

CardioDB: A Relational Database and Knowledge Graph of Cardiovascular Disease

Tapan Kumar BEHERA¹, Siddhartha SATHIA², and Pradeep Kumar NAIK^{1,*}

¹ Research Centre of Excellence in Natural Products and Therapeutics, Department of Biotechnology and Bioinformatics, Sambalpur University, Jyoti Vihar, Burla, Sambalpur-768019, Odisha, India.

² Department of Cardiothoracic Surgery (CTVS), All India Institute of Medical Sciences, Sijua, Patrapada, Bhubaneswar, 751019 Odisha, India.

E-mail(*): pknaik1973@gmail.com

* Author to whom correspondence should be addressed; Tel.: +91 9479268802

Received: 6 November 2025 / Accepted: 23 September 2026 / Published online: 30 September 2026

Abstract

Background: Cardiovascular disease (CVD) poses a significant and persistent challenge to global public health. Worldwide, heart attacks and strokes account for 80% of cardiovascular disease fatalities, many of which can be averted with proactive lifestyle modifications and timely medical interventions. Given the significant mortality rates, the clinical health sector is producing substantial volumes of data. These healthcare sectors face significant challenges in managing multidimensional data, encompassing everything from handwritten notes to high-resolution images to medical images. Consequently, an effective database is required to address the challenges of data storage and accessibility. **Method:** This database consists clinical records of patients with CVD, with the highest number of feature parameters recorded. In addition, a CVD calculator was developed to analyze user risk based on 13 major clinical parameters, and the risk was measured via the weighted assisted scoring (WAS) function. **Results:** The database features a user-friendly web interface that allows users to easily search, navigate, and discover knowledge. This database was designed to increase accessibility, support researchers in addressing high-end feature profiles, improve treatment decision-making, enable time-series pattern analysis, and provide health professionals and researchers with specific reports with current patient information to refine future diagnostic accuracy.

Keywords: Cardiovascular database (CardioDB); Cardiovascular disease (CVD); Structure query language (SQL); Electrocardiogram (ECG); Electrocardiography (ECHO)

Introduction

Cardiovascular disease (CVD) is a leading cause of death worldwide, accounting for millions of deaths annually. The total number of CVD deaths has increased by 60%, from 12.1 million in 1990 to 20.5 million in 2021, driven by an aging population and rising risk factors such as obesity and diabetes [1]. Several well-established risk factors for CVD exist, including smoking, poor diet, physical inactivity, excessive alcohol consumption, high blood pressure, high cholesterol, diabetes, and obesity [2, 3]. Cardiovascular disease does not affect all populations equally. Addressing health disparities is a critical aspect of public health. The costs associated with hospitalization, medications, and productivity loss due to CVD are significant, leading to increased healthcare expenditures. Detecting CVD at an early stage allows early prevention of complications, is more cost-efficient than treating advanced disease, and reduces long-term complications, and hence reduces mortality.

Sex differences in coronary artery disease (CAD) are well documented but are complex. A study by Wan et al. (2014) of 1,961 women and 8,593 men from the Malaysian NCVDPCI Registry (2007–2009) revealed that women

undergoing percutaneous coronary intervention (PCI) were older and had higher rates of diabetes, hypertension, and severe heart failure, whereas men were more likely to smoke and have previously experienced myocardial infarction [Error! Bookmark not defined.]. Women had more left main stem disease, smaller vessels, and longer door-to-balloon times, contributing to higher in-hospital mortality across PCI, ST-segment elevation myocardial infarction (STEMI), and non-STEMI patients (odds ratios: 2.06–6.18). The adjusted six-month mortality differences were not significant, suggesting that comorbidities and delays drive outcomes, which is consistent with previous findings [4]. However, another author reported no mortality gap after post-adjustment, highlighting variability [5]. Biological and access-to-care factors, necessitate tailored strategies for women to improve CAD outcomes [6].

Microbe-induced cardiovascular diseases (CVDs), including endocarditis, myocarditis, and pericarditis, are understudied despite their clinical significance. The MorCVD database addresses this gap by consolidating 23,377 host–pathogen interactions (HPIs) between humans and 432 unique microbial proteins linked to 19 CVDs, sourced from scattered molecular interaction databases [Error! Bookmark not defined.]. MorCVD is accessible via a user-friendly web interface (<http://morcvd.sblab-nsit.net>) and enriches HPI data with hyperlinked protein IDs, PubMed references, gene ontology annotations, drug target details, and interactor information across species. The database supports searches by disease, protein ID, pathogen, or interaction detection methods, with results exportable in a tabular format. Analysis revealed that viruses dominate microbial CVD interactions more than bacteria, highlighting the need for further bacterial HPI research, as noted by MorCVD (2023). The comprehensive documentation and 116 detection method descriptions of MorCVD enhance its utility, building on earlier efforts, such as earlier effects to catalogue microbial interactions [7]. This unified resource facilitates molecular studies of microbe-induced CVDs, offering insights into their biochemical properties and potential therapeutic targets for CVD treatment.

