RESEARCH ARTICLE

pISSN: 0126-074X | eISSN: 2338-6223
https://doi.org/10.15395/mkb.v58.4876
Majalah Kedokteran Bandung. 2026;58(2):103–114

Majalah Kedokteran Bandung (MKB)

Received: April 1, 2026
Accepted: June 15, 2026
Available online: June 30, 2026

Development of the Diabetic Health Impact Indicator (DHII) as Composite Biomarker-Based Score for Diabetes Health Assessment

Aasma Riaz, Nasrullah Khan
College of Statistical Sciences, University of the Punjab, Lahore, Pakistan

Abstract

Diabetes mellitus is a global significant health concern that affects various body systems. Current assessment tools mostly rely on individual biomarkers, which do not fully capture the overall health burden of the disease. Therefore, a comprehensive and understandable measure is needed to evaluate the health impact associated with diabetes. This study aimed to develop the Diabetic Health Impact Indicator (DHII), a multidimensional, biomarker-based summary score designed to reflect the overall effects of diabetes, by employing a cross-sectional design with secondary data from 1,000 individuals, including diabetic, prediabetic, and non-diabetic participants. The dataset encompassed biochemical, physiological, and demographic variables related to diabetes. Exploratory Factor Analysis (EFA) was utilized to identify underlying health domains, with factor loadings used to assign weights. The empirical cumulative distribution function (ECDF) scaled the composite scores within a 0-10 range. Gaussian mixture modelling showed that DHII followed heterogeneous distribution patterns across subgroups, with K-CGMD providing the best fit in most cases. Predictive validation using logistic regression showed that DHII alone achieved an AUC of 0.721, while a baseline model with age, sex, and BMI achieved an AUC of 0.961. Adding DHII to the baseline model resulted in a marginal increase in AUC to 0.968, which was not statistically significant (DeLong's test, p=0.33). In summary, DHII is a data-driven composite index with multiple biomarker domains that demonstrates moderate discriminatory ability for diabetes status. Further validation in independent populations is required to assess its clinical use.

Keywords: Composite health indicator, diabetes mellitus, exploratory factor analysis, gaussian mixture model, risk stratification

Introduction

Diabetes mellitus is a chronic, multifactorial, metabolic disorder that represents one of the greatest challenges to public health worldwide. Estimates from the International Diabetes Federation indicate that more than 500 million adults are currently living with diabetes, and this number is expected to rise significantly in the coming decades.[1] The disease affects multiple body systems and is associated with serious complications such as cardiovascular disease, kidney dysfunction, neuropathy, and retinopathy.[2,3] Consequently, diabetes imposes a substantial health burden on individuals and healthcare systems, highlighting the need for effective and efficient tools to assess the overall impact.

Corresponding Author:
Aasma Riaz
College of Statistical Sciences, University of the Punjab, Lahore, Pakistan
Email: aasma.stat@pu.edu.pk

Given the complex and systematic nature of diabetes, accurately assessing its health burden requires consideration of multiple physiological pathways rather than reliance on isolated clinical indicators. Contemporary evidence suggests that type 2 diabetes is characterized by interconnected metabolic, inflammatory, and cardiometabolic processes that collectively influence disease progression and outcomes.[4,5] However, in both clinical practice and research, diabetes assessment is commonly based on individual biomarkers, such as fasting plasma glucose, glycated hemoglobin (HbA1c), lipid profiles, or blood pressure measurements. Although these indicators provide valuable clinical information, they may not adequately capture the multidimensional nature of the health burden associated with diabetes or the interactions among multiple physiological systems.[6]

To address this challenge, several approaches have been developed to combine multiple risk factors into a single measure. Examples

include metabolic syndrome criteria, composite risk scores, and prediction models, such as the Framingham Risk Score.[5] These methods have contributed significantly to disease-risk prediction and clinical decision-making by integrating information from multiple variables. Nevertheless, their primary objective is generally to predict future disease events rather than to quantify the current health burden of diabetes. Furthermore, many composite measures rely on predetermined clinical thresholds or weighting schemes derived from expert judgment (composite indicator methodology references, such as OECD/JRC Handbook), which may limit transparency and fail to fully reflect the underlying data structure. Conventional approaches may also overlook population heterogeneity, potentially obscuring important variations among individuals with diabetes..

Despite the availability of existing risk scores and composite measures, a gap remains in the development of a comprehensive, data-driven indicator specifically designed to summarize the multidimensional health burden of diabetes. A measure that integrates information from multiple biomarker domains, derives weights empirically from observed data, and accounts for population heterogeneity may provide a more informative assessment of overall health status. Such an indicator could be valuable for clinical assessment, population health monitoring, risk stratification, and evaluation of disease-management interventions.

