

A Dynamic Bayesian Network Model for Predicting Cardiovascular Disease Incidence in the Tehran Lipid and Glucose Study

maryam mahdavi

Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University

Anoshirvan Kazemnejad

Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University

Abbas Asosheh

`asosheh@modares.ac.ir`

Department of Medical Informatics, Faculty of Medical Sciences, Tarbiat Modares University

Davood Khalili

Prevention of Metabolic Disorders Research Center, Research Institute for Metabolic and Obesity Disorders, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences

Ahmadreza Tajari

Institute for Biological Information Processes (IBI), Molecular and Cellular Physiology (IBI-1), Forschungszentrum Jülich

Kamyab Hosseinpour

Department of Electrical Engineering, Sharif University of Technology

Research Article

Keywords: Dynamic Bayesian Network, CVD incidence, risk score

Posted Date: September 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10943682/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Cardiovascular disease (CVD) remains a leading cause of morbidity and mortality worldwide, necessitating accurate risk prediction tools for early prevention. Traditional survival models, such as Cox regression, often assume independence among predictors and may not fully capture temporal dependencies or handle censored data efficiently. Dynamic Bayesian Networks (DBNs) offer a probabilistic framework capable of modeling longitudinal health data, managing missing values, and incorporating censoring, making them promising for CVD risk assessment.

Methods

We analyzed data from 7,839 CVD-free participants (3,239 men and 4,600 women) in the Tehran Lipid and Glucose Study (TLGS) with up to 10 years of follow-up. Sociodemographic, behavioral, obesity-related, and cardio-metabolic variables were collected at multiple time points. A predefined sex-specific DBN structure, validated by a clinical expert, was applied to model temporal dependencies and survival outcomes while accounting for censoring.

Results

Over 10 years, 632 participants (8.1%) developed CVD. In men, CVD incidence was 7.2% in the low-risk group and 21.6% in the high-risk group (HR = 3.10, 95% CI: 1.91–5.04; AUC = 0.612). In women, CVD incidence was 6.6% in the low-risk group and 20.0% in the high-risk group (HR = 3.10, 95% CI: 1.58–6.08; AUC = 0.674). The DBN effectively incorporated temporal patterns and censoring, revealing clinically meaningful risk stratification.

Conclusions

Our sex-specific DBN framework demonstrated the ability to integrate longitudinal health data, account for censoring, and identify high-risk individuals with more than a threefold increased hazard of developing CVD. This approach provides a flexible and interpretable alternative to traditional survival models and may enhance preventive strategies when applied to large-scale cohort data. The proposed DBN can serve as a data mining technique in the context of survival data, addressing limitations of existing machine learning algorithms.

Background

Cardiovascular disease (CVD) stands as one of the primary causes of death throughout the world since it resulted in 17.9 million deaths in 2016 which represents 31% of total global deaths (1). The majority of

these fatalities take place in developing and middle-income nations (2). The Organization for Economic Cooperation and Development (OECD) member countries documented 4,791 deaths per 100,000 population from heart attack and ischemic heart disease during 2017 (3). Different CVD subtypes contribute to mortality rates at varying levels across different nations (4). The prevention of CVD depends on early risk assessment for people who need specific dietary and lifestyle or pharmacological interventions (5, 6). Several prediction models have emerged during the past two decades to calculate CVD risk through mathematical combinations of multiple risk factors (7–9).

Traditionally, Cox proportional hazards regression has been widely used for CVD risk estimation; however, it assumes independence among predictors and relies on a pre-specified set of risk factors (10). By contrast, Bayesian networks (BNs) incorporate dependencies among predictors by combining probability theory with graphical modeling (11). Dynamic Bayesian Networks (DBNs) can be applied to longitudinal data to capture temporal dependencies; however, they are not directly compatible with traditional survival analysis frameworks (10).

In survival models, especially those used in human studies, researchers often encounter incomplete data. This is largely due to censoring, which occurs when individuals do not experience the event of interest during the study period. This issue may arise from loss to follow-up, withdrawal, or the study ending before the event occurs (12). In this study, we propose a DBN-based approach for survival data analysis, focusing on the time to first incidence of CVD. Our method integrates the principles of survival analysis into the DBN framework, allowing for robust modeling despite the presence of censored data and complex temporal dependencies.

Longitudinal health data introduce additional challenges, such as correlations between repeated measures, irregular sampling intervals, and missing values. Many traditional machine learning models assume independent and identically distributed data, making them unsuitable for such settings (13). Furthermore, patient-specific variability and uncertainty further complicate the modeling process (14). DBNs offer a flexible probabilistic framework that can capture temporal correlations, manage missing data, and model uncertainties in repeated measurements (10). By incorporating survival analysis into the DBN, our approach enables accurate risk prediction for CVD while accounting for censoring and the timing of first events.