IndiaBioDB represents a significant effort to catalogue and centralize the growing contributions of Indian researchers in the global bioinformatics landscape. As a manually curated repository, it compiles detailed information on over 550 freely accessible bioinformatics resources (of which 263 are biological databases) across academic and research institutions in India [8]. Each entry includes essential metadata, such as resource names, web links, publication details, corresponding authors, host institutes, funding agencies, and resource types, thereby offering a comprehensive snapshot of India's bioinformatics infrastructure. The platform enhances accessibility through a user-friendly search interface and browsing functionality, and extends its utility via a dedicated mobile application. By including only functional resources, IndiaBioDB ensures practical relevance for wet-lab scientists and computational researchers, enabling efficient experimental design, cost-effective workflows, and opportunities for methodological innovation through data reuse. This initiative not only underscores India's emergence as a key player in bioinformatics, ranked 4th globally in database development, but also serves as a vital tool for mapping national research capacity and fostering future collaborations.

Data of patients with CVDs routinely captured in a variety of formats, including clinical notes, structured test results, and diagnostic images such as ECG and ECHO. Because these records contain critical insights for longitudinal patient care, the systematic organization and storage of this information are vital for identifying clinical patterns that improve future treatment decisions. To meet this need, this study aims to design, build, and evaluate a comprehensive database system tailored for the secure management and analysis of complex cardiovascular health records.

Materials and Methods

Data Sources

With the increasing incidence of CVD, hospitals worldwide are generating extensive data, making its management challenging [Error! Bookmark not defined., Error! Bookmark not defined.]. Our aim is to store this data in a database, retrieve knowledge, and make overwhelming decisions in the context of diagnosis [Error! Bookmark not defined., 9]. This high-end CVD database provides clinical information such as patient ID, sex, age, date of visit, consent details, medical history, previous medications, family history, current etiology, comorbidities, clinical examination findings, symptoms, diagnostic test results, ECG, ECHO, angiography, chest

X-ray, treatments administered along with diagnostic outcomes, and follow-up recommendations. These data consist of high-resolution scanned copies of patient case files maintained daily from the day of admission to the day of discharge. CardioDB comprises 1,001 raw case files of patients with CVD that were compiled from the All India Institute of Medical Sciences (AIIMS), Bhubaneswar, over a two year span from 2021 to 2023 using mobile document scanner application version 10.0.69. The Institutional Ethics Committee (IEC) of AIIMS approved the ethical clearance for data collection. Figure 1 shows the study workflow from data collection to database development. The majority of these cases originated from various regions across India.

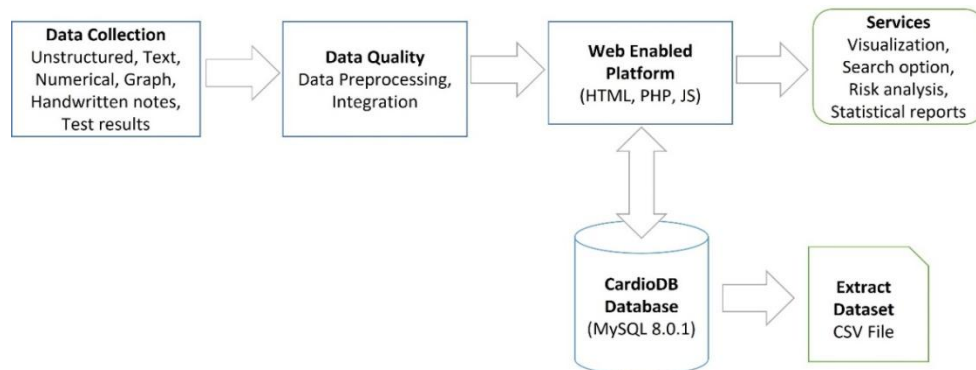


Figure 1. Study workflow for developing the cardiovascular database (CardioDB).