Given these limitations, the present study developed the Diabetic Health Impact Indicator (DHII), a composite biomarker-based measure constructed using data-driven statistical techniques.. The proposed methodology first identifies latent health domains through exploratory factor analysis, and assigns weights according to the proportion of explained variance. The resulting composite score is than standardized using the empirical cumulative distribution function and further modelled using Gaussian mixture distributions to capture population heterogeneity (papers discussing heterogeneity in diabetes populations). By combining various biomarker dimensions, assigning weights based on data, and explicitly considering heterogeneity, DHII offers a new approach to condense the health impact of diabetes into a single, understandable, and scalable metric.

Therefore, this study aimed to develop and validate the Diabetic Health Impact Indicator (DHII), a composite biomarker-based measure designed to summarize the multidimensional health burden associated with diabetes mellitus.

Methods

This cross-sectional study used secondary data obtained from the Multiclass Diabetes Dataset uploaded by Ahlam Rashid to Mendeley Data in 2024 (DOI: 10.17632/wj9rwkp9c2.1). The dataset was collected from the Laboratory of Medical City Hospital and the Specialized Center for Endocrinology and Diabetes at Al-Kindy Teaching Hospital, Iraq. As this publicly available data contained no personally identifiable information and no direct involvement of human participants was required, ethical approval and informed consent were not necessary for this secondary data analysis. s. The final analytical dataset comprised 1,000 participants classified as diabetic, prediabetic, or non-diabetic according to standard clinical criteria used in the source dataset, including fasting plasma glucose (FPG), glycated hemoglobin (HbA1c), and body mass index (BMI).

The dataset included biochemical, physiological, and demographic variables associated with diabetes, including HbA1c, lipid profile components (total cholesterol, triglycerides, high-density lipoprotein [HDL], low-density lipoprotein [LDL], and very-low-density lipoprotein [VLDL]), renal biomarkers, and body mass index (BMI).

Participants were eligible if they were aged ≥18 years, had complete data for all biomarkers required for DHII construction, and had a clearly defined glycemic status.. Records containing duplicate entries, missing biomarker values, inconsistent diagnostic classifications, or physiologically implausible measurements were excluded. Implausible values were defined according to established clinical reference ranges, including HbA1c <3% or >20%, FPG <40 mg/dL or >600 mg/dL, BMI <10 kg/m^2 or >70 kg/m^2, total cholesterol <50 mg/dL or >500 mg/dL, triglycerides <20 mg/dL or >2,000 mg/dL, HDL <5 mg/dL or >150 mg/dL, and LDL <10 mg/dL or >400 mg/dL. Because the study used a publicly available, de-identified dataset and involved no direct participant contact, biological sample collection, or access to personally identifiable information, ethical approval and informed consent were not require.

Participants were further categorized (<30, 30–44, 45–59, and ≥60 years) for descriptive purposes. Body mass index (BMI) was classified

according to World Health Organization criteria as underweight (<18.5 kg/m^2), normal weight (18.5–24.9 kg/m^2), overweight (25.0–29.9 kg/m^2), and obese (≥30 kg/m^2).

The continuous biomarkers were screened for extreme values before analysis. To mitigate the effect of potentially erroneous or influential outliers while keeping all data points, winsorization was used. The outliers were managed by winsorizing at percentile extremes to reduce their impact. Continuous variables were subsequently standardized to facilitate comparison across different measurement scales and support multivariable analyses..

The Diabetic Health Impact Indicator (DHII) was constructed using exploratory factor analysis (EFA) to identify latent health domains among the selected biomarkers. The suitability of the data for factor analysis was assessed using the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's test of sphericity. Factors were retained based on eigenvalues, scree plot evaluation, and interpretability of the resulting domains.[7] This approach reduced mutiple correlated biomarkers into smaller number of clinical meaningful dimensions representing major physiological systems affected by diabetes.

The retained latent domains were interpreted as key components of diabetes-related health burden, including metabolic control, lipid metabolism, and renal function. The final DHII score was calculated using variance-based weights derived from the proportion of total variance explained by each factor..

Let F1, F2, ...,Fk denote the retained latent factor scores obtained from exploratory factor analysis and let w1, w2, ..., wk represent the corresponding proportions of explained variance. The unscaled DHII score for participant *i* was calculated as: $DHII\_i = \Sigma (wj \times Fij)$

where $\Sigma wj = 1$.

To improve interpretability, the composite score was standardized to a 0–10 scale using the empirical cumulative distribution function (ECDF):For participant i:

$DHII\_standardized = 10 \times ECDF(DHII\_i)$

This transformation maintained the rank order of individuals while generating a clinically interpretable score from 0 (lowest burden) to 10 (highest burden). Unlike simple min–max normalization, the ECDF preserves the empirical rank ordering and accounts for distributional skew, thereby anchoring scores to observed population quantiles.

