scientific reports

Check for updates

OPEN

Integrating ECG-derived features with conventional CVD risk models

Maryam Mahdavi[1], Anoshirvan Kazemnejad[1]✉, Abbas Asosheh[2]✉, Davood Khalili[3], Kamyab Hosseinpour[4] & Ahmadreza Tajari[5]

Non-communicable diseases (NCDs), particularly cardiovascular diseases (CVDs), have become the leading cause of mortality worldwide, with Iran exhibiting higher-than-average incidence and mortality rates. Early detection of high-risk individuals is critical, as CVD often progresses silently. Electrocardiogram (ECG) signals may enhance risk prediction beyond Framingham risk score (FRS). This study aimed to evaluate the predictive performance of ECG signal features for incident CVD using signal processing in a large population-based cohort from the Tehran Lipid and Glucose Study (TLGS). A total of 4,637 adults aged 40 years devoid of past CVD at baseline (2006–2008) were followed up until 2018. Baseline characteristics, laboratory measurements, and ECG signal features were collected. CVD events were defined as coronary heart disease (CHD) or stroke. A recalibrated FRS (baseline) model assessed the association between ECG features and incident CVD, with model performance evaluated using Harrell's C-index, Net Reclassification Index (NRI), and Integrated Discrimination Improvement (IDI). Over a 10-year follow-up, 483 participants (10.4%) developed CVD. The introduction of ECG signal features improved risk prediction in women, increasing the Harrell's C-index from 0.84 to 0.85 and demonstrating significant reclassification improvement (NRI: 55.7%, IDI: 2.8%). However, no meaningful improvement was observed in men. ECG-based modeling outperformed FRS, particularly for intermediate-risk categories among women. Incorporating ECG signal features into risk models significantly enhanced CVD prediction performance in women, suggesting potential utility for improving individualized preventive strategies. Further research is warranted to refine ECG-based risk stratification tools for broader clinical application.

Keywords ECG signal, NRI and IDI, CVD, Prediction model

Non-communicable diseases (NCDs) have emerged as a major global public health concern, accounting for nearly 60% of total annual mortality worldwide[1]. Among these, cardiovascular diseases (CVDs) represent a leading contributor, affecting a substantial portion of the global population[2]. Cardiovascular diseases (CVDs) constitute the predominant cause of mortality in Iran[3], with age-standardized incidence and mortality rates, as well as disability-adjusted life years (DALYs) attributed to CVDs, surpassing global averages[4]. The increasing prevalence of sedentary lifestyles, rising obesity rates, and declining physical activity levels among Iranians are expected to exacerbate the burden of CVDs in the coming years[5]. Demographic transitions, including an aging population, are likely to further elevate the prevalence of these conditions.

CVDs count as one of the primary causes of morbidity and mortality on a global scale. In 2015, they were responsible for approximately 17.9 million deaths and 347.5 million DALYs[6,7]. Although mortality rates from CVDs are declining in high-income countries, a significant proportion of deaths still occur in low- to middle-income countries, particularly in the Eastern Mediterranean region[7]. In Iran, there has been a notable shift in mortality patterns from infectious diseases to NCDs over recent decades, with CVDs being highly prevalent[8]. Iran's high prevalence of CVDs presents significant healthcare and economic challenges[9]. Cardiovascular complications such as stroke, myocardial infarction (MI), and coronary artery disease (CAD) can result in both fatal and non-fatal outcomes. While advancements in treatment have contributed to reduced mortality, many individuals continue to suffer from long-term complications, including psychological distress, fatigue,

[1]Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran. [2]Department of Medical Informatics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran. [3]Prevention of Metabolic Disorders Research Center, Research Institute for Metabolic and Obesity Disorders, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [4]Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran. [5]Institute for Biological Information Processes (IBI), Molecular and Cellular Physiology (IBI-1), Forschungszentrum Jülich, Jülich, Germany. ✉email: kazem_an@modares.ac.ir; asosheh@modares.ac.ir

sleep disturbances, dyspnea, and reduced quality of life[10]. In severe cases, CVDs may progress to heart failure, recurrent MI, or sudden cardiac death[11].

Cardiovascular disease (CVD), a multifactorial and chronic condition stemming from various heart and vascular system disorders, remains the primary cause of premature mortality and long-term disability globally. Although pharmacological and surgical interventions are routinely employed to manage CVD, they do not offer a permanent cure and often have a considerable impact on patients' quality of life. Consequently, contemporary approaches to CVD management place significant emphasis on prevention. Emerging evidence suggests that up to 80% of premature deaths attributed to CVD may be preventable through timely interventions[12]. Since CVD typically progresses slowly and may remain asymptomatic for extended periods, it is frequently diagnosed in its advanced stages, when treatment becomes more complex. Thus, early detection of individuals at high risk is essential for implementing effective preventive strategies[13]. In line with this, recent clinical guidelines increasingly advocate for the use of CVD risk prediction models to identify individuals who may benefit from early, targeted preventive measures. Notably, Framingham risk score (FRS) have formed the basis for most of these models[14].

Advancements in computer science, coupled with the demand for precision medicine, have led to an increase in multidimensional data from various sources. This necessitates the development of advanced tools and models capable of processing, understanding, and analyzing this vast and intricate data. These tools must also accurately forecast outcomes and predict risks. The most effective predictive model, which delivers optimal performance, is determined by several key factors. These include the specific objectives and purposes for which the models are developed, their ability to generalize across different datasets, their robustness in handling diverse conditions, and their capacity to produce consistent and reproducible results when applied in real clinical settings[15]. Our goal in doing this research was to quantify the association between ECG signals and incident CVD in a population-based cohort called the Tehran and Lipid Glucose Study (TLGS) during more than a decade of follow-up. This study explores whether integrating ECG signal data can improve cardiovascular disease prediction.