Materials and Methods

Participants

The Tehran Lipid and Glucose Study (TLGS) is a large-scale, long-term prospective cohort study initiated in 1999 with the goal of uncovering risk factors for non-communicable diseases (NCDs) in District 13 of Tehran. The study design has been detailed in previous reports (15). In short, the first phase was a cross-sectional survey that included 15,005 participants aged 3 years and older. To recruit participants, phases one (1999–2001) and two (2002–2005) used a multistage random cluster sampling approach.

Subsequent follow-up phases (three to seven) were conducted at approximately 3.5-year intervals, extending through 2020. From the 9,887 participants aged 20 years and older who were recruited in phase 3 of the TLGS, a total of 1,695 individuals were excluded. The reasons for exclusion were prevalent cardiovascular disease at baseline (n = 833), loss to follow-up (n = 31), having only one measurement during follow-up (n = 580), having only one measurement before the occurrence of cardiovascular disease (n = 140), and two consecutive missing observations before cardiovascular disease (n = 464). The final analytic sample therefore consisted of 7,839 participants, including 3,239 men and 4,600 women. Participants were followed up for incident CVD events until March 2020." (Fig. 1).

Measurements

Eligible participants underwent standardized interviews to collect sociodemographic and medical information. Blood pressure measurements were performed by trained staff using a mercury sphygmomanometer after a 15-minute rest, with systolic and diastolic pressures recorded as the average of two seated readings. Participants fasted for 12–14 hours prior to blood sampling. Samples were promptly analyzed at the TLGS laboratory using Selectra autoanalyzers. Fasting plasma glucose was measured by an enzymatic colorimetric method based on glucose oxidation. Lipid parameters were assessed using commercial enzymatic assays, with HDL-C measured after precipitation of apolipoprotein B-containing lipoproteins. LDL-C was calculated using the Friedewald formula for participants with triglyceride levels below 400 mg/dL (16). Serum creatinine was measured on the day of sampling at the TLGS laboratory using the Jaffe–kinetic method under acceptable quality control conditions, with established assay performance characteristics and manufacturer-defined reference ranges. Further details on the measurement methods can be found in previous studies (17).

Definitions of outcome and terms

Participants were followed annually for CVD events through structured telephone interviews, with reported events verified by physician review of medical records and confirmed by the Cohort Outcome Panel (18). Coronary heart disease (CHD) was defined as the occurrence of myocardial infarction (MI), probable MI, unstable angina, or coronary artery disease (CAD) confirmed by angiography. Stroke was classified according to the World Health Organization (WHO) criteria (19). For the purposes of this study, CVD was defined as the first occurrence of either stroke or CHD, with the corresponding time point recorded as the time-to-event.

Any CVD events in first-grade female family members aged 65 years or lower or first-grade male family members aged 55 years or lower were considered a premature family history of CVD (20). Diabetes mellitus was defined according to American Diabetes Association criteria, including fasting plasma glucose ≥ 126 mg/dL, 2-hour plasma glucose ≥ 200 mg/dL, or use of glucose-lowering medication (21). Hypertension was defined as systolic blood pressure ≥ 140 mmHg, diastolic blood pressure ≥ 90 mmHg, or use of antihypertensive medication (22).

Dyslipidemia was defined as having total cholesterol greater than 200 mg/dL, HDL cholesterol below 40 mg/dL in men and below 50 mg/dL in women, triglycerides above 150 mg/dL, LDL cholesterol above 100 mg/dL, or current use of lipid-lowering medications (23). Smoking status was categorized as current smoker or non-smoker. Information on physical activity was collected using the Persian version of the Lipid Research Clinic (LRC) questionnaire in the first phase (24), and the Modifiable Activity Questionnaire (MAQ) in later follow-ups (25). Participants were considered physically active if they achieved at least 600 metabolic equivalent task (MET) minutes per week (26).

Body mass index (BMI) was calculated as weight (Kilogram) divided by height (meter) squared, with overweight/obesity defined as $BMI \geq 25 \text{ kg/m}^2$. Abdominal obesity was defined using Iranian-specific waist circumference cutoffs ($\geq 89 \text{ cm}$ in men and $\geq 91 \text{ cm}$ in women) (27). Education was categorized as ≤ 12 or > 12 years, age as < 50 or ≥ 50 years, occupation as full-time or not, and marital status as married or living alone. The estimated glomerular filtration rate (eGFR), expressed in ml/min/1.73 m^2 , was calculated using the CKD-EPI Eq. (28). Besides, kidney dysfunction was defined as $eGFR < 60 \text{ mL/min/1.73m}^2$ according to equation.

Analysis Method

Normally distributed continuous variables were summarized as mean $\pm$ SD, and categorical variables as frequencies (%). Baseline characteristics were compared between participants with and without CVD using independent t tests for continuous variables and chi-square tests for categorical variables. The event date was defined as the date of incident CVD. Participants were censored at the time of non-CVD death, loss to follow-up, relocation from the study area, or end of follow-up.