Latent subgroup identification was performed using finite Gaussian mixture modelling applied to the standardized DHII scores.[8] Models containing one to six latent classes were estimated under the assumption that observations within each class followed a normal distribution. The optimal number of latent classes was determined using the Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), entropy statistics, and clinical interpretability.[9] The model with the lowest BIC, adequate class separation, and meaningful clinical interpretation was selected as the final model.

For clinical interpretation, the standardized DHII scores were categorized into seven ordinal groups representing increasing levels of diabetes-related health burden..

The predictive validity of DHII was evaluated using multivariable logistic regression.[10-12] Diabetes status (diabetic versus non-diabetic/prediabetic) was specified as the dependent variable, while DHII served as the primary predictor. Age, sex, and body mass index were included as covariates because of their established associations with diabetes. Results were reported as odds ratios (ORs) with 95% confidence intervals (CIs) and p-values.

Model discriminative was assessed using the area under the receiver operating characteristic curve (AUC), whereas calibration was evaluated using goodness-of-fit tests and graphical calibration plots.[13]. Internal validation was performed using stratified 10-fold cross-validation, in which the dataset was divided into ten approximately equal subsets while preserving outcome-class proportions. In each iteration, nine subset were used for model training, and one subset for validation. The model's performance was summarized by the mean and standard deviation of the AUC, sensitivity, specificity, and calibration measure across all validation folds.

All statistical analyses were performed using R software version 4.3.1 (R Foundation for Statistical Computing, Vienna, Austria). Exploratory factor analysis, Gaussian mixture modelling, and validation procedures were conducted using appropriate R packages. The development and validation of the predictive models were performed in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) recommendations.[10]

Results

The analytic cohort comprised 1,000 individuals,

Table 1 Characteristics of the Study Population (n=1,000)

Variable	Category	n	%	Mean ± SD
Gender	Male	565	56.5	
	Female	435	43.5	
Age (years)	Young (20–40)	101	10.1	
	Middle-aged (41–50)	137	13.7	
	Older Adults (51–60)	614	61.4	
	Elderly (61–80)	148	14.8	
	—			53.53±8.80
BMI (kg/m^2)	Normal (18.5–24.9)	190	19	
	Overweight (25–29.9)	284	28.4	
	Obese (30–34.9)	393	39.3	
	Severely Obese (≥35)	133	13.3	
	—			29.58±4.96
DHII (0–10)	—			5.00±2.89

of whom 56.5% were male and 43.5% were female. The mean age was 53.5±8.8 years, and most participants (61.4%) were aged 51–60 years. The mean BMI was 29.6±5.0 kg/m^2, with the majority of participants classified as overweight or obese. The mean DHII score was 5.0±2.9 on the standardized 0–10 scale (Table 1). Biomarker characteristics across normal, prediabetic, and diabetic groups are presented in Table 2. One-way ANOVA demonstrated significant differences in HbA1c ($p<0.001$), total cholesterol ($p<0.001$), triglycerides ($p<0.001$), VLDL ($p=0.004$), and BMI ($p<0.001$). No significant differences were observed for urea ($p=0.063$), creatinine ($p=0.496$), HDL-C ($p=0.652$), or LDL-C ($p=0.735$).

Post-hoc Tukey HSD analysis showed significant differences among all three groups for HbA1c, between normal and diabetic groups for total cholesterol, triglycerides, and VLDL-C, and between diabetic and both normal and prediabetic groups for BMI. BMI was

Table 2 Biomarker Characteristics Across Normal, Prediabetic, and Diabetic Groups

Biomarker	Normal (N)	Prediabetic (P)	Y (Diabetic)	ANOVA p-value
Urea (mmol/L)	4.68±2.52	4.51±2.02	5.22±3.02	0.063
Creatinine (µmol/L)	62.8±30.02	66.08±41.57	69.87±63.58	0.496
HbA1c (%)	4.56±0.92	6±0.19	8.88±2.26	<0.001
Total Cholesterol (mmol/L)	4.27±1.28	4.58±1.04	4.95±1.3	<0.001
Triglycerides (mmol/L)	1.63±1.03	2.13±1.06	2.45±1.43	<0.001
HDL-C (mmol/L)	1.23±0.51	1.13±0.38	1.21±0.69	0.652
LDL-C (mmol/L)	2.63±0.98	2.49±0.87	2.62±1.14	0.735
VLDL-C (mmol/L)	0.94±1.48	0.98±0.5	2.02±3.93	0.004
BMI (kg/m^2)	22.37±1.42	23.93±2.71	30.81±4.32	<0.001