Materials and methods

Participants

The Tehran Lipid and Glucose Study (TLGS), launched in 1999, is a population-based prospective cohort study conducted in Tehran's District 13, aimed at examining risk factors associated with non-communicable diseases (NCDs). The study design has been described in previous publications[16]. In brief, the first phase (1999–2001) was cross-sectional, enrolling 15,005 individuals aged ≥ 3 years through a multistage random sampling method. The study continued as a longitudinal follow-up, with data from 8,071 participants aged 40 to 79 years in the third phase used for this analysis. The sixth phase (2015–2018) follow-up included cardiovascular disease (CVD) events such as coronary heart disease (CHD) or stroke, along with follow-up duration. The study was conducted in multiple phases:

- Phase one (1999–2001) and phase two (2002–2005) followed a multistage, random cluster sampling method.
- Follow-ups occurred at approximately 3.5-year intervals, continuing through phases three to seven (2006–2020), with an average 73% participation rate per phase.

For the current analysis, 5,479 adults aged 40–79 years who took part in the third examination cycle were initially considered. We entered the participants in phase III of TLGS for data analysis because ECG signals were gathered in this phase onwards. After excluding individuals based on specific criteria, history of CVD ($n = 772$), lost to follow-up for CVD events ($n = 15$), and missing ECG data ($n = 2207$). The final study population included 4,637 adults without any established CVD (CHD or stroke). Participants had no previous record of coronary artery disease (including angina, myocardial infarction, coronary artery bypass grafting, or percutaneous coronary intervention), cerebrovascular conditions (such as stroke or transient ischemic attack), or peripheral arterial disease (such as claudication) at the third examination cycle, which served as the baseline, with follow-up extending until March 2018 (Fig. 1). This large-scale, population-based study continues to be a valuable resource for monitoring risk factors that correspond to chronic diseases and cardiovascular conditions.

Measurements

Eligible participants underwent initial interviews to collect socio-demographic and medical data. All measurements were conducted by trained staff following standardized study protocols. Further details on the measurement methods can be found in previous studies [19]. Participants were seated, and their diastolic blood pressure (DBP) and systolic blood pressure (SBP) were assessed twice by a general physician after they had rested for 15 min. A standard mercury sphygmomanometer was used consistently to measure blood pressure, with the first reading determining the maximum inflation level and the average of two readings recorded. For laboratory assessments, participants were directed to fast for 12 to 14 h before blood sampling. Samples were immediately transported to the TLGS laboratory for analysis using Selectra autoanalyzers (Vital Scientific, Spankeren, Netherlands). Fasting plasma glucose (FPG) was measured using an enzymatic colorimetric method based on glucose oxidation. Lipid profile assessments were performed using Pars Azmoon (Tehran, Iran) commercial kits. Total cholesterol (TC) was measured through enzymatic colorimetric tests utilizing cholesterol esterase, cholesterol oxidase, and glycerol phosphate oxidase. High-density lipoprotein cholesterol (HDL-C) was measured using phosphotungstic acid precipitation. Laboratory tests were conducted only when internal quality control values fell within the acceptable range. For more information on the measurement methods, please refer to other studies[17].

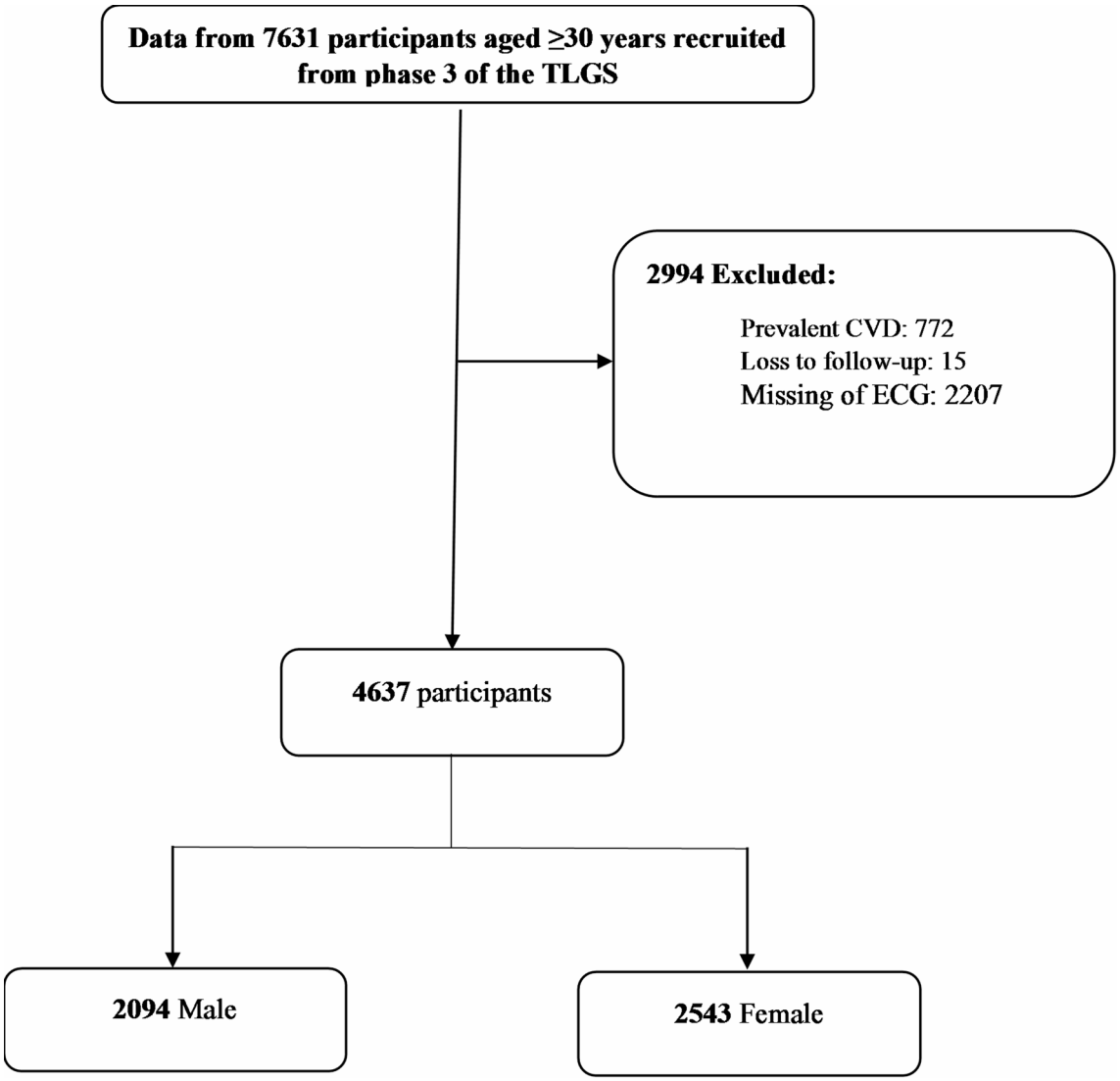

Fig. 1. Study flowchart illustrating the selection of participants from the Tehran Lipid and Glucose Study (TLGS) cohort.