To address missing data for covariates at each study phase, single imputation was performed using the multiple imputation by chained equations (MICE) framework with $M = 1$. This approach is suitable for large datasets and allows simultaneous imputation of different variable types. Missing values were imputed through simple random sampling with replacement from the observed data. Linear regression models were applied for continuous variables and logistic regression for binary variables. While multiple imputation is generally preferred over listwise deletion due to its ability to reduce bias and preserve statistical power, a single imputation ($M = 1$) strategy was adopted in this study for computational efficiency.

Model training and evaluation followed a stratified data split: 20% of the dataset was reserved as an unbiased test set. Stratified sampling ensured preservation of the original class distribution across all subsets. The training set was used for model parameter optimization, the validation set guided hyper-parameter tuning and performance monitoring, and the test set was used exclusively for final model evaluation to assess its generalization ability. Feature selection for male and female participants was informed by clinical expert opinion and existing literature, acknowledging sex-specific differences in CVD risk factors and incidence (29). The DBN models were designed to capture patterns over the 10-year observation window. With missing data imputed, the risk score of CVD were comprehensively assessed.

Model performance was evaluated primarily using the area under the ROC curve (AUC), as accuracy can be misleading in the context of imbalanced data.

The classical BN is not adopted to address time-dependent processes like survival analysis. Therefore, DBN was introduced to extend this process. In this context, time-dependent random variables $(X_t)_{t \geq 1} = (X_{1,t}, \dots, X_{D,t})_{t \geq 1}$ are defined where t is a discrete index time formally called slice. DBN uses Markov property which indicates the future of a stochastic process is independent of its past, given current status or several lags before it. The number of lags determines the order of the Markov process. Survival analysis is a subfield of statistics where the goal is to analyze and model data where the outcome is the time until an event of interest occurs (10).

We adopted a predefined DBN structure informed by the literature and ultimately validated by a clinical expert, prioritizing simplicity in the model design. One of the main challenges in this context is the presence of instances whose event outcomes become unobservable after a certain time point or when some instances do not experience any event during the monitoring period. The incomplete data known as censoring is expected because sample withdrawal or loss of follow-up exceeds the event time of the study period. We handle this so-called censoring by deletion of participants who have two missing consecutive observations. We evaluate the proposed model's performance using both training and test datasets. All analyses were conducted using the programming language Python, and a two-tailed $p < 0.05$ was considered significant.

Results

A total of 3,239 men and 4,600 women were included in the study and followed over a 10-year period. Among the 7,839 participants who were initially free of CVD in the Tehran Lipid and Glucose Study (TLGS), 632 individuals (8.1%) went on to develop CVD. Table 1 summarizes the descriptive statistics of the study participants, with analyses conducted separately for men and women. It provides a comparison of baseline characteristics between individuals who developed CVD during follow-up and those who remained CVD-free.

Table 1
Baseline characteristics of Participants

Variables	Male			Female		
	Non-CVD (n = 2903)	CVD (n = 336)	P-value	Non-CVD (n = 4304)	CVD (n = 296)	P-value
Age group, ≥ 40 year	383(13.2)	139(41.4)	< .001	429(10.0)	152(51.4)	< .001
Low PA	1149(39.6)	135(40.2)	< .832	1490(34.6)	110(37.2)	< .374
FH-CVD	519(17.9)	89(26.5)	< .001	871(20.2)	90(30.4)	< .001
Occupation, employed	2297(79.1)	206(61.3)	< .001	654(15.2)	12(4.1)	< .001
Spouse, married	2203(75.9)	326(97.0)	< .001	3353(77.9)	204(68.9)	< .001
Education, ≤ 12 years	2139(73.7)	285(84.8)	< .001	3534(82.1)	288(97.3)	< .001
Current Smoker	667(23.0)	72(21.4)	.522	118(2.7)	9(3.0)	.761
Obesity	1851(63.8)	249(74.1)	< .001	2940(68.3)	261(88.2)	< .001
Abdominal obesity	1966(67.7)	267(79.5)	< .001	1836(42.7)	233(78.7)	< .001
Dyslipidemia	2614(90.0)	324(96.4)	< .001	3907(90.8)	293(99.0)	< .001
Kidney dysfunction	208(7.2)	63(18.8)	< .001	547(12.7)	112(37.8)	< .001
Hypertension	348(12.0)	95(28.3)	< .001	477(11.1)	114(38.5)	< .001
Diabetes melitus	194(6.7)	61(18.2)	< .001	316(7.3)	97(32.8)	< .001
Abbreviations: FH-CVD, family history of CVD; PA, physical activity. Data are given as the n(%)						