Table 3 Quantile-Based Classification of DHII into Seven Ordered Categories (n=1,000)

Class	Lower Bound	Upper Bound	Count	Percentage
Very Low	0.01	1.41	140	14.0
Low	1.41	2.81	140	14.0
Moderate Low	2.81	4.21	140	14.0
Moderate	4.21	5.71	151	15.1
Moderate High	5.71	7.11	139	13.9
High	7.11	8.50	140	14.0
Very High	8.50	10.00	150	15.0

substantially higher in diabetic participants than in normal and prediabetic participants. Similarly, total cholesterol, triglycerides, and VLDL-C levels were significantly elevated in the diabetic group compared with the normal group.

Exploratory factor analysis (EFA) was performed using nine biomarkers (urea, creatinine, HbA1c, Cholesterol, TG, HDL, LDL, VLDL, and BMI). Sampling adequacy was acceptable (KMO=0.78), and Bartlett's test of sphericity was significant ($p<0.001$), supporting the suitability of the data for factor analysis. A three-factor solution with eigenvalues >1 was retained and explained 40.4% of the total variance. Factor 1, was characterized primarily by total cholesterol and LDL with additional contribution from BMI, representing a lipid–obesity domain. Factor 2 loaded strongly on Urea and Creatinine, with moderate HbA1c contribution, representing a renal–glycemic domain. Factor 3, was defined primarily mainly by BMI and HbA1c, reflected a metabolic–glycemic domain. Correlations among the retained factors were low (all $r<0.10$). The retained factors accounted for 39%, 35%, and 27% of the explained variance, respectively. The Diabetic Health Impact Indicator (DHII) was constructed as a weighted sum of factor scores.

Accordingly, the lipid–obesity domain contributed the greatest weight to the DHII, followed by the renal–glycemic and metabolic–glycemic domains. To facilitate interpretation, raw DHII scores were standardized to a 0–10 scale using the empirical cumulative distribution function (ECDF).For descriptive purposes, standardized DHII score were categorized into seven quantile-based categories: Very Low, Low, Moderate Low, Moderate, Moderate High, High, and Very High. Each category contained approximately 13–15% of participats, with score cutoffs ranging from 0.01–1.41 for Very Low category to 8.50–10.00 for Very High category

Table 4 Mean DHII Scores According to Glycemic Status, Age Group, and BMI Category

Class	Age Group	Normal	Overweight	Obese	Severely Obese
Normal	Young (20–40)	2.98±2.45	–	–	–
	Middle-aged (41–50)	2.63±2.63	–	–	–
	Older Adults (51–60)	4.60±2.48	–	–	–
	Elderly (61–80)	4.27±3.49	–	–	–
Prediabetic	Young (20–40)	4.23±2.96	3.05±0.00	3.97±4.16	–
	Middle-aged (41–50)	3.04±2.21	3.32±2.02	NA	–
	Older Adults (51–60)	1.42±NA	5.02±NA	NA	–
Diabetic	Young (20–40)	6.42±2.59	5.04±2.20	4.25±3.38	7.17±1.81
	Middle-aged (41–50)	6.29±2.27	2.90±1.83	6.12±2.63	–
	Older Adults (51–60)	5.92±1.95	5.07±2.77	5.16±2.79	5.49±2.90
	Elderly (61–80)	5.18±6.17	5.11±2.73	6.45±2.91	5.91±3.19

Note: Data are presented as mean ± standard deviation (SD). — = no participants in the subgroup

Table 5 Comparison of KCGM and Best-Fitting Univariate Distributions Across Age and BMI Categories

Category	KCGM BIC	Best Univariate Distribution	BIC
Young (20–40)	-495.65	Arcsine	-56.30
Middle-aged (41–50)	-639.69	Exponential	-66.42
Older Adults (51–60)	-2929.25	4P Beta	-109.73
Elderly (61–80)	-718.31	Johnson SB	-75.46
Normal	-898.91	Exponential	-65.07
Overweight	-1345.95	Triangular	-71.10
Obese	-1895.85	4P Beta	-81.49
Severely Obese	-640.21	Johnson SB	-60.93

Note: KCGM = Kernel-Copula Gaussian Mixture Model; BIC = Bayesian Information Criterion. Lower BIC values indicate better model fit

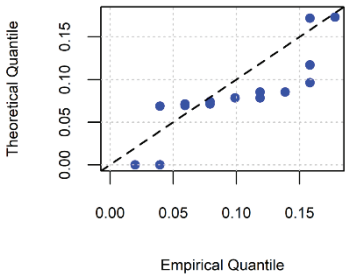
(Table 3). DHII values across glycemic status, age groups, and BMI categories are presented in Table 4).