Database Development

After the collection of clinical patient data, a web-based portal was developed to enter high-dimensional data into the database through the designed web pages. This web platform was designed using Hypertext Markup Language (HTML) version 5.0, and the database connection pages were developed using Hypertext Preprocessor (PHP) version 8.0.2 and the Java scripting language, as shown in Figure 1. The CardioDB database was developed using MySQL version 8.0.2 on MariaDB server for the storage of clinical data of patients with CVD [14]. These HTML web pages are actively connected to their respective database tables and are used for long-term storage and easy access. In the database, the data types for all clinical variables (text, numeric, graph) were identified and selected accordingly.

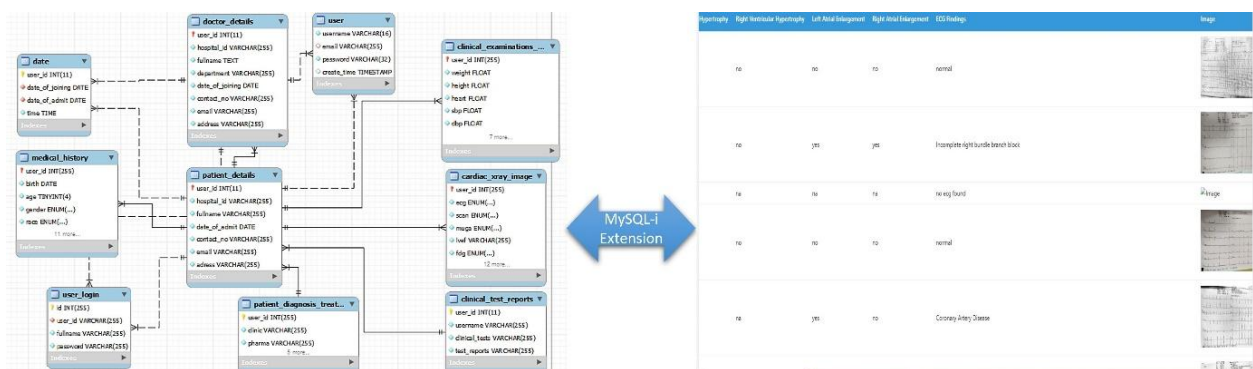


Figure 2. Schematic representation of data processing from the front end (PHP, JS) to the back end of the MySQL database via the MySQL-i extension

Figure 2 depicts the entity-relationship (ER) diagram of the CardioDB database. The connections are between the patient demographic data, date of joining, examination report, and final diagnosis. Strong relationships are represented by solid lines, whereas weak relationships are represented by dotted lines [3, 6, 13]. The aforementioned ER diagram contains six fact tables, namely patient_details, patient_diagnosis, medical_history, clinical_examinations, cardiac_X-ray_imaging, and clinical_test_reports, which are strongly related. Similarly, on the right-hand side, the ECG data of patients include the following details: patient ID, date, type of rhythm,

presence or absence of a bundle branch block, QRS width, QT interval, presence or absence of any of the following: a) right ventricular hypertrophy, b) left ventricular hypertrophy, c) left atrial enlargement, and d) right atrial enlargement, findings with ECG images.

Preprocessing of Raw Data

Data collection initially generated a raw repository of 1,065 individual patient files. Upon initial screening and quality control, out of which 64 files were found incomplete baseline clinical profiles so discarded, refining the cohort to 1,001 rows and 124 columns for subsequent detailed analysis. The data preprocessing step involves removing the columns with missing variables, correcting wrong entries, removing duplicate entries, and removing direct identifiers on the first step [10]. Few parameter values were standardized according to their measuring units, namely, the test results, date, weight, and height. Patient_id was selected as the primary key in the database, and all the data were integrated to produce a machine-readable dataset.

Data Insertion

Patient data were entered into the database using web forms. Figure 3 shows the corresponding major data categories for patients with CVD, namely, patient demographic data, clinical history, clinical examinations, clinical test reports, electrocardiograms, echocardiograms, coronary angiogram, cardiac imaging results, and treatment and diagnosis. After all the forms are filled out correctly, the submit button is clicked to insert the record into the database. These nine forms were separately connected to nine different tables in the CVD database and the data were recorded accordingly. The form “clinical examination” consists of parameters such as patient_ID, weight, height, heart rate, systolic BP, diastolic BP, NYHA score, signs and symptoms, underlying diseases, records of any life-threatening events, diagnosed diseases, and date. Similarly, the corresponding attributes of the angiography results were the left main coronary artery (LMCA), right coronary artery (RCA), left anterior descending artery (LADA), left circumflex artery (LCX), and the final impressions of the physician via angiography. The attributes of the treatment and diagnosis page include current diagnosis, pharmacological/implantable treatment, pharmacological treatment at the time of discharge/OPD, and non pharmacological treatment.