Table 2: Incidence of Cardiovascular Diseases Across Different DBN Risk Groups			
DBN Risk Group	CVD event, n(%)	HR(%95CI)	AUC
Men			0.612
Low risk (n = 514)	37 (7.2)	1 (reference)	
High risk (n = 134)	29 (21.6)	3.10 (1.91,5.04)	
Women			0.674
Low risk (n = 870)	57 (6.6)	1 (reference)	
High risk (n = 50)	10 (20.0)	3.10 (1.58,6.08)	
Abbreviations: HR, hazard ratio; CI, confidence interval; AUC, area under the curve			

Baseline characteristics of participants by CVD status are summarized in Table 1. In both men and women, who developed CVD were significantly older, with a larger proportion aged 40 years or older compared to those who remained CVD-free (men: 41.4% vs. 13.2%; women: 51.4% vs. 10.0%; both $P < 0.001$). A family history of CVD was also more frequently reported among participants who developed the disease (men: 26.5% vs. 17.9%; women: 30.4% vs. 20.2%; both $P < 0.001$). Employment status showed a notable difference: a lower proportion of individuals with CVD were employed (men: 61.3% vs. 79.1%; women: 4.1% vs. 15.2%; both $P < 0.001$). Similarly, educational attainment was lower among those with CVD, with a greater percentage having 12 or fewer years of education (men: 84.8% vs. 73.7%; women: 97.3% vs. 82.1%; both $P < 0.001$).

As shown in Table 1, no significant differences were observed in physical activity levels or current smoking status between individuals with and without CVD, in either men or women. In contrast, obesity and abdominal obesity were significantly more prevalent among participants who developed CVD (men: 74.1% vs. 63.8% and 79.5% vs. 67.7%; women: 88.2% vs. 68.3% and 78.7% vs. 42.7%; all $P < 0.001$). Additionally, cardio-metabolic conditions, including dyslipidemia, kidney dysfunction, hypertension, and diabetes mellitus, were consistently more common in the CVD group across both sexes (all $P < 0.001$).

As shown in Table 2, among men, the incidence of CVD was 7.2% in the low-risk group and increased markedly to 21.6% in the high-risk group. This translated to a significantly elevated hazard ratio (HR = 3.10; 95% CI: 1.91–5.04), with AUC of 0.612. Similarly, in women, the incidence of CVD rose from 6.6% in the low-risk group to 20.0% in the high-risk group, corresponding to an HR of 3.10 (95% CI: 1.58–6.08) and an AUC of 0.674.

Figure 2 illustrates the structure of the DBN. In Fig. 2, the “sociodemographic” block is disaggregated into its individual components, including age group, family history of CVD, education, and occupation. The “behavioral” block is similarly expanded to show its components, namely smoking status and physical activity group. The “obesity” block is divided into obesity based on body mass index and

abdominal obesity. The “cardio-metabolic disorder” block is further detailed to include hypertension, dyslipidemia, diabetes, and kidney dysfunction. This directed acyclic graph (DAG) is presented separately for each sex within the framework of the dynamic Bayesian network.

Discussion

In this study, we built a sex-specific DBN model to predict the 10-year risk of developing CVD in 7,839 adults participating in the TLGS. Our findings revealed that individuals in the high-risk group were more than three times as likely to develop CVD as those in the low-risk group, this held true for both men (HR = 3.10, 95% CI: 1.91–5.04; AUC = 0.612) and women (HR = 3.10, 95% CI: 1.58–6.08; AUC = 0.674). Participants who eventually developed CVD tended to share common baseline characteristics, such as being older, having a family history of CVD, lower educational attainment, unemployment, obesity (general or abdominal), dyslipidemia, hypertension, diabetes, and impaired kidney function. These findings highlight the DBN model's strength in capturing changes over time its potential to help identify those at higher risk in a real-world longitudinal setting.

In this study, we present a DBN model for survival analysis. The proposed DBN can serve as a data mining technique in the context of survival data, addressing limitations of existing machine learning algorithms, which are often complex and rarely account for the censoring characteristic of such data. While a conventional Bayesian network offers a straightforward framework, the DBN extends this by explicitly incorporating censoring information into its inferences. Moreover, Bayesian network tools enable a depth of data exploration not achievable with traditional methods. For instance, regression-based approaches typically focus on the outcome variable while neglecting the interrelationships among covariates (10). In contrast, the graphical representation and conditional probability distributions of a Bayesian network effectively convey rich information in a concise and interpretable manner (30).

In our DBN model, the variables are categorical with two levels. The relatively low AUC can be attributed to the nature of such categorical variables, as they inherently lead to information loss compared to continuous variables. To address this limitation, we propose incorporating Electrocardiogram (ECG) signal data into the DBN model, as well as exploring the integration of continuous data through a mixture DBN or a fully continuous DBN model. This recommendation is motivated by the work of Hughes et al., who developed SEER (the Stanford Estimator of ECG Risk), a deep neural network designed to estimate long-term cardiovascular disease and mortality risk from 12-lead ECG data. Their study demonstrated the utility of this risk score in enhancing 10-year ASCVD risk stratification when used alongside the Pooled Cohort Equations (PCE) (31).