In general, mean DHII scores were higher among diabetic participants than among prediabetic and non-diabetic participants across most age and BMI strata. Several subgroup estimates were based on a limited number of observations, resulting in missing values or unstable standard deviations in some cells. Therefore, these subgroup results should be interpreted descriptively..

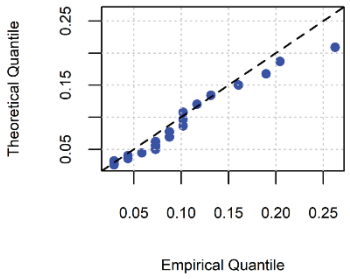
To further characterize the distribution of DHII, finite Gaussian mixture models and alternative parametric distributions. Model selection based on thr Bayesian Information Criterion (BIC) indicated that the Gaussian mixture model provided a better fit than several competing parametric distributions. The best-fitting univariate distribution identified for each age and BMI subgroup, together with its corresponding BIC value, is presented in Table 5

The best-fitting distributions varied across age and BMI strata. Among age groups, the Arcsine distribution provided the best fit for young adults (20–40 years), the Exponential distribution for middle-aged adults (41–50 years), the 4-parameter Beta distribution for older adults (51–60 years), and the Johnson SB distribution for elderly participants (61–80 years).. Across BMI categories, the Exponential, Triangular, 4-parameter Beta, and Johnson SB distributions provided the best fit for normal-weight, overweight, obese, and severely obese participants, respectively. Probability plots (Figure 1) and quantile–quantile plots (Figure 2) demonstrated distinct distributional patterns across the evaluated subgroups.

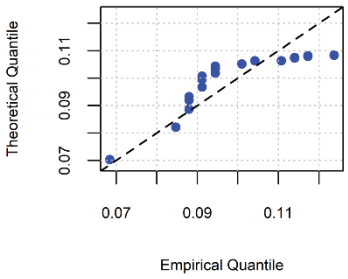
Distribution fitting analyses were performed across all study subgroups using both mixture-



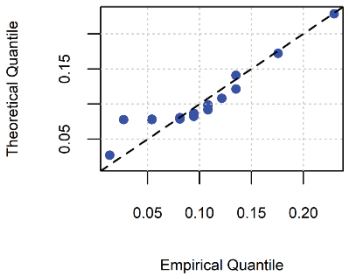
a) Young (20–40)



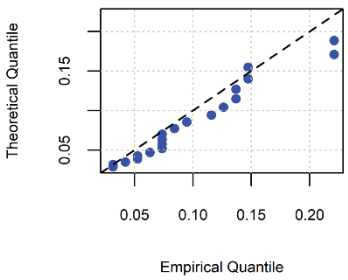
b) Middle-aged (41–50)



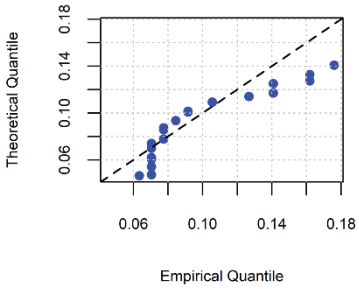
c) Older Adults (51–60)



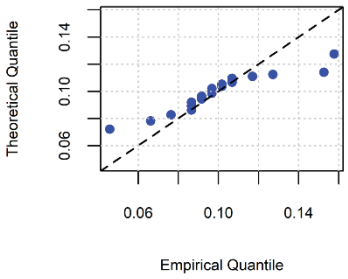
d) Elderly (61–80)



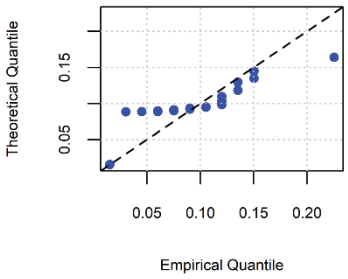
e) Normal



f) Overweight



g) Obese

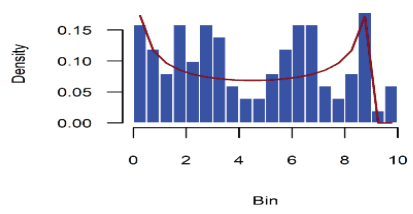


h) Severely Obese

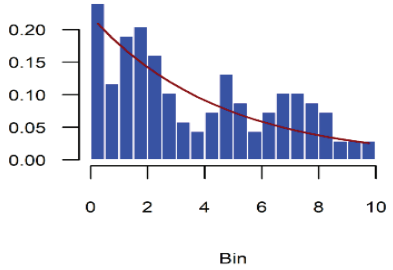
Figure 1 Probability Plots of the Best-Fitting Univariate Distributions for DHII Across Age and BMI Categories

based and classical probability distributions. Among the evaluated models, the K-CGMD provided the best fit for most subgroups according to the selected goodness-of-fit criteria.