Major Risk Factors	Clinical Tests Reports	Categorical Features
☐ Age	☐ Angiography	☐ NYHA
☐ Gender	☐ Echocardiography	☐ Diabetes
☐ SysBP	☐ Electrocardiogram	☐ Hypertension
☐ Diabetes	☐ CBC	☐ Current Smoker
☐ Smoking	☐ Lipid Profile	☐ Clinical Symptoms
☐ High BMI	☐ Thyroid	☐ Etiologic report
☐ Triglycerides	☐ Electrolyte	☐ Gender
☐ Creatinine		
☐ Fasting_sugar		
☐ NYHA		

Figure 3. Major risk factors, clinical test reports and categorical features of the CardioDB database.

To ensure patient privacy and ethical compliance, all direct personal identifiers (names, addresses, contact numbers, government IDs) were permanently removed from clinical records prior to database entry. Each patient was assigned a unique, randomized alphanumeric study ID to securely link disparate data modalities, such as laboratory results, ECG, and ECHO findings, without revealing their identity. The master mapping file linking study IDs to patient identities is stored separately on an encrypted, access-restricted server, accessible only to authorized personnel. The complete database features are described according to their specifications: patient demographic data, clinical examination data, clinical test report data, and final diagnoses are briefly presented in Tables 1 and 2. These tables provide detailed descriptions of the data features.

Table 1. Patients' complete characterization of demographic data.

Parameters	Data Type	Description
Age	Integer	Patient's living year count
Race	Text	Classification on self-identification
Sex	Boolean	Patient is Male or female
Marital status	Boolean	Whether patient is married
Weight (in k.g.)	Float	Total Weight of the patient
Height (in c.m.)	Float	Total height of the patient
Address	Large Text	Patients address region
Regd. No.	Integer	This is 15 digit number to which the patient is registered in the Hospital
OPD, IPD	Integer	These are also 15 digit Hospital OPD and IPD numbers
Consent signed	Boolean	Does patient sign the consent form
Alcohol consumption	Boolean	Did the patient consume alcohol? If yes, then how frequently?
Smoking	Boolean	Did the patient smoke? If yes, then how frequently?
Tobacco	Boolean	Did the patient consume tobacco? If yes, then how frequently?
Etiology	Text	The study for the causes of disease. This field basically covers CVD related other illness in patient. Ex. Coronary Artery Disease, Coronary Heart Disease, etc.
Comorbidities	Text	The medical condition that coexists alongside to the primary diagnosis and affects your health. This field covers other secondary illness present at patient. Ex. Diabetes, Hypertension, kidney disease, etc.
Clinical Examinations	Float	This is the basic observations by the physician to know potential of the patient.
Clinical symptoms	Text	Objective or subjective related to disease that are mentioned by the patient. This field covers symptoms accumulated to the CVD. Ex. Palpitation, chest pain, shortness of breath, etc.
Any underlying disease	Text	Does the patient have any other diseases? Ex. Diabetes, anaemia, asthma, etc.
Patient inclusion criteria	Text	The specific characteristics, conditions, or parameters that a patient must have to be eligible for inclusion in our study.
Patient exclusion criteria	Text	The specific characteristics or conditions that disqualify a patient from the study.

CardioDB functions as a specialized clinical repository restricted entirely to coronary artery disease (CAD) cohorts. It systematically catalogs multidimensional patient data encompassing all major CAD phenotypes: stable and unstable angina, acute coronary syndrome (ACS), ST-elevation and non-ST-elevation myocardial infarctions (STEMI and NSTEMI), and post-revascularization status following percutaneous coronary intervention (PCI) or coronary artery bypass grafting (CABG).

Cardiovascular Risk Calculator

The backend algorithm employs the WAS technique, a three-stage prediction logic grounded in multivariable logistic regression. First, the model identifies which clinical variables independently predict CAD by evaluating multiple risk factors simultaneously while controlling for confounding effects. Second, logistic regression quantifies each factor's predictive strength through regression coefficients, where larger values indicate stronger CAD