Definitions of outcome

Participants were followed up annually for cardiovascular disease (CVD) events by trained nurses through telephone interviews. If an event was reported, a trained physician collected additional data through home visits or hospital record reviews. The final diagnosis was confirmed by the Cohort Outcome Panel[18]. Coronary heart disease (CHD) was defined as the occurrence of myocardial infarction (MI), probable MI, unstable angina pectoris, or angiography-confirmed coronary artery disease (CAD). In this study, cardiovascular disease (CVD) was characterized as the incidence of either stroke or coronary heart disease (CHD), with the time until the initial occurrence of either event recorded as the time-to-event. Further details of CVD outcomes are available in previous publications[19].

ECG signal

The electrocardiogram (ECG) captures variations in electrical potentials across chest surface electrodes, reflecting cardiac activity. Each heartbeat appears as a series of deviations from the ECG baseline, corresponding to the heart's electrical activity that drives muscle contraction[20]. Key ECG components include the P wave, QRS complex, and T wave. Features were extracted for 12 leads in ECG signals across multiple domains, including time[21], frequency[21], time–frequency[22], peak-R and morphological distances[23], and Heart Rate Variability (HRV)[24,25] (Table 1). The selection of features for men and women participants was driven by clinical expert opinion and literature, taking into account the sex-specific differences in CVD risk factors and incidence[26].

Analysis method

Normally distributed continuous variables were presented as mean and standard deviation (SD), and categorical baseline characteristics were described as frequency (%). The independent T-test was applied to analyze

Domain	Feature	Description
Frequency	The amplitude of the ECG signal over time	These features use the Fast Fourier Transform (FFT) to analyze the signal in the frequency domain
	The average and midpoint of the signal values	
	The amount of signal dispersion around the mean	
	The minimum and maximum signal amplitudes and the distance between them	
	The amount of asymmetry in the distribution of amplitudes	
	The peak intensity of the distribution of the number of outliers	
	The root mean square value describes the signal's overall power and indicates the signal's overall strength	
Time	The total energy of the signal across all frequencies	A general statistical representation of the amplitude of the ECG signal over time
	Identifying the frequency that has the most energy	
	The center of gravity is the frequency spectrum associated with the heart rate. It determines whether the energy is concentrated at low or high frequencies. These features are useful for detecting arrhythmias or heart rhythm abnormalities	
Wavelet	The energy of the wavelets at each level, which is useful for identifying sharp changes such as the QRS complex	Wavelet decomposition decomposes the ECG signal into different levels (up to level 3) to extract time-frequency information from the signal
	Average wavelet coefficients at each level, which are very useful for identifying local changes in the signal structure and play an important role in detecting ectopic beats or ischemia	
Morphological distances	The maximum amplitude of one of the peak-R waves	These features are extracted based on R-peaks (peak points of the heart rate)
	The average duration of the QRS complex indicates the duration of ventricular activation. The time interval from the beginning of the P wave to the peak of the R wave	
	The time interval from the beginning of the Q wave to the end of the T wave is very important for diagnosing electrical disorders of the heart	
	The fluctuation of each of these intervals	
Heart Rate Variability (HRV)	The average of the R-R intervals, which is the inverse of the heart rate	HRV features measure the time variation between successive R-peaks and are indicative of autonomic nervous system activity
	The fluctuation of these intervals	
	The root mean square of the differences between consecutive R-R intervals and is suitable for assessing the variability of the heart rate over short intervals	
	The percentage of R-R intervals whose difference is greater than 50 milliseconds and indicates the activity of the parasympathetic branch of the nervous system	

Table 1. ECG signal analysis features.

continuous variables, while the chi-squared test was used for categorical variables. The study participants' baseline characteristics were compared between those with and without CVD using these tests. The event date was established as the date of the CVD incident. Individuals who met the following criteria were excluded: leaving the residential area, deaths not related to CVD, loss of follow-up, or end of follow-up.

The univariate cox regression model explored the relationship between potential features and CVD incidence. The univariate model's significance threshold for feature selection was set at an alpha level of 0.2. These features were then entered into the forward stepwise Cox regression model, and important features for CVD incidence were selected in both men and women.

We used the FRS for the analysis of CVD outcome. Standard risk factors, such as age, smoking, systolic blood pressure, antihypertensive medication use, total and HDL cholesterol levels, and diabetes, were employed to develop the recalibrated FRS (baseline) model. The baseline Framingham risk score were combined with the features of the ECG signal to generate an improved recalibrated FRS (baseline) model. In survival analysis, Harrell's C statistic is used to measure discrimination performance. The model's performance was evaluated on the data set through Harrell's concordance index (C-index), specifically focusing on its ability to rank subjects by risk. It determined the likelihood that, given a randomly selected pair of individuals who did or did not experience the event of interest, the individual who did experience the event at a given time would have a higher risk score compared to the randomly selected pair who did not experience the event during the same follow-up interval[27]. Bootstrap resampling was used to estimate the 95% confidence intervals for Harrell's C statistics of various models.

The discrimination of the model was determined by comparing the Net Reclassification Index (cNRI), and Integrated Discrimination Improvement Index (IDI) between models. Paraclinical parameters were employed to incorporate absolute and relative IDI, as well as cut-point-based and cut-point-free NRI, into the baseline survival-based regression model as predictive ability measures[28]. Cut-points for NRI were considered as low risk: <5%, borderline risk: 5%-7.5%, intermediate risk: 7.5%-20%, and high risk: ≥20% based on the 10-year ASCVD-PCE score classification. All analyses were conducted via Python, with a two-tailed p-value < 0.05 deemed statistically significant.