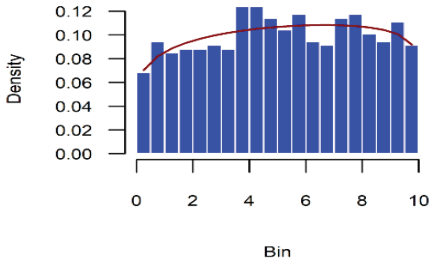
The distribution of DHII was subsequently examined across non-diabetic, prediabetic, and diabetic groups. Predictive performance was evaluated using logistic regression models.. A model including only DHII (Model B) achieved an AUC of 0.721. A baseline model incorporating age, sex, and BMI (Model A) yielded an AUC of 0.961. Addition of DHII to the baseline



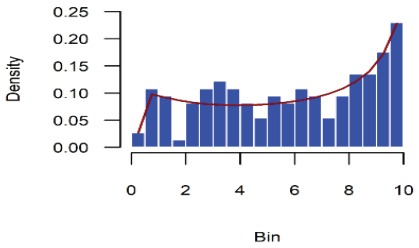
a) Young (20–40)



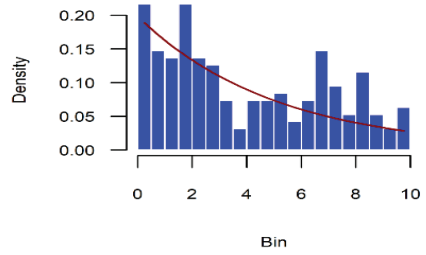
b) Middle-aged (41–50)



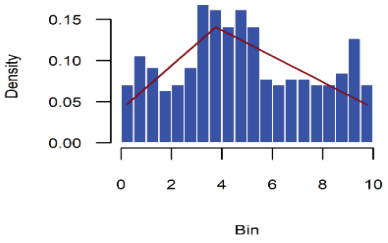
c) Older Adults (51–60)



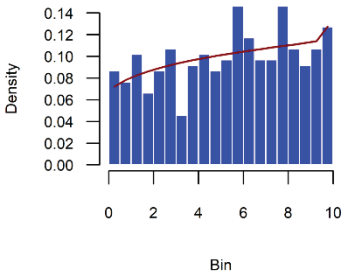
d) Elderly (61–80)



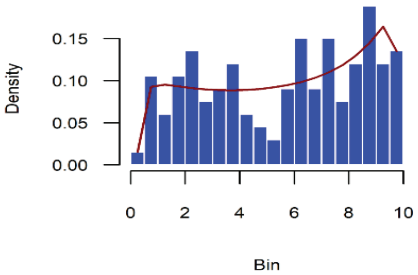
e) Normal



f) Overweight



g) Obese



h) Severely Obese

Figure 2 Quantile–Quantile Plots of the Best-Fitting Univariate Distributions for DHII Across Age and BMI Categories

model (Model C), increased the AUC to 0.968; however, this improvement was not statistically significant according to DeLong's test (Z −0.97, p=0.33). The mode incorporating individual biomarkers (Model D) demonstrated the highest discrimination, with an AUC of 0.983.

Model calibration was assessed using the Hosmer–Lemeshow goodness-of-fit test, which showed acceptable agreement between predicted and observed probabilities (χ^2=7.96, df=8, p=0.44). Calibration plots showed close correspondence between predicted and observed event probabilities. Cross-validation of Model C yielded a mean ROC value of 0.960, with sensitivity of 0.78 and specificity of 0.96.

The predictive performance analyses showed that DHII alone achieved moderate discriminatory ability (AUC=0.721). In comparison, a model based on conventional demographic predictors (age, sex, and BMI) achieved higher discrimination (AUC=0.961). The addition of DHII to the conventional model resulted in a modest increase in predictive performance (AUC=0.968); however, this improvement was not statistically significant according to DeLong's test (p=0.33).