associations. Third, these coefficients are converted into integer-based weights via standardized scaling: the smallest non-zero coefficient (0.2 for LDL cholesterol) serves as the reference unit, with all others proportionally scaled and rounded to the nearest integer. The resulting additive risk points are: age >40 years (+0.5), male sex (+1), smoking (+3), tobacco use (0), diabetes (+4), hypertension (+3), systolic BP >140 mmHg (+0.1 per point), LDL >160 mg/dL (+0.2), total cholesterol >240 mg/dL (+0.2), BMI >25 (+0.2), family history of CAD (+2), and other medical conditions (+1.5). During clinical assessment, the calculator evaluates patient parameters against these thresholds, accumulates the corresponding weights for all exceeded values, and generates a cumulative score that maps to the 0.1–30% CAD probability range for rapid, evidence-based risk stratification. The final risk score, which ranged between 0.1% and 30%, had four categories: A. 0.1 - 4.9 (Low risk)§ B. 5.0 - 7.4 (Borderline)§ C. 7.5 - 19.9 (Intermediate)§ and D. 20.0 - 30.0 (High Risk). This record window holds the patient's overall data during examination and treatment at the hospital. The records are connected to the underlying MySQL database via PHP connection pages [Error! Bookmark not defined., 8].

Table 2. Patients' complete characterization of clinical test reports and diagnosis

Parameters	Data Types	Description
CBC	Float	Patient's complete blood test report. This test report covers haemoglobin(g/dl), rbc(million/mm ³), lymphocytes(%), monocytes(%), neutrophils(%), eosinophils(%), basophils(%), platelets(lakhs/mm ³).
LIPID profile	Float	Patient's cholesterol and other fats present in blood report. This profile covers cholesterol (mg/dl), triglycerides (mg/dl), HDL cholesterol (mg/dl), LDL cholesterol (mg/dl).
Thyroid	Float	Measures patient's thyroid gland function, hormonal levels in blood. This test report obtains T3(mg/dl), T4(mg/dl) and TSH(mIU/L).
Electrolyte	Float	Measure the amount of electrolyte or minerals in blood. It contains urea(mg/dl), creatinine(mg/dl), sodium(mEq/L), potassium(mEq/L), chloride(mMol/L).
ECG	Image	This test conducted to know the patient's heart electrical activity and usually collected in image format with some interval values. Ex. Rhythm type, bundle branch type, QT(ms), QRS(ms), etc.
ECHO	Enum and Numerical	It is an ultrasound test that show heart and its nearby blood vessels and is collected in text-numerical form. Ex. Ejection fraction(%), ECHO dimensions: LVDV(mm), LVSD(mm), IVS(mm), LVPWD(mm).
ANGIOGRAM	Enum	It is an X-ray visualization of blood vessels nearby heart and is collected in text-numerical form. Artery type: LMCA, RCA, LADA, LCx, Ramus intermedius.
CAD	Binary	The clinical diagnostic standards used in study to classify a patient as having CAD.

Functioning of the Database and Website

The CardioDB database was developed in a web-based system. The database web pages contain a unique patient_id through which patient records are retrieved throughout the database. All these data were stored in MariaDB database version 10.4.17 in phpMyAdmin version 5.2.2, an open-source software tool that is portable for web database applications. The CardioDB webserver was developed using HTML, PHP, and CSS. The database website uses an Apache server to interact with the client, with HyperText Transfer Protocol Secure (HTTPS) security, and is hosted with the root domain.

Results

Table 3 shows the description parameters with their measuring units for the entire CVD dataset. The mean, standard deviation, and number of unique features in the database are also described. Also the non-normal distribution features were reported as median (Q1-Q3), the interquartile range using Mann-Whitney U test. Platelets had the highest mean and standard deviation, and neutrophils had the highest number of unique features in the dataset. The features, namely basophil, diabetes, and hypertension, had the lowest means. Similarly, basophils and smoking had the least standard deviation in the dataset, whereas the least number of unique features were sex, smoking, clinical symptoms, etiology, diabetes, and NYHA.

Table 4 explains the basic characteristics of the dataset features associated with CAD status. The non-normally distributed continuous variables are expressed as median (interquartile range) and compared using the Mann-Whitney U test. Categorical variables are expressed as frequencies and percentages. Variables demonstrating a statistically significant difference between CAD and non-CAD patients ($p < 0.05$) include systolic and diastolic blood pressure, red blood cell count, lymphocyte, monocyte, and neutrophil percentages, serum total cholesterol, serum chloride, TSH, fasting blood sugar, smoking status, clinical symptoms, etiology, diabetes, and hypertension.