Result

A total of 2094 men and 2543 women were followed up. During a 10-year follow-up, out of 4637 non-CVD participants in the TLGS, 483 (10.4%) developed CVD. Table 2 presents the descriptive data regarding research participants, with analyses conducted separately for male and female subjects. A comparison of the baseline characteristics between participants developing CVD and those not developing CVD is illustrated in Table 2. At baseline, subjects with CVD had higher age, systolic and diastolic blood pressure, FPG and 2hPG levels, cholesterol, and LDL compared to those without cardiovascular disease, across both genders. Mean values of the age (57.76 vs. 47 years) were significantly higher in participants developing CVD than those not developing CVD in men. Mean values of the age (58.67 vs. 45.69 years) were significantly higher in participants developing CVD than those not developing CVD in women. No difference was observed between participants developing CVD and those not developing CVD in the mean HDL in men. There was no significant difference in smokers between participants developing CVD and those not developing CVD, regardless of gender.

In this study, specific electrocardiogram (ECG) features were selected for analysis, with different sets chosen for male and female subjects. For male subjects, three features were selected: the maximum amplitude in lead aVR, which is the highest amplitude recorded in that lead; the root mean square in lead I, a measure of the signal magnitude calculated as the square root of the mean of squared values; and the minimum amplitude in lead V6, the lowest amplitude recorded in that lead. For female subjects, a broader set of thirteen features was utilized. These include the spectral centroid from the Fast Fourier Transform (FFT) in lead aVF, representing the center of mass of the frequency spectrum; kurtosis in lead aVR, a statistical measure of the signal distribution; the mean of wavelet coefficients at level 1 in lead aVR, which is the average of the wavelet decomposition coefficients at the first level; the spectral centroid from FFT in lead I; the standard deviation in lead II, indicating the variability of the signal amplitude; the dominant frequency from FFT in lead II, identifying the most prominent frequency

Variables	Women			Men		
	Non-CVD (*n* = 2353)	CVD (*n* = 190)	P-value	Non-CVD (*n* = 1801)	CVD (*n* = 293)	P-value
Age (year)	45.96 ± 11.26	58.67 ± 9.94	< 0.001	47.00 ± 12.66	57.76 ± 11.69	< 0.001
SBP (mmHg)	112.59 ± 17.79	130.90 ± 20.95	< 0.001	117.77 ± 16.84	127.81 ± 20.06	< 0.001
DBP (mmHg)	73.09 ± 10.18	78.53 ± 11.61	< 0.001	76.05 ± 10.11	79.15 ± 11.49	< 0.001
TC (mg/dl)	196.94 ± 39.41	218.26 ± 41.85	< 0.001	190.92 ± 36.46	200.95 ± 38.05	< 0.001
HDL-C (mg/dl)	45.00 ± 10.52	42.03 ± 9.49	< 0.001	37.64 ± 8.53	37.94 ± 8.50	0.578
LDL-C (mg/dl)	121.69 ± 32.83	136.87 ± 35.63	< 0.001	119.67 ± 31.55	127.35 ± 34.40	< 0.001
FPG (mg/dl)	95.84 ± 28.89	121.58 ± 52.27	< 0.001	96.95 ± 28.31	107.72 ± 41.08	< 0.001
2-h PG (mg/dl)	110.92 ± 44.60	134.98 ± 56.52	< 0.001	107.98 ± 52.80	120.41 ± 58.55	0.001
DM	260 (11.3)	67 (37.2)	< 0.001	165 (9.5)	57 (20.2)	< 0.001
DM medication	139 (5.9)	47 (24.7)	< 0.001	72 (4.0)	29 (9.9)	< 0.001
HTN	335 (14.5)	83 (45.1)	< 0.001	285 (16.1)	92 (31.9)	< 0.001
HTN medication	111 (4.7)	29 (15.3)	< 0.001	46 (2.6)	20 (6.8)	< 0.001
Current smoker	65 (2.8)	6 (3.3)	0.708	449 (25.4)	76 (26.4)	0.732

Table 2. Baseline characteristics of participants. SBP, systolic blood pressure; DBP, diastolic blood pressure; TC, Total Cholesterol; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; FPG, fasting plasma glucose; 2-hPG, 2-hour post-challenge plasma glucose; DM, diabetes melitus; HTN, Hypertension. Data are given as the mean ± SD for continuous variables and data are given as the n(%) for categorical variables.

component; the minimum amplitude in lead V3; the standard deviation of the PR interval in lead V3, reflecting the variability of the time from the P wave's start to the QRS complex's start; the spectral centroid from FFT in lead V4; the maximum R-peak amplitude in lead V4, the highest amplitude of the QRS complex's peak; kurtosis in lead V5; the mean QT interval in lead V5, the average duration from the Q wave's start to the T wave's end; and the energy of wavelet coefficients at level 1 in lead V6, representing the energy content of the wavelet decomposition coefficients at the first level.

Harrell's C index of discrimination can provide helpful information on the predictive performance of a predictive model. As illustrated in Table 3, Harrell's C index for models with and without ECG signal features in men were 0.77 (CI: 0.75–0.79) and 0.77 (95% CI: 0.74–0.79), respectively. There was a slight difference in the goodness of fit as indicated by AIC between these two risk algorithms (AIC: 3896 vs. 3899) in men. Also, the Harrell's C index for models with and without ECG signal features in women was 0.85 (CI: 0.83–0.88) and 0.84 (95% CI: 0.81–0.86), respectively (Fig. 2). A significant difference in goodness of fit, as measured by AIC, was observed between the two risk algorithms in women (AIC: 2375 vs. 2395) in women. Assessing the clinical significance of a new risk biomarker involves analyzing the predictive capability of an existing predictive model enhanced by the addition of new biomarker(s). Generally, the introduction of ECG signal features to the Framingham risk score in women significantly improved risk classification as indicated by cut point-free NRI of 55.7% (95% CIs: 46.5–65.0%), absolute IDI of 2.8% (95% CIs: 1.0–4.6%) but the addition of ECG signal features to the Framingham risk score in men significantly didn't improve risk classification.