Yang (2020) proposed a DBN-based framework integrated with IoT-enabled medical data acquisition to classify medical multimedia big data and assess health status using a multivariate Gaussian distribution model, demonstrating higher sensitivity and stability than ANN and comparable performance to CNN on the MIAS dataset and physiological signals. While the approach effectively extracts high-level features from raw data, its validation on relatively small datasets limits generalizability, underscoring the need for

large-scale, multi-institutional evaluation and exploration of additional DBN architectures (32). In our work, we hypothesize that incorporating ECG signals into the model could enhance discrimination of CVD incidence. Our study applies a DBN model to a large, longitudinal cohort TLGS to predict the first incidence of CVD, explicitly incorporating censoring into the modeling process and capturing temporal dependencies in repeated measurements.

Sheidaei et al. (10) proposed a two-slice temporal Bayesian network model for survival data, in which survival and censoring status at each observed time point were represented as dynamic states. Their approach considered only categorical variables at baseline and the censoring status over time. In contrast, our DBN model accounts for the longitudinal nature of the data and incorporates continuous learning up to the occurrence of the first event. The DAG of the BN model is the only mechanism for demonstrating the intra-relationship of covariates which is not available in the alternative approaches. For instance, interventions could be designed basis on the roots of the network or the parents of unchangeable nodes. Finally, our DBN forced the covariates to be discretized. Our BN has two primary assumptions. At first, the variables follow the Markov property, and then conditional multinomial or conditional binomial distribution is appropriate for discrete nodes. Both of these assumptions are logical in practice.

In the present study, it is important to acknowledge that survival distributions in longitudinal health data are often skewed, and the common assumption of normality for all continuous variables in a DBN may not fully capture their underlying characteristics. This issue becomes more pronounced when the number of complete cases is limited, as was evident in the work of Marshall et al. (2010), where reducing the dataset to only 94 complete records likely altered the distribution of survival times. Under such conditions, a Gaussian network structure may inadequately represent survival dynamics, particularly for extreme survival values. Previous research has demonstrated that when modeling survival within a continuous BN framework, alternative distributions such as the Coxian phase-type distribution can more effectively accommodate skewness and uncover latent survival subgroups (32). Applying such distributional flexibility within a DBN framework could enhance predictive accuracy and robustness, particularly when modeling time-to-event outcomes such as the first incidence of CVD.

A major strength of this study is its foundation in TLGS, a large, well-characterized, long-term prospective cohort with standardized data collection protocols. This design enabled robust modeling of CVD incidence over a 10-year follow-up. The application of a DBN framework allowed for the explicit modeling of temporal dependencies and censoring, factors often neglected in traditional machine learning and survival analysis approaches. Moreover, stratifying analyses by sex and incorporating expert-informed model structures contributed to the clinical relevance and interpretability of the results.

Nonetheless, several limitations should be acknowledged. First, converting all covariates into categorical variables may have resulted in information loss, potentially contributing to the relatively modest AUC values observed. Second, the current DBN model does not yet include continuous high-resolution clinical data, such as ECG waveforms, which could enhance predictive accuracy. Lastly, since the study was

based on a single urban Iranian population, external validation in more diverse populations and healthcare settings is needed to assess generalizability.

Conclusions

In this study, our sex-specific DBN model, applied to 7,839 CVD-free adults from the TLGS demonstrated effective prediction of first-time CVD events over a 10-year follow-up period. Individuals classified as high-risk exhibited more than a threefold increase in CVD hazard in both men and women. Moreover, significant baseline differences in sociodemographic, lifestyle, and cardio-metabolic characteristics were observed between those who developed CVD and those who did not. These findings highlight the DBN model's capability to integrate longitudinal data, account for censoring, and deliver clinically meaningful risk stratification within a large, population-based cohort.

Abbreviations

The following abbreviations are used in this manuscript:

Cardiovascular disease: CVD

Organization for Economic Cooperation and Development: OECD

Bayesian network: BN

Dynamic Bayesian Network: DBN

Tehran Lipid and Glucose Study: TLGS

Non-communicable disease: NCD

Total cholesterol: TC

High-density lipoprotein cholesterol: HDL-C

Triglyceride: TG

low-density lipoprotein cholesterol: LDL-C

Coronary heart disease: CHD

Myocardial infarction: MI

Coronary artery disease: CAD

World Health Organization: WHO

Lipid Research Clinic: LRC

Modifiable Activity Questionnaire: MAQ

Metabolic equivalent task: MET

Body mass index: BMI

Glomerular filtration rate: GFR

Standard deviation: SD

Multiple imputation by chained equations: MICE

Area under the ROC curve: AUC

Electrocardiogram: ECG

Stanford Estimator of ECG Risk: SEER

Pooled Cohort Equations: PCE

Artificial Neural Network: ANN

Convolutional Neural Networks: CNN

Mammographic Image Analysis Society: MIAS

Directed Acyclic Graph: DAG

Declarations

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 permission.

