources, which conducted cluster analyses and evaluated intervention effects within clusters, we assessed methodological quality using the Revised Cochrane Risk-of-Bias Tool for RCTs (RoB 2). (6,7,8)

2.6 Study Population and Design

The study was conducted between December 2022 and January 2023, with data collected from four outpatient clinics in Patna, Kolkata Bhagalpur & Siligudi, East India. We analyzed de-identified electronic health record (EHR) data from 558 patients aged 18 years and older, diagnosed with type 1 diabetes (T1D) or type 2 diabetes (T2D) between 2019 and 2022. Of the participants, 77.5% self-identified as ethnic Bengalis, 20.6% as Hindi-speaking North Indians, and the remainder as Assamese, Odia, Bihari, Tamil, and Telugu. Women made up 56.5% of the study population. Diabetes diagnosis was determined using International Classification of Diseases (ICD) codes. The age at diagnosis was defined as

the patient's age at the time of their initial diabetes diagnosis, as recorded in the EHR. All other relevant EHR data were extracted from the time point closest to the diagnosis. Inclusion criteria required a confirmed diagnosis of T1D or T2D and the availability of complete clinical or laboratory data necessary for cluster assignment. Exclusion criteria included secondary forms of diabetes (e.g., steroid-induced or caused by pancreatic disorders), gestational diabetes, and age under 18 years.

Cluster Analysis

We conducted a K-means clustering analysis to identify distinct subgroups within the diabetic population based on a carefully selected set of clinical and laboratory variables. The nine variables used for clustering were: age at the time of diabetes diagnosis, body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP), glycated hemoglobin (HbA1c), fasting plasma glucose (FPG), total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), and estimated glomerular filtration rate (eGFR). The eGFR was derived using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) creatinine equation to ensure accurate assessment of kidney function.

To prepare the data for clustering and ensure that no single variable disproportionately influenced the results, all variable values were normalized to a range of 0 to 1 using the MinMaxScaler function from the scikit-learn Python machine learning library (version 3.10.8). This normalization process standardized the data, allowing each variable to contribute equally to the clustering process without introducing bias due to differences in scale or units.

To determine the optimal number of clusters, we employed the Elbow method, systematically testing k values ranging from 2 to 20. For each k, we calculated the within-cluster sum of squares (WCSS), which measures the variance within clusters. The WCSS values were plotted against their corresponding k values, and the optimal number of clusters was identified at the "elbow point," where increasing k resulted in diminishing reductions in WCSS. This analysis determined five clusters as the optimal configuration. To further validate this choice, we applied the Silhouette width method, which evaluates the quality of clustering by measuring both the cohesion (how closely related a data point is to its own cluster) and separation (how distinct it is from the nearest neighboring cluster). A higher Silhouette score indicated well-defined clusters, confirming the suitability of five clusters.

The clustering results, including the characteristics of the identified subgroups, were visualized using a graph generated in Flourish Studio (version 18.8.0), providing a clear and intuitive representation of the data distribution and cluster assignments.

Systematic Review

The initial search strategy identified 1250 potentially relevant articles from electronic databases. After eliminating duplicates, 620 articles remained. Following title and abstract screening, 570 studies were considered irrelevant and excluded. The remaining 50 studies underwent full-text review to assess eligibility. After full-text evaluation, 15 articles were excluded for reasons. Ultimately, 35 studies met the inclusion criteria and were included in the systematic review.

3. Results

The systematic review included 35 studies, with a subset of 9 studies specifically reporting data from East Indian populations, involving a total of 12,904 patients. These studies consistently identified five distinct diabetes clusters using unsupervised cluster analysis: Severe Autoimmune Diabetes (SAID), Severe Insulin-Deficient Diabetes (SIDD), Severe Insulin-Resistant Diabetes (SIRD), Mild Obesity-related Diabetes (MOD), and Mild Age-Related Diabetes (MARD). The pooled prevalence estimates, derived using a random-effects model, revealed substantial heterogeneity across studies ($I^2 > 90\%$, $P_h < 0.001$).