Table 4 presents the improvement in the reclassification of people across risk categories after complementing Framingham Risk Score (FRS) with ECG signal features. In each of the four FRS categories (i.e., 0–4.9%, 5–7.4%, 7.5–19.9%, and ≥ 20%), 18.5%, 9.1%, 10.3%, and − 16.1% of women were correctly reclassified, respectively. In each of the four FRS categories (i.e., 0–4.9%, 5–7.4%, 7.5–19.9%, and ≥ 20%), 0.0%, 5.9%, 1.7%, and − 3.0% of men were correctly reclassified, respectively (Fig. 3). Participants without ECGs were excluded, which could provide selection bias. Supplementary Table S1 compares included and excluded participants by gender. Although some of the differences are statistically significant, the study's findings are unlikely to be impacted by their clinically negligible amounts.

Discussion

This study represents the first evaluation of electrocardiogram (ECG) signal features within the Tehran Lipid and Glucose Study (TLGS) for predicting cardiovascular disease (CVD) risk. Our findings demonstrate that ECG signal features significantly enhance the predictive capacity for CVD in women, both statistically and clinically. These sex-specific results align with emerging evidence of differential immune-metabolic regulation between males and females, as described by Pei et al. (2024), which may influence cardiac electrophysiology and ECG-based prediction models[29]. Mahdavi et al. evaluated the American Heart Association (AHA) risk score classification in TLGS participants, noting its effectiveness in identifying high-risk individuals but limited ability to accurately distinguish those with severe cardiovascular outcomes[30]. This suggests a need for refined risk stratification approaches to improve predictive accuracy.

Khalili et al. previously highlighted the predictive utility of abnormal resting ECGs compared to Rose Questionnaire angina in estimating 10-year coronary heart disease (CHD) risk in an urban Iranian population. Their study categorized participants into four groups based on Rose Angina and ECG ischemia status, finding that adding abnormal ECG findings to angina did not significantly increase CHD event risk prediction. However, their analysis relied on the Minnesota Coding (MC) system, a standardized ECG classification method widely

	Basic model	Enhanced model
Women		
Harrell's C index (95% CIs)	0.84 (0.81,0.86)	0.85 (0.83,0.88)
Akaike information criterion	2395	2375
Added predictive values		
Absolute IDI (95% CIs)	0.0280 (0.0100, 0.0460)	P-value = 0.002
Relative IDI (95% CIs)	0.2139 (− 0.0421, 0.4699)	P-value = 0.102
Cutpoint-based NRI (95% CIs)	0.1541 (0.0416, 0.2665)	P-value = 0.007
Cutpoint-free NRI (95% CIs)	0.5575 (0.4653, 0.6496)	P-value < 0.001
Men		
Harrell's C index (95% CIs)	0.77 (0.74,0.79)	0.77 (0.75,0.79)
Akaike information criterion	3899	3896
Added predictive values		
Absolute IDI (95% CIs)	0.0030 (− 0.0022, 0.0081)	P-value = 0.261
Relative IDI (95% CIs)	0.0242 (− 0.0501, 0.0986)	P-value = 0.523
Cutpoint-based NRI (95% CIs)	0.0154 (− 0.0601, 0.0908)	P-value = 0.690
Cutpoint-free NRI (95% CIs)	0.1138 (− 0.0322, 0.2598)	P-value = 0.127

Table 3. Predictive performances of the basic framingham's "general CVD risk" algorithm vs. enhanced model. Akaike information criterion (AIC) was used as a measure of model fit. The lower is the AIC the better will be the model fitness. Difference in AIC > 10 was considered significant.

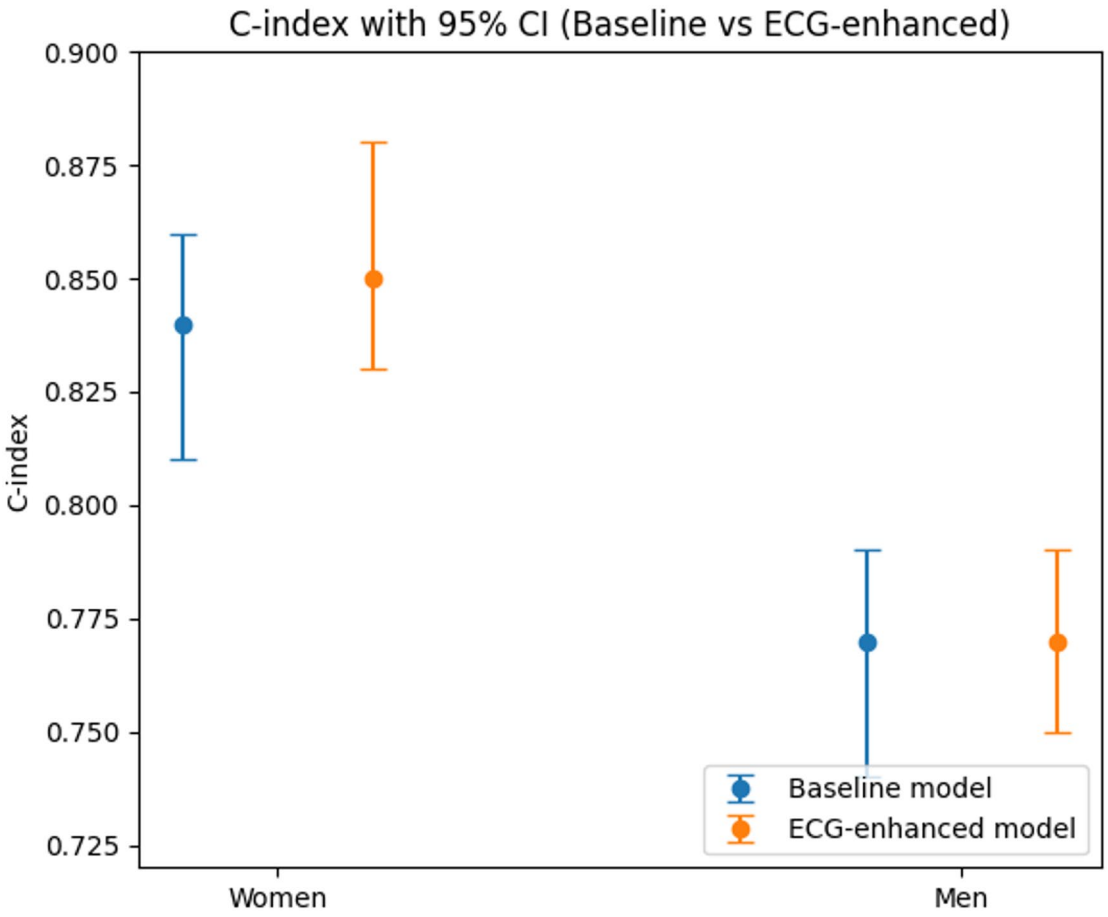

Fig. 2. Forest plot of C-index with 95% confidence intervals for men and women for the baseline model versus the ECG-enhanced model.

used in epidemiological studies, rather than raw ECG signal data[31]. The MC system, introduced in 1960 and expanded in 1983 to include serial comparisons, involves complex measurement protocols that make visual coding time-consuming and prone to errors. In contrast, the current study leverages ECG signal feature extraction to minimize measurement and coding errors, offering a more robust approach to risk prediction[26].