Discussion

This study developed and evaluated the Diabetic Health Impact Indicator (DHII), a composite biomarker-based measure designed to summarize the multidimensional health burden associated with diabetes mellitus.. The analytic pathway integrated latent structure extraction, weighted aggregation, probabilistic distribution modeling, and validation through predictive performance, thereby providing both methodological rigor and applied relevance.[-14]
Overall, the results indicate that DHII offers a composite score based on various metabolic and renal biomarkers. The index varied across different glycemic, age, and BMI groups and demonstrated acceptable calibration and reproducibility. However, due to the strong predictive power of traditional predictors and only slight improvements with DHII, further validation is needed to confirm its additional clinical and public health benefits in other populations.
The demographic and clinical characteristics of the study population were consistent with the known epidemiology of diabetes, with most participants belonging to older age groups and higher BMI categories. Biomarker profiles differed significantly across glycemic groups, particularly for HbA1c, triglycerides, VLDL-C, total cholesterol, and BMI. These findings support the biological relevance of the selected biomarkers as components of a composite measure intended to reflect diabetes-related health burden..

Exploratory factor analysis reduced the correlated biomarker set into three latent domains representing lipid–obesity, renal–glycemic, and metabolic–glycemic dimensions. Together, these domains explained approximately 40% of the total variance and provided the basis for the weighted construction of DHII. The use of variance-based weighting allowed the contribution of each domain to be determined empirically rather than through predefined assumptions, consistent with previously reported approaches for developing composite health indicators..[15]

Following scaling to a 0–10 scale, DHII scores were categorized into seven quantile-based groups to facilitate interpretation. Diabetic participants generally demonstrated higher DHII scores than prediabetic and non-diabetic participants across most age and BMI strata. Higher scores were particularly observed among obese and older individuals, which may reflect the combined effects of adiposity, aging, and metabolic dysregulation on diabetes-related health burden.16–19 These findings provide preliminary support for the ability of DHII to differentiate levels of health burden across clinically relevant population subgroups..[16,17,19]

Distributional analyses demonstrated heterogeneity in DHII across age and BMI groups. Although mixture models provided a better statistical fit than alternative distributions, the principal finding is that diabetes-related health burden is not uniformly distributed across individuals.. This heterogeneity supports the use of multidimensional measures such as DHII, which integrate information from multiple biomarker domains and may capture aspects of health burden that are not reflected by individual biomarkers alone. The distribution-fitting analyses should therefore be interpreted primarily as methodological support for the underlying structure of DHII rather than as direct evidence of clinical outcomess.[18]

The observed variation in DHII across different age and BMI groups suggests that this index may be useful for identifying subgroups with differing levels of diabetes-related health burden. Higher DHII values were generally observed among older and obese individuals,

consistent with the established associations between aging, adiposity, metabolic dysfunction, and diabetes-related complications. These findings indicate that DHII may have potential applications in population health assessment and risk stratification; however, further validation in independent populations is required before clinical implementation can be recommended. s.

Predictive validation showed that DHII alone provided moderate discrimination between glycemic groups. In contrast, conventional predictors, including age, sex, and BMI, achieved substantially higher predictive performance. Furthermire, the addition of DHII to the conventional model resulted in only a modest and statistically non-significant improvement in discriminationdiscrimination.These findings suggest that the primary value of DHII's may lie in its ability to summarize multidimensional diabetes-related health burden into a single composite measure rather than in improving prediction beyond established clinical risk factors. Accordingly, DHII should be considered a complementary assessment tool rather than a replacement for established predictors or comprehensive biomarker evaluation.

When compared with established diabetes assessment tools, DHII differs in both structure and scope. HbA1c primarily reflects long-term glycemic control and does not capture extra-glycemic metabolic or organ-related complications. Similarly, metabolic syndrome criteria summarize clustered cardiovascular risk factors but remain threshold-based and do not provide a continuous burden score. The Diabetes Complications Severity Index (DCSI) focuses on the severity of established complications, whereas risk prediction tools such as the UKPDS Risk Engine and Framingham Risk Score are designed to estimate future cardiovascular events rather than quantify the current physiological burden of diabetes. In contrast, DHII integrates metabolic, lipid, and renal biomarkers into a continuous composite score, providing a multidimensional summary of diabetes-related health burden.

Compared with traditional approaches to diabetes assessment, DHII serves a different purpose. While biomarkers such as HbA1c primarily reflect glycemic control and conventional risk models estimate the likelihood of future disease outcomes, DHII integrates metabolic, renal, and lipid-related biomarkers into a single composite measure. This multidimensional approach is intended to provide a broader representation of the overall health burden associated with diabetes. Such a measure may facilitate communication of disease burden, enable comparisons across population subgroups, and support population-level monitoring.

Potential applications of DHII include the assessment of overall diabetes-related health burden among individuals with diabetes or prediabetes. By summarizing information from multiple physiological domains into a standardized score, DHII may complement routine biomarker evaluation and provide an integrated overview of metabolic health status.. However, because the addition of DHII to conventional predictors resulted in only a modest and statistically non-significant improvement in predictive performance, its role may be more appropriate as a descriptive and integrative measure rather than as a standalone predictive tool. Further validation in independent populations is required before its clinical or public health utility can be established..