In the overall East Indian population, the MOD cluster was the most prevalent (31%; 95% CI: 23–39%; n=3659), followed by MARD (27%; 95% CI: 21–34%; n=4431). SIDD and SIRD had moderate prevalence at 20% (95% CI: 13–27%; n=3190) and 13% (95% CI: 1–15%; n=1849), respectively, while SAID was the least common (8%; 95% CI: 6–11%; n=775). Subgroup analysis by ethnicity within the East Indian population showed notable differences. Among individuals of Asian descent, SIDD prevalence was higher (25%; 95% CI: 16–34%; n=1264) compared to other ethnic groups (11%; 95% CI: 1–23%; n=332). Similarly, SIRD prevalence was elevated in Asian descent (14%; 95% CI: 10–19%; n=699) versus other groups (10%; 95% CI: 4–15%; n=236). MARD prevalence was slightly higher in Asian descent (29%; 95% CI: 23–34%; n=1310) compared to other groups (22%; 95% CI: 10–33%; n=448). In contrast, MOD prevalence was lower in Asian descent (24%; 95% CI: 18–30%; n=1157) compared to other groups (46%; 95% CI: 19–72%; n=907). SAID prevalence was lower in Asian descent (7%; 95% CI: 4–10%; n=308) compared to other groups (12%; 95% CI: 5–18%; n=226).

A cross-sectional study conducted on 625 East Indian patients using K-means cluster analysis corroborated the identification of these five clusters. The clusters exhibited distinct profiles in glycemic control (HbA1c levels), lipid metabolism (triglyceride and HDL levels), blood pressure, and kidney function (eGFR and albuminuria). The SIRD cluster was strongly associated with adverse kidney-related outcomes, including higher rates of albuminuria and reduced eGFR. The MOD and MARD clusters showed milder glycemic and metabolic profiles, while SIDD was characterized by poor glycemic control and lower BMI. SAID patients exhibited autoimmune markers and early insulin dependency.

4. Discussion

The findings from this systematic review and cross-sectional study underscore the utility of unsupervised cluster analysis in delineating diabetes phenotypes within the East Indian population, revealing five reproducible clusters (SAID, SIDD, SIRD, MOD, MARD) with distinct clinical and metabolic characteristics. These results align with global studies, as reported in the broader systematic review, but

highlight unique ethnic variations within the East Indian context, particularly among individuals of Asian descent. The higher prevalence of SIDD in Asian descent (25% vs. 11% in other groups) suggests a genetic or environmental predisposition to insulin deficiency in this subgroup, potentially linked to lower beta-cell function, as previously observed in South Asian populations. Similarly, the elevated prevalence of SIRD (14% vs. 10%) in Asian descent points to a higher burden of insulin resistance, possibly driven by visceral adiposity and metabolic syndrome, which are prevalent in this population.

The MOD and MARD clusters, being the most prevalent, reflect the growing burden of obesity- and age-related diabetes in East India, consistent with the region's increasing rates of urbanization and lifestyle changes. However, the lower prevalence of MOD in Asian descent (24% vs. 46%) compared to other ethnic groups suggests that obesity-related diabetes may be less dominant in certain East Indian subgroups, potentially due to differences in body composition or socioeconomic factors. The low prevalence of SAID (7–12%) across subgroups aligns with the autoimmune nature of this cluster, which is less common in Asian populations compared to Western cohorts.

The strong association of the SIRD cluster with kidney-related outcomes is particularly relevant for East India, where diabetic kidney disease is a major public health concern. This finding emphasizes the need for targeted screening and management strategies for SIRD patients, who may benefit from therapies addressing insulin resistance and renal protection. The cross-sectional study's confirmation of distinct cluster profiles in glycemic control, lipid metabolism, and kidney function further supports the clinical relevance of these subtypes for personalized diabetes management.

Limitations of this study include the substantial heterogeneity across studies ($I^2 > 90\%$), which may reflect variations in clustering methodologies, patient demographics, or clinical data quality. Additionally, the cross-sectional study's sample size (n=625) limits generalizability, and longitudinal data are needed to assess complication progression and treatment outcomes. Future research should explore alternative clustering techniques, such as hierarchical or model-based approaches, to uncover additional subtypes and validate these findings in larger East Indian cohorts.