Most population-based studies employ the MC system for ECG coding, which can be performed manually or through automated methods. Both approaches, however, are susceptible to errors, and neither manual coding by a single individual nor automated techniques can be considered fully reliable[27]. Hadaegh et al. demonstrated the additional value of ECG abnormalities beyond the FRS for CHD risk stratification in Middle Eastern women. Their findings indicated that incorporating ECG abnormalities—specifically ST depression or T-wave changes—into the FRS did not significantly improve C-statistics but enhanced predictive performance by 20.8% (95% CI 5.0–38.9) using the cut-point-free Net Reclassification Improvement (NRI). Notably, among women, ECG abnormalities were independently linked to intensified CHD risk only within the intermediate risk category; however, their investigation used MC for ECG evaluation[28].

The reliance on manual ECG pattern recognition and MC in population-based and clinical studies is increasingly being questioned, as these methods are labor-intensive and error-prone[24]. ECG signal processing, as applied in the current study, offers a promising alternative. By extracting signal features directly from ECG data, this approach reduces errors associated with visual coding and enhances the precision of CVD risk prediction. This study underscores the potential of ECG signal analysis to refine risk stratification and improve outcomes in epidemiological research, particularly for women in all risk groups.

Although this study is focused on adult patients, signal-based ECG analysis could potentially benefit rare pediatric cardiac cases. Zhang et al. (2024) report a successful implantable cardioverter defibrillator intervention in a child with Timothy syndrome, underlining the importance of accurate ECG interpretation in complex congenital diseases[32]. While this this finding is outside the primary scope of our integrative CVD risk modeling, novel hardware developments may provide additional advantages. One such innovation is the use of phase-modulated pump light combined with external Gaussian noise to improve the sensitivity of SERF magnetometers (Ma et al., 2024). This approach could significantly enhance ECG capture by increasing the signal-to-noise ratio and enabling the detection of low-amplitude components. As a result, it may increase the number of ECG-derived features accessible for future modeling attempts[33].

	Model with ECG signal features				Reclassified		
	< %5	%5–%7.5	%7.5–%20	≥%20	Increased risk	Decreased risk	Net correctly reclassified %
Women							
Event							
0–5%	22	2	3	0	5	0	18.5
5–7.5%	4	2	5	0	5	4	9.1
7.5–20%	4	9	44	21	21	13	10.3
≥20%	0	0	9	47	0	9	– 16.1
Non-evnet							
0–5%	1464	70	16	1	87	0	5.6
5–7.5%	82	79	53	1	54	82	– 13.0
7.5–20%	34	78	210	31	31	112	– 22.9
≥20%	1	1	46	90	0	48	– 34.8
Men							
Event							
0–5%	7	0	0	0	0	0	0.0
5–7.5%	2	12	3	0	3	2	5.9
7.5–20%	0	3	110	5	5	3	1.7
≥20%	0	0	4	130	0	4	– 3.0
Non-evnet							
0–5%	402	41	0	0	41	0	9.3
5–7.5%	57	283	41	0	41	57	– 4.2
7.5–20%	1	41	530	21	21	42	– 3.5
≥20%	0	0	22	267	0	22	– 7.6

Table 4. Reclassification table comparing risk strata for models incorporating CVD risk factors with and without ECG signal features. ACC/AHA, American College of Cardiology/American Heart Association.

Although our work focuses on improving CVD risk prediction using ECG-derived features, developing non-invasive treatments strategies are also worth considering. A recent meta-analysis identified Shenfu injection as a potential treatment for bradyarrhythmia (Wei et al., 2025). Improved risk stratification might assist identify patients most likely to benefit from such medicines, emphasizing the relevance of integrating prediction models to not only prognosis but also treatment selection and outcomes[34]. Similarly to our integration of clinical features and ECG, biosensor approaches such as those developed by Li et al. (2025) provide highly sensitive molecular diagnostics, suggesting a multidisciplinary future for the early detection of CVD[35].

Zhang et al. (2024) identified a protective effect of lncRNA AK083884 via PKM2/HIF-1α–mediated macrophage reprogramming in viral myocarditis, demonstrating the increasing recognition of the function of inflammatory and metabolic signaling in shaping ECG patterns. Our present model does not integrate molecular biomarkers, but future research may benefit from including such mechanistic layers to improve both interpretability and predictive performance[36].

To improve model generalizability, additional variables like inflammatory markers should be incorporated. Chen et al. (2024) found that the neutrophil-lymphocyte ratio predicts all-cause and cardiovascular mortality in individuals with COPD[37]. We acknowledge that these characteristics enhance the predictive power of the model and increase generalizability. Although we acknowledge that these elements improve the model's generalizability and predictive power, they were not assessed in our study and were thus identified as a limitation of our study. While the findings are valuable, a brief comment on the potential need for validation in other ethnic populations would be appropriate. We acknowledging this limitation and suggesting the need for external validation of the model and the selected features in diverse cohorts. Furthermore, missing ECG data is a study limitation that could introduce selection bias.

Conclusions

In this study, the Minnesota Coding (MC) system and select ECG features were not employed; instead, comprehensive ECG signal data were analyzed. This approach significantly enhanced the statistical and clinical prediction of cardiovascular disease (CVD) risk in women, but no such effect was observed in men.