Table 3. Patient dataset features and their means and standard deviations

CVD Dataset Parameters	Feature Description with measuring unit	Mean with Std. dev. (n=1001)	Median (Q1 - Q3)	No. of unique features
age	Age (in years)	55.9 (12.64)	57.00 (49.00 - 65.00)	67
sex	Gender (binary 0, 1)	0.72 (0.44)	729 (72.8%)	2
heart	Vital physiological function (number)	73.42 (19.9)	75.00 (64.00 - 84.00)	72
sbp	Systolic BP (mmHg)	117.43 (29.28)	118.00 (110.00 - 130.00)	84
dbp	Diastolic BP (mmHg)	73.11 (17.84)	75.00 (70.00 - 80.00)	56
bmi	Body mass index from person's weight and height (kg/m ²)	21.3 (8.11)	23.00 (19.50 - 25.80)	156
Hemo	Haemoglobin: Protein in RBC (g/dL)	11.48 (4.28)	12.60 (11.00 - 14.10)	90
rbc	Red blood cells in blood (c/mcL)	4.2 (1.6)	4.55 (4.02 - 5.07)	214
lympho	Lymphocyte: White blood cells (WBC) (μL)	24.45 (11.89)	26.60 (19.40 - 31.80)	262
mono	Monocyte: WBC produces bone marrow (mm ³)	3.64 (2.14)	3.70 (2.80 - 4.60)	60
neutro	Neutrophils: WBC that act as first line defence (μL)	57.68 (22.03)	62.70 (53.60 - 71.20)	284
eosin	Eosinophils: WBC part of immune system (cells/μL)	-	2.90 (1.30 - 5.40)	135
baso	Basophils: WBC participate in allergic reactions (mm ³)	0.3 (0.22)	0.30 (0.20 - 0.40)	14
plate	Platelets: blood fragments form clots (platelets/μL)	235.32 (121.87)	244.00 (166.00 - 309.00)	262
serum	Serum total cholesterol (mg/dL)	123.28 (81.29)	135.00 (79.00 - 178.00)	155
serumt	Serum triglycerides: fat found in blood (mg/dL)	114.54 (93.36)	118.00 (55.00 - 163.00)	183
serumh	Serum HDL cholesterol (mg/dL)	37.84 (24.49)	42.00 (24.00 - 57.00)	58
seruml	Serum LDL cholesterol (mg/dL)	63.65 (51.8)	71.00 (0.00 - 101.00)	124
t3	Level of Triiodothyronine in blood (ng/dL)	-	1.90 (0.00 - 2.96)	179
t4	Level of Thyroxine in blood (mcg/dL)	-	1.24 (0.00 - 5.00)	143
tsh	Thyroid Stimulating Hormone (mIU/L)	-	1.22 (0.00 - 2.44)	258
blood	Blood urea nitrogen (mg/dL)	23.32 (13.51)	22.00 (17.00 - 28.00)	62
creat	Serum creatinine in blood (mg/dL)	0.91 (0.4)	0.90 (0.80 - 1.10)	27
sodium	Serum sodium in blood (mmol/L)	122.75 (39.65)	135.00 (133.00 - 138.00)	37
potassium	Serum potassium in blood (mEq/L)	-()	4.17 (3.90 - 4.48)	163

CVD Dataset Parameters	Feature Description with measuring unit	Mean with Std. dev. (n=1001)	Median (Q1 - Q3)	No. of unique features
chloride	Serum chloride in blood (mEq/L)	92.31 (29.29)	101.00 (97.00 - 104.00)	35
fasting	Sugar level in blood before eating (mmol/L)	83.86 (8.37)	95.00 (0.00 - 109.00)	127
smoking	Smoking (binary 0, 1)	0.07 (0.26)	73 (7.3%)	2
Clinical symptoms	Active clinical symptoms (binary 0, 1)	0.69 (0.45)	699 (69.8%)	2
etiology	Active etiologic report (binary 0, 1)	0.73 (0.44)	736 (73.5%)	2
diabetes	Diabetes (binary 0, 1)	-	308 (30.8%)	2
hypertension	Hypertension (categorical 0, 1, 2, 3)	-	307 (30.7%)	4
nyha	NYHA (categorical 0, 1, 2)	1.26 (0.48)	758 (75.7%)	2

Figure 4 shows the homepage of the CardioDB database and the CVD patient record window. The home page features three menus: (a) CVD patient records, (b) patient search via a unique key, and (c) a CVD risk calculator. The final CVD patient dataset contains 1,001 patient cases with 124 clinical attributes, where each patient_id is unique, and all this information is displayed in a tabular format. The patient search option was provided to quickly access a specific patient within the database using the patient_id. Finally, the CVD risk calculator, which analyzes the 5-year heart disease risk for patients based on 13 major key risk factors, is integrated into the system. The calculated results are shown on a scale ranging from 0 to 30 and the risk scores are also categorized accordingly.