5. Conclusion

In summary, this systematic review and cross-sectional study highlight the transformative role of unsupervised cluster analysis in dissecting the heterogeneity of diabetes mellitus within the East Indian population. By consistently identifying five key clusters—SAID, SIDD, SIRD, MOD, and MARD—across diverse subgroups, our findings reveal distinct phenotypic profiles that align with global observations while underscoring ethnic-specific variations, such as the elevated prevalence of SIDD and SIRD among individuals of Asian descent. These clusters not only exhibit differences in prevalence (e.g., MOD at 31% overall and MARD at 27%) but also demonstrate unique associations with clinical outcomes, particularly the heightened risk of kidney-related complications in the SIRD subtype, which is especially pertinent given the rising burden of diabetic nephropathy in South Asian populations. The cross-sectional analysis on 625 East Indian patients further validates these subtypes, showing variations in glycemic control, lipid metabolism, blood pressure, and renal function that could inform targeted interventions.

The implications of these findings extend beyond academic interest, offering a pathway toward precision medicine in East India, where diabetes prevalence is projected to exceed 100 million cases by 2030. Recognizing clusters like SIDD, which may stem from genetic predispositions to insulin deficiency prevalent in South Asians, enables early identification and tailored therapies, such as intensified beta-cell preservation strategies for high-risk Asian subgroups. Similarly, the prominence of MOD and MARD clusters reflects the influence of lifestyle factors like urbanization and obesity, suggesting public health initiatives focused on metabolic syndrome screening and lifestyle modifications could mitigate their impact. For SIRD patients, enhanced renal monitoring and insulin-sensitizing agents could reduce complication rates, addressing the unique aggressive phenotypes observed in Asian Indians.

Despite the study's strengths, including the integration of real-world data and subgroup analyses, challenges such as high heterogeneity ($I^2 > 90\%$) and the cross-sectional design limit causal inferences. Future longitudinal studies in larger East Indian cohorts, incorporating genetic markers and alternative clustering methods, are essential to refine these subtypes and evaluate long-term treatment responses. Ultimately, embracing unsupervised cluster analysis in clinical practice could revolutionize diabetes management in East India, fostering personalized care that accounts for ethnic diversity and reduces the socioeconomic burden of this epidemic. By bridging phenotypic insights with actionable strategies, we move closer to equitable, effective interventions for one of the world's most vulnerable populations.

References

[1] International Diabetes Federation. (2021). IDF Diabetes Atlas, 10th Edition.
[2] Chung, W. K., et al. (2020). Precision medicine in diabetes: a consensus report. Diabetes Care, 43(7), 1617-1635.
[3] Tuomi, T., et al. (2014). The many faces of diabetes: a disease with increasing heterogeneity. The Lancet, 383(9922), 1084-1094.
[4] Ahlqvist, E., et al. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. The Lancet Diabetes & Endocrinology, 6(5), 361-369.
[5] Sarría-Santamera, A.; Orazumbekova, B.; Maulenkul, T.; Gaipov, A.; Atageldiyeva, K. The Identification of Diabetes Mellitus Subtypes Applying Cluster Analysis Techniques: A Systematic Review. Int. J. Environ. Res. Public Health 2020, 17, 9523.
[6] National Heart, Lung, and Blood Institute. Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies.National Institutes of Health. Available online: https://www.nhlbi.nih.gov/health-pro/guidelines/in-develop/cardiovascularrisk-

reduction/tools/cohort (accessed on 27 November 2024).

[7] Moola, S.; Munn, Z.; Tufanaru, C.; Aromataris, E.; Sears, K.; Sfetcu, R.; Currie, M.; Qureshi, R.; Mattis, P.; Lisy, K.; et al. Chapter 7: Systematic reviews of etiology and risk. In JBI Manual for Evidence Synthesis; Aromataris, E., Munn, Z., Eds.; JBI: Adelaide, Australia, 2020. Available online: https://synthesismanual.jbi.global (accessed on 27 November 2024).

[8] Sterne, J.A.C.; Savovi´c, J.; Page, M.J.; Elbers, R.G.; Blencowe, N.S.; Boutron, I.; Cates, C.J.; Cheng, H.-Y.; Corbett, M.S.; Eldridge, S.M.; et al. RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ 2019, 366, l4898.