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.

Consent for publication: Not applicable.

Clinical trial number: Not applicable.

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

Competing interest: The authors declare that they own no conflicting interests, either financial or non-financial, that may affect the integrity of this study.

Funding: None

Author's 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. The final draft was evaluated and approved by all authors.

Acknowledgment: The data for this study were sourced from the Tehran Lipid and Glucose Study (TLGS), executed by the Endocrine Research Center at Shahid Beheshti University of Medical Sciences. The authors like to present their appreciation to all those engaged in the design and data gathering of the TLGS, along with the research participants. The Ethics Committee of Tarbiat Modares University approved this study with the code IR.MODARES.REC.1403.100.

References

1. Reddy KS, Mathur MR. Global Burden of CVD: Prevalence, Pattern, and Trends. Handbook of Global Health. Springer; 2021. pp. 423–37.
2. Tabutin D, Masquelier B, Grieve M, Reeve P. Mortality inequalities and trends in low-and middle-income countries, 1990–2015. *Population*. 2017;72(2):221–96.
3. Shevatekar GK. Using the CIPP model of evaluation on a health disparities curriculum delivered to medical residents in 12 residency programs. 2023.
4. Ezzati M, Obermeyer Z, Tzoulaki I, Mayosi BM, Elliott P, Leon DA. Contributions of risk factors and medical care to cardiovascular mortality trends. *Nat Reviews Cardiol*. 2015;12(9):508–30.
5. Haskell WL. Cardiovascular disease prevention and lifestyle interventions: effectiveness and efficacy. *J Cardiovasc Nurs*. 2003;18(4):245–55.
6. Narkhede M, Pardeshi A, Bhagat R, Dharme G. Review on emerging therapeutic strategies for managing cardiovascular disease. *Curr Cardiol Rev*. 2024;20(4):86–100.
7. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353.
8. Stevens D, Lane DA, Harrison SL, Lip GY, Kolamunnage-Dona R. Modelling of longitudinal data to predict cardiovascular disease risk: a methodological review. *BMC Med Res Methodol*. 2021;21(1):283.
9. Rahim A, Rasheed Y, Azam F, Anwar MW, Rahim MA, Muzaffar AW. An integrated machine learning framework for effective prediction of cardiovascular diseases. *IEEE access*. 2021;9:106575–88.

10. Sheidaei A, Foroushani AR, Gohari K, Zeraati H. A novel dynamic Bayesian network approach for data mining and survival data analysis. *BMC Med Inf Decis Mak.* 2022;22(1):251.
11. Marcot BG, Penman TD. Advances in Bayesian network modelling: Integration of modelling technologies. *Environ Model Softw.* 2019;111:386–93.
12. Barkhordari M, Padyab M, Sardarinia M, Hadaegh F, Azizi F, Bozorgmanesh M. Survival regression modeling strategies in CVD prediction. *Int J Endocrinol metabolism.* 2016;14(2):e32156.
13. Cascarano A, Mur-Petit J, Hernandez-Gonzalez J, Camacho M, de Toro Eadie N, Gkontra P, et al. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. *Artif Intell Rev.* 2023;56(Suppl 2):1711–71.
14. Gray RA, Pathmanathan P. Patient-specific cardiovascular computational modeling: diversity of personalization and challenges. *J Cardiovasc Transl Res.* 2018;11(2):80–8.
15. Azizi F, Madjid M, Rahmani M, Emami H, Mirmiran P, Hadjipour R. Tehran Lipid and Glucose Study (TLGS): rationale and design. *Iran J Endocrinol metabolism.* 2000;2(2):77–86.
16. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem.* 1972;18(6):499–502.
17. Azizi F, Madjid M, Rahmani M, Emami H, MIRMIRAN P, Hadjipour R. Tehran Lipid and Glucose Study (TLGS): rationale and design. 2000.
18. Saatchi M, Mansournia MA, 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.* 2020:1015–28.
19. Organization WH. Cerebrovascular disorders: a clinical and research classification. *Cerebrovascular disorders: a clinical and research classification* 1978.
20. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'agostino RB, Gibbons R, 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. *J Am Coll Cardiol.* 2014;63:2935–59. (25 Part B).
21. Association AD. Diagnosis and classification of diabetes mellitus. *Diabetes Care.* 2010;33(Supplement1):S62–9.
22. James PA, Oparil S, Carter BL, Cushman WC, Dennison-Himmelfarb C, Handler J, et al. 2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the Eighth Joint National Committee (JNC 8). *JAMA.* 2014;311(5):507–20.
23. Brethauer SA, Kim J, El Chaar M, Papasavas P, Eisenberg D, Rogers A, et al. Standardized outcomes reporting in metabolic and bariatric surgery. *Obes Surg.* 2015;25:587–606.
24. Ainsworth BE, Jacobs DR Jr, Leon AS. Validity and reliability of self-reported physical activity status: the Lipid Research Clinics questionnaire. *Med Sci Sports Exerc.* 1993;25(1):92–8.