Taken together, these findings suggest that DHII may provide as a a useful composite measure for summarizing diabetes-related health burden. It captures variance across multiple biomarker domains, stratifies individuals into interpretable risk categories, accommodates heterogeneity through flexible distribution modeling, and demonstrates predictive validity for diabetes status. While recalibration in external populations is recommended, the DHII provides a strong foundation for both epidemiological research and healthcare policy applications aimed at quantifying and monitoring the health burden of diabetes.[13,17-19]

This study has several limitations. First, the cross-sectional design precludes causal inference and does not permit evaluation of changes in health burden over time. Secondly, the analysis was based on a publicly available secondary dataset comprising 1,000 individuals classified as non-diabetic, prediabetic, or diabetic; therefore, the findings may reflect characteristics specific to the source population and may not be generalizable to other populations, ethnic groups, or healthcare settings. Third, external validation was not performedand remains necessary before broader application of DHII can be considered.. Furthermore, although DHII was designed as an integrative measure, its addition to conventional predictors resulted in only a modest and statistically non-significant improvement in predictive performance. Consequently, DHII should be viewed as a complementary measure rather than a replacement for established clinical

assessment tools.

Future studies should evaluate its reproducibility, calibration, clinical utility, and comparative performance against established measures such as HbA1c-based models, metabolic syndrome classifications, and the Diabetes Complications Severity Index (DCSI). In conclusion, the Diabetic Health Impact Indicator (DHII) is a data-driven composite score that integrates information from multiple biomarker domains into a standardized measure of diabetes-related health burden. DHII varied across glycemic categories, age groups, and BMI categories and demonstrated moderate ability to discriminate diabetes status. Although DHII could be useful for summarizing disease burden and aiding risk stratification at the population level, its additional predictive value beyond standard clinical predictors was limited in this dataset. Further validation in independent populations is required before routine clinical or public health implementation can be recommended.

References

1. International Diabetes Federation. IDF Diabetes Atlas. 10th ed. Brussels: IDF; 2021.
2. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, et al. Global and regional diabetes prevalence estimates for 2019 and projections. Diabetes Res Clin Pract. 2019;157:107843. doi:10.1016/j.diabres.2019.107843
3. O'Hearn M, Lauren BN, Wong JB, Kim DD, Mozaffarian D. Trends and disparities in cardiometabolic health among U.S. adults, 1999–2018. J Am Coll Cardiol. 2022;80(2):138–51. doi:10.1016/j.jacc.2022.04.046
4. Rawshani A, Rawshani A, Franzén S, Sattar N, Eliasson B, Svensson AM, et al. Risk factors, mortality, and cardiovascular outcomes in patients with type 2 diabetes. N Engl J Med. 2018;379(7):633–44. doi:10.1056/NEJMoa1800256
5. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus. Nat Rev Endocrinol. 2018;14(2):88–98. doi:10.1038/nrendo.2017.151
6. Pearson ER. Type 2 diabetes: a multifaceted disease. Diabetologia. 2019;62(7):1107–12. doi:10.1007/s00125–019–4909-y
7. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374(2065):20150202. doi:10.1098/rsta.2015.0202
8. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J. 2016;8(1):289–317. doi:10.32614/RJ-2016–021
9. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis. Lancet Diabetes Endocrinol. 2018;6(5):361–9. doi:10.1016/S2213–8587(18)30051–2
10. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594. doi:10.1136/bmj.g7594
11. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8. doi:10.7326/M18–1376.
12. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245–7. doi:10.1016/j.jclinepi.2015.04.005.
13. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi:10.1186/s12916–019–1466–7
14. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal. BMJ. 2020;369:m1328. doi:10.1136/bmj.m1328.
15. Riley RD, van der Windt D, Croft P, Moons KGM. Prognosis research strategy (PROGRESS) framework. BMJ. 2019;364:l459. doi:10.1136/bmj.l459
16. Grundy SM. Metabolic syndrome update. Circulation. 2016;134(1):e10–e19. doi:10.1161/CIRCULATIONAHA.115.018585
17. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support. J Med Internet Res. 2018;20(5):e10775. doi:10.2196/10775
18. Dennis JM, Shields BM, Henley WE, Jones AG, Hattersley AT. Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with

models based on simple clinical features: an analysis using clinical trial data. Lancet Diabetes Endocrinol. 2019;7(6):442–51. doi:10.1016/S2213–8587(19)30087–7

19. Gregg EW, Sattar N, Ali MK. The changing face of diabetes complications. Lancet Diabetes Endocrinol. 2016;4(6):537–47. doi:10.1016/S2213–8587(16)30010–9