Table 4. Standard characteristics of patient datasets with and without CVD

CVD Parameter	Patients with CAD (n = 722)	Patients without CAD (n = 279)	p-value
Age (years)	57.0 (48.0 – 65.0)	57.0 (50.0 – 65.0)	0.215
Gender (Male), n (%)	507 (70.2%)	222 (79.6%)	0.004
Heart beat (bpm)	76.0 (65.0 – 84.0)	71.0 (64.0 – 82.0)	0.068
Systolic BP (mmHg)	120.0 (110.0 – 136.0)	110.0 (108.0 – 122.0)	<0.001
Diastolic BP (mmHg)	77.0 (70.0 – 83.0)	70.0 (68.0 – 79.0)	<0.001
BMI (kg/m ²)	23.2 (19.7 – 26.0)	22.9 (19.5 – 25.6)	0.791
Haemoglobin (g/dL)	12.1 (10.4 – 13.7)	13.2 (12.6 – 14.8)	<0.001
RBC (10 ⁶ /μL)	4.5 (3.9 – 5.0)	4.8 (4.4 – 5.2)	<0.001
Lymphocyte (%)	24.8 (15.5 – 31.4)	29.2 (24.5 – 32.6)	<0.001
Monocyte (%)	3.5 (2.7 – 4.3)	4.1 (3.3 – 5.3)	<0.001
Neutrophils (%)	62.7 (52.3 – 71.4)	62.7 (54.7 – 68.9)	0.207
Eosinophils (%)	2.8 (0.9 – 5.6)	3.3 (1.8 – 4.8)	0.079
Basophils (%)	0.3 (0.1 – 0.4)	0.3 (0.2 – 0.5)	<0.001
Platelets (10 ³ /μL)	238.0 (155.0 – 296.0)	252.0 (180.5 – 334.0)	<0.001
Serum total chol. (mg/dL)	124.0 (0.0 – 167.0)	166.0 (128.0 – 198.0)	<0.001
Serum triglycerides (mg/dL)	105.5 (0.0 – 166.0)	133.0 (105.0 – 161.0)	<0.001
Serum HDL chol. (mg/dL)	36.0 (0.0 – 45.0)	62.0 (54.0 – 69.5)	<0.001
Serum LDL chol. (mg/dL)	62.0 (0.0 – 102.0)	80.0 (57.0 – 99.0)	<0.001
T3 (ng/dL)	2.7 (0.0 – 3.1)	1.0 (0.8 – 1.9)	<0.001

T4 (µg/dL)	1.1 (0.0 – 1.3)	8.5 (6.9 – 10.0)	<0.001
TSH (mIU/L)	0.8 (0.0 – 2.3)	2.0 (1.0 – 2.6)	<0.001
Blood urea nitrogen (mg/dL)	21.0 (17.0 – 28.0)	22.0 (19.0 – 29.0)	0.009
Serum creatinine (mg/dL)	0.9 (0.8 – 1.1)	0.9 (0.8 – 1.1)	0.590
Serum sodium (mEq/L)	135.0 (131.0 – 137.0)	137.0 (135.0 – 139.0)	<0.001
Serum potassium (mEq/L)	4.1 (3.8 – 4.4)	4.3 (4.0 – 4.5)	<0.001
Serum chloride (mEq/L)	99.0 (95.0 – 102.0)	104.0 (102.0 – 107.0)	<0.001
FBS (mg/dL)	93.0 (0.0 – 114.0)	99.0 (87.0 – 106.0)	0.004
Smoking, n (%)	73 (10.1%)	0 (0.0%)	<0.001
Clinical symptoms, n (%)	695 (96.3%)	4 (1.4%)	<0.001
Etiologic report, n (%)	716 (99.2%)	20 (7.2%)	<0.001
Diabetes, n (%)	308 (42.7%)	0 (0.0%)	<0.001
Hypertension, n (%)	307 (42.5%)	0 (0.0%)	<0.001
Nyha	558 (77.3%)	200 (71.7%)	0.0485