25. Azizi F, Ghanbarian A, Momenan AA, Hadaegh F, Mirmiran P, Hedayati M, et al. Prevention of non-communicable disease in a population in nutrition transition: Tehran Lipid and Glucose Study phase II. *Trials*. 2009;10(1):5.
26. Committee IR. Guidelines for data processing and analysis of the International Physical Activity Questionnaire (IPAQ)-short and long forms. <http://www.ipaq.ki.se/scoring.pdf>; 2005. <http://www.ipaq.ki.se/scoring.pdf>.
27. Delavari A, Forouzanfar MH, Alikhani S, Sharifian A, Kelishadi R. First nationwide study of the prevalence of the metabolic syndrome and optimal cutoff points of waist circumference in the Middle East: the national survey of risk factors for noncommunicable diseases of Iran. *Diabetes Care*. 2009;32(6):1092–7.
28. Levey AS, Stevens LA, Schmid CH, Zhang Y, Castro AF III, Feldman HI, et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med*. 2009;150(9):604–12.
29. Rajendran A, Minhas AS, Kazzi B, Varma B, Choi E, Thakkar A, et al. Sex-specific differences in cardiovascular risk factors and implications for cardiovascular disease prevention in women. *Atherosclerosis*. 2023;384:117269.
30. Shiguihara P, Lopes ADA, Mauricio D. Dynamic Bayesian network modeling, learning, and inference: A survey. *IEEE Access*. 2021;9:117639–48.
31. Hughes JW, Tooley J, Torres Soto J, Ostropolets A, Poterucha T, Christensen MK, et al. A deep learning-based electrocardiogram risk score for long term cardiovascular death and disease. *NPJ Digit Med*. 2023;6(1):169.
32. Yang Y. Medical multimedia big data analysis modeling based on DBN algorithm. *IEEE Access*. 2020;8:16350–61.