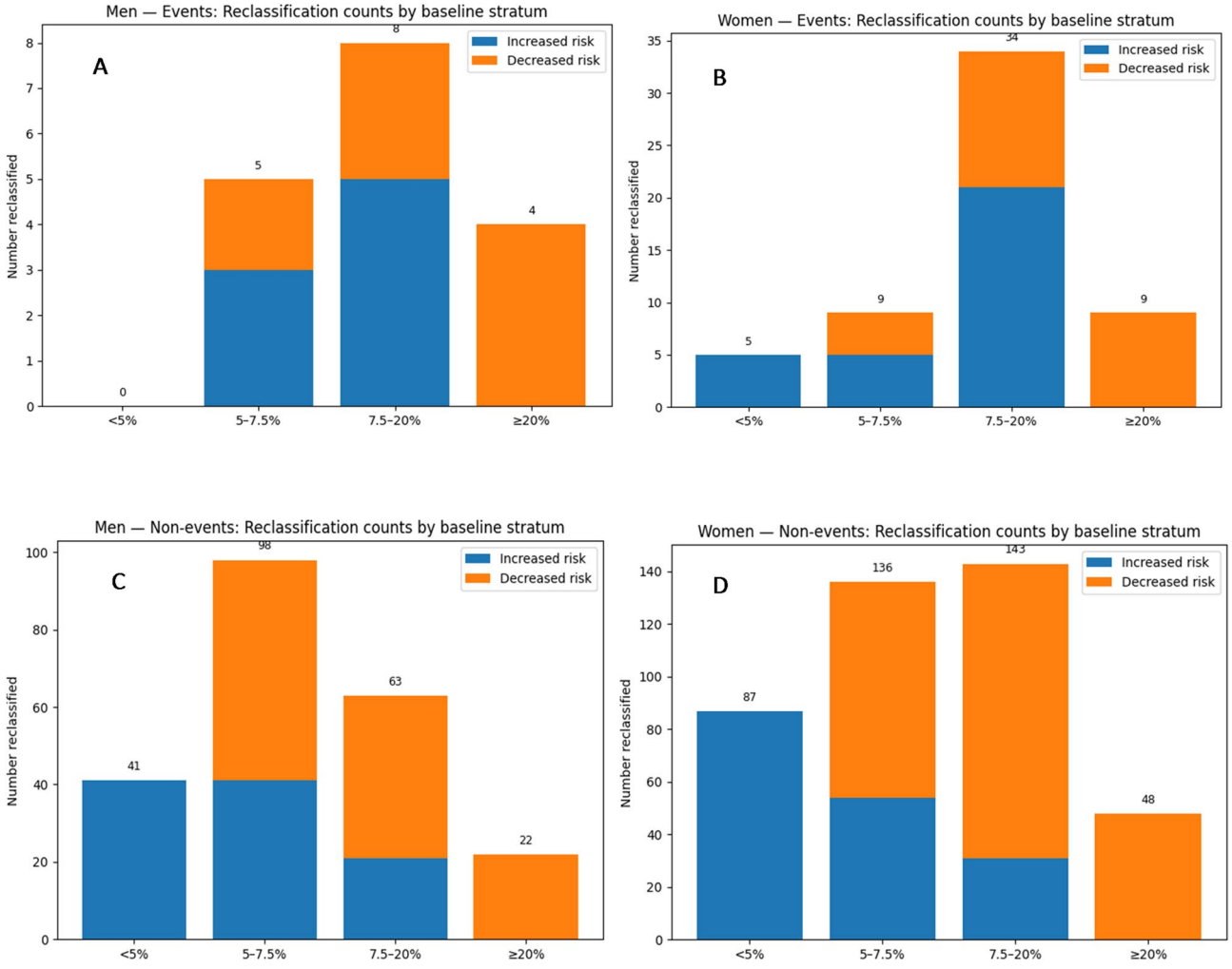

Fig. 3. Risk reclassification counts of CVD after adding ECG features for men and women with events.

Data availability

The datasets used and examined in this investigation could be obtained from the corresponding author upon a reasonable request.

Received: 27 April 2025; Accepted: 29 October 2025
Published online: 07 November 2025

References

1. Taheri Soodejani, M. Non-communicable diseases in the world over the past century: a secondary data analysis. *Front. Public. Health*. **12**, 1436236 (2024).
2. Roth, G. A. et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *J. Am. Coll. Cardiol.* **76** (25), 2982–3021 (2020).
3. Aminorroaya, A. et al. Global, regional, and National quality of care of ischaemic heart disease from 1990 to 2017: a systematic analysis for the global burden of disease study 2017. *Eur. J. Prev. Cardiol.* **29** (2), 371–379 (2022).
4. Organization, W. H. *Noncommunicable Diseases Country Profiles 2018* (2018).
5. Rahmani, A. et al. Investigation of the prevalence of obesity in iran: a systematic review and meta-analysis study. *Acta Med. Iranica* **2015**, 596–607 (2015).
6. Uthman, O. A. Global, regional, and National disability-adjusted life years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2015: a systematic analysis for the global burden of Diseases, Injuries, and risk factors (GBD) 2015 study. *Lancet* **388** (10053), 1603–1658 (2016).
7. Roth, G. A. et al. Global, regional, and National burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J. Am. Coll. Cardiol.* **70** (1), 1–25 (2017).
8. Danaei, G. et al. Iran in transition. *Lancet* **393** (10184), 1984–2005 (2019).
9. Mensah, G. A. & Brown, D. W. An overview of cardiovascular disease burden in the united States. *Health Aff.* **26** (1), 38–48 (2007).
10. Jabir, R. et al. Current updates on therapeutic advances in the management of cardiovascular diseases. *Curr. Pharm. Design.* **22** (5), 566–571 (2016).
11. Guidry, U. C. et al. Temporal trends in event rates after Q-wave myocardial infarction: the Framingham heart study. *Circulation* **100** (20), 2054–2059 (1999).