Figures

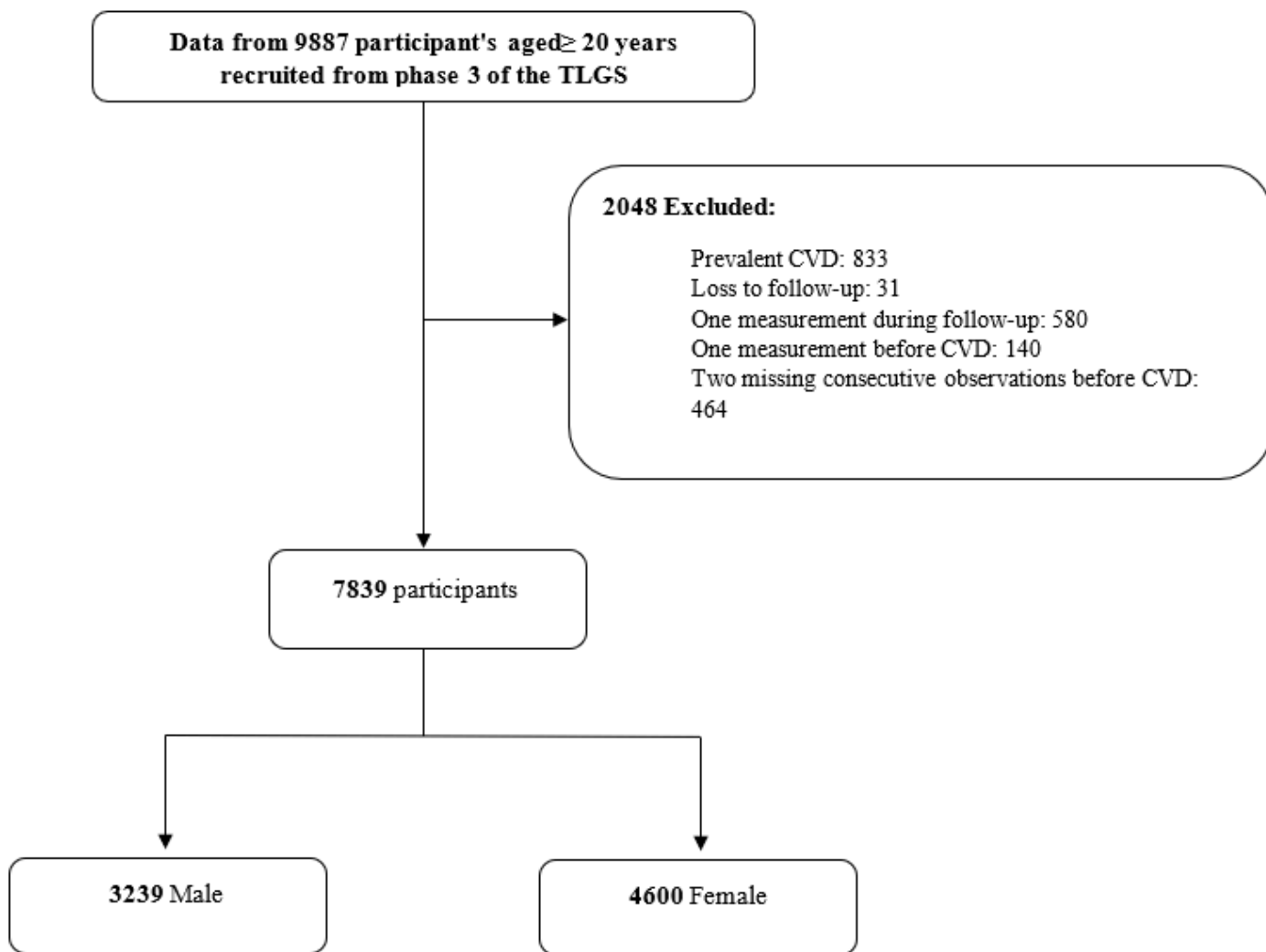


Figure 1

Figure 1

Study flowchart. Tehran Lipid and Glucose Study; TLGS

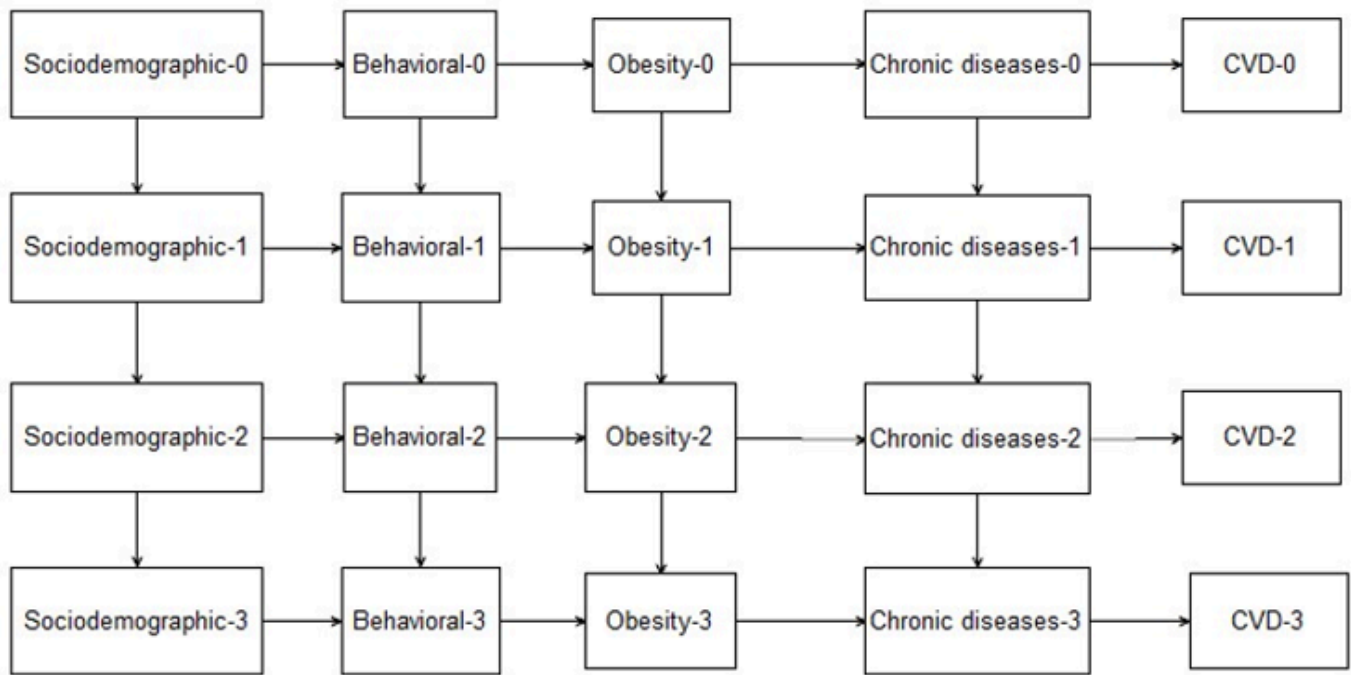


Figure 2

Figure 2

Representation of DAG corresponds to DBN of TLGS participants separated by sex. Sociodemographic factors including: spouse, education, age group, family history of CVD, occupation. Behavioral factors including: smoking and physical activity. Obesity including general and central adiposity. Cardio-metabolic Disorders including: diabetes mellitus, kidney dysfunction, hypertension, and dyslipidemia. In the figure, variables like obesity-0 to obesity-3 and so on denote measurements taken at multiple time points, with the numerical suffix indicating the time of assessment.