12. Piepoli, M. F. et al. Guidelines: editor's choice: 2016 European guidelines on cardiovascular disease prevention in clinical practice: the sixth joint task force of the European society of cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of 10 societies and by invited experts) developed with the special contribution of the European association for cardiovascular prevention & rehabilitation (EACPR). *Eur. Heart J.* **37** (29), 2315 (2016).
13. Liu, S. et al. Burden of cardiovascular diseases in China, 1990–2016: findings from the 2016 global burden of disease study. *JAMA Cardiol.* **4** (4), 342–352 (2019).
14. Goff, D. C. Jr et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American college of Cardiology/American heart association task force on practice guidelines. *Circulation* **129** (25_suppl_2), S49–S73 (2014).
15. Faizal, A. S. M., Thevarajah, T. M., Khor, S. M. & Chang, S-W. A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach. *Comput. Methods Programs Biomed.* **207**, 106190 (2021).
16. Azizi, F. et al. Tehran lipid and glucose study (TLGS): rationale and design. *Iran. J. Endocrinol. Metabolism*. **2** (2), 77–86 (2000).
17. Azizi, F. et al. Prevention of non-communicable disease in a population in nutrition transition: Tehran lipid and glucose study phase II. *Trials* **10**, 1–15 (2009).
18. Saatchi, M., Mansournia, M. A., Khalili, D., Daroudi, R. & Yazdani, K. Estimation of generalized impact fraction and population attributable fraction of hypertension based on JNC-IV and 2017 ACC/AHA guidelines for cardiovascular diseases using parametric G-formula: Tehran lipid and glucose study (TLGS). *Risk Manage. Healthc. Policy* **2020**, 1015–1028 (2020).
19. Organization, W. H. Cerebrovascular disorders: a clinical and research classification. In *Cerebrovascular Disorders: A Clinical and Research Classification* (1978).
20. Afsar, F. A., Arif, M. & Yang, J. Detection of ST segment deviation episodes in ECG using KLT with an ensemble neural classifier. *Physiol. Meas.* **29** (7), 747 (2008).
21. Singh, A. K. & Krishnan, S. ECG signal feature extraction trends in methods and applications. *Biomed. Eng. Online*. **22** (1), 22 (2023).
22. Addison, P. S. Wavelet transforms and the ECG: a review. *Physiol. Meas.* **26** (5), R155 (2005).
23. Srinivasulu, A. & Sriraam, N. Signal processing framework for the detection of ventricular ectopic beat episodes. *J. Med. Signals Sens.* **13** (3), 239–251 (2023).
24. Electrophysiology TFotESoCtNASoP. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation* **93** (5), 1043–1065 (1996).
25. Shaffer, F. & Ginsberg, J. P. An overview of heart rate variability metrics and norms. *Front. Public. Health*. **5**, 258 (2017).
26. Rajendran, A. et al. Sex-specific differences in cardiovascular risk factors and implications for cardiovascular disease prevention in women. *Atherosclerosis* **384**, 117269 (2023).
27. Harrell, F. E. Jr, Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15** (4), 361–387 (1996).
28. Pencina, M. J., D'Agostino Sr, R. B., D'Agostino, R. B. Jr & Vasan, R. S. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* **27** (2), 157–172 (2008).
29. Pei, W. et al. Multitargeted Immunomodulatory therapy for viral myocarditis by engineered extracellular vesicles. *ACS Nano*. **18** (4), 2782–2799 (2024).
30. Mahdavi, M., Kazemnejad, A., Asosheh, A. & Khalili, D. Cardiovascular risk patterns through AI-enhanced clustering of longitudinal health data. *J. Diabetes Metabolic Disorders*. **24** (1), 1–10 (2025).
31. Prineas, R. J., Crow, R. S. & Zhang, Z-M. *The Minnesota Code Manual of Electrocardiographic Findings* (Springer Science & Business Media, 2009).
32. Zhang, Z. et al. A case of pioneering subcutaneous implantable cardioverter defibrillator intervention in timothy syndrome. *BMC Pediatr.* **24** (1), 729 (2024).
33. Ma, N. et al. Enhancing the sensitivity of spin-exchange relaxation-free magnetometers using phase-modulated pump light with external Gaussian noise. *Opt. Express*. **32** (19), 33378–33390 (2024).
34. Wei, Y. et al. Efficacy and safety of Shenfu injection on bradyarrhythmia: a systematic review and meta-analysis. *Medicine* **104** (18), e41779 (2025).
35. Li, L. et al. Nanozyme-enhanced tyramine signal amplification probe for preamplification-free myocarditis-related MiRNAs detection. *Chem. Eng. J.* **503**, 158093 (2025).
36. Zhang, Y. et al. M2 macrophage exosome-derived LncRNA AK083884 protects mice from CVB3-induced viral myocarditis through regulating PKM2/HIF-1α axis mediated metabolic reprogramming of macrophages. *Redox Biol.* **69**, 103016 (2023).
37. Chen, Z. et al. The neutrophil-lymphocyte ratio predicts all-cause and cardiovascular mortality among united States adults with COPD: results from NHANES 1999–2018. *Front. Med.* **11**, 1443749 (2024).

Acknowledgements

The data for this research were obtained from the Tehran Lipid and Glucose Study (TLGS), conducted by the Endocrine Research Center at Shahid Beheshti University of Medical Sciences. The authors would like to express their gratitude to everyone involved in the design and data collection of the TLGS, as well as to the study participants. This project was approved by the Ethics Committee of Tarbiat Modares University under the code IR.MODARES.REC.1403.100.

Author contributions

'MM': Data collection, literature review, and manuscript preparation. 'DK': Study design, revising the manuscript, and final approval of the manuscript. 'MM', 'AT', and 'KH': Data analysis and data interpretation. 'AA' and 'AK': Study design, manuscript preparation, revising the manuscript, and the final approval of the manuscript. All authors reviewed and approved the final draft.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Data for this investigation were acquired from the Tehran Lipid and Glucose Study (TLGS), which was conducted by the Endocrine Research Center at Shahid Beheshti University of Medical Sciences. The authors are grateful to the study participants and all those who contributed to the design and data acquisition of the TLGS. The project was authorized by the Ethics Committee of Tarbiat Modares University under the code IR.MODARES.REC.1403.100.

Informed consent

Following the ethical guidelines of the Helsinki Declaration, this study was conducted; all procedures comprising human subjects were approved by the Human Research Review Committee of the Endocrine Research Center, Shahid Beheshti University, Tehran, Iran. Every subject provided written informed consent.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-26471-6.

Correspondence and requests for materials should be addressed to A.K. or A.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.