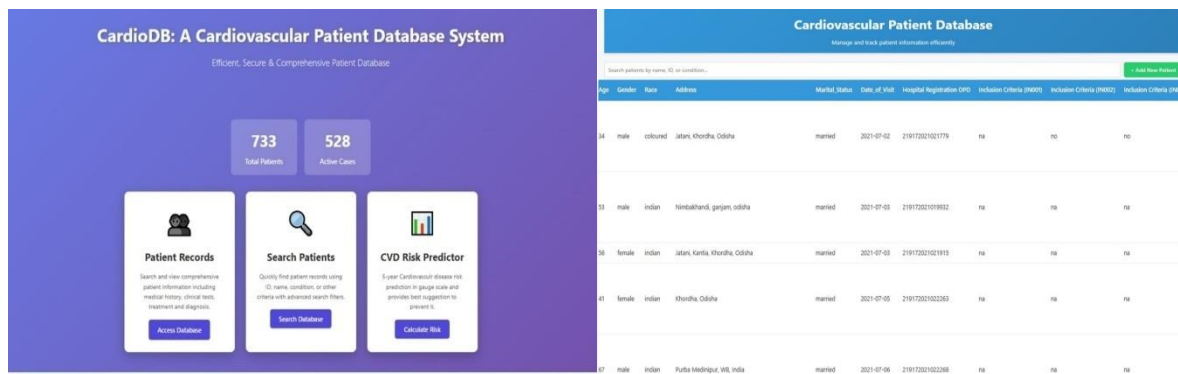


Figure 4. Screenshot of the home page and patient records in the database

CVD Risk Calculator

Figure 5 shows a screenshot of the CVD risk calculator. The left-hand side of the figure shows the blank key fields, and the right-hand side shows the outcome on the gauge scale with the preferred actions. This CVD calculator provides a general 5-year risk prediction by considering 13 major clinical risk parameters as inputs and calculates the next 5 years of risk. Based on the risk score generated, certain follow-ups are recommended for the user. The interface was designed using PHP, HTML, and JavaScript.

The predictive performance of the proposed risk calculator was evaluated using multivariable logistic regression on the curated cohort of 1,001 patients. The model incorporated key clinical features, including age, sex, smoking status, alcohol status, tobacco consumption, systolic blood pressure, total and LDL cholesterol, BMI, diabetes, hypertension, NYHA class, and presence of clinical symptoms. The model exhibited exceptional discrimination (AUROC = 0.997) and overall accuracy (98.8%), achieving 98.6% sensitivity and 99.3% specificity at the optimal Youden cutoff. Additional metrics included a 0.992 F1-score, 99.7% PPV, and 96.5% NPV. Calibration was excellent, reflected by a Brier score of 0.009 and close alignment with the ideal diagonal Figure 6(a). Furthermore, decision curve analysis confirmed strong clinical utility, showing superior net benefit over default treat-all or treat-none strategies across relevant thresholds Figure 6(b).

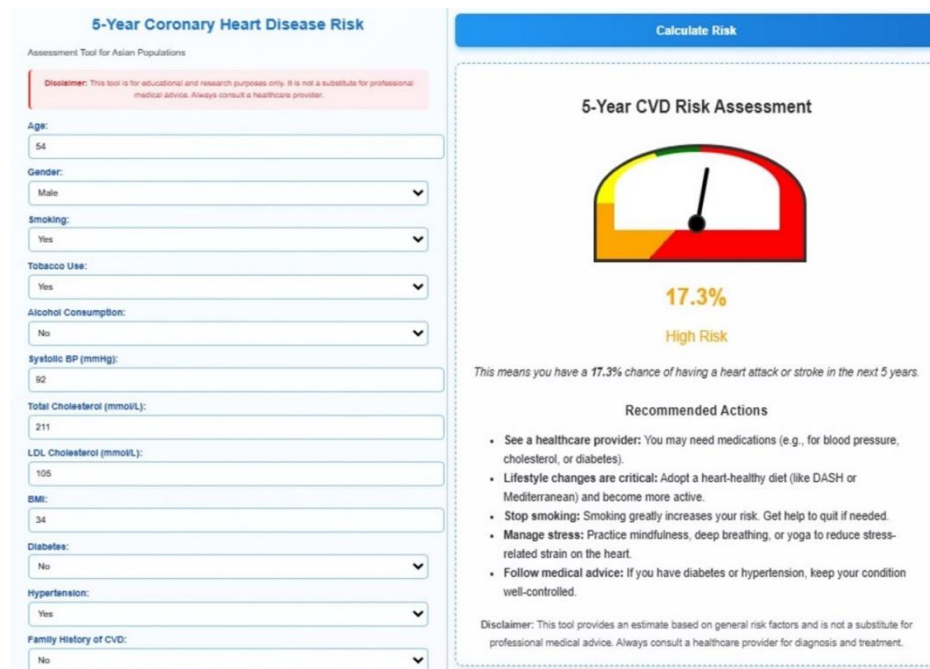


Figure 5. Five-year CVD risk analysis via the Weighted Additive Scoring (WAS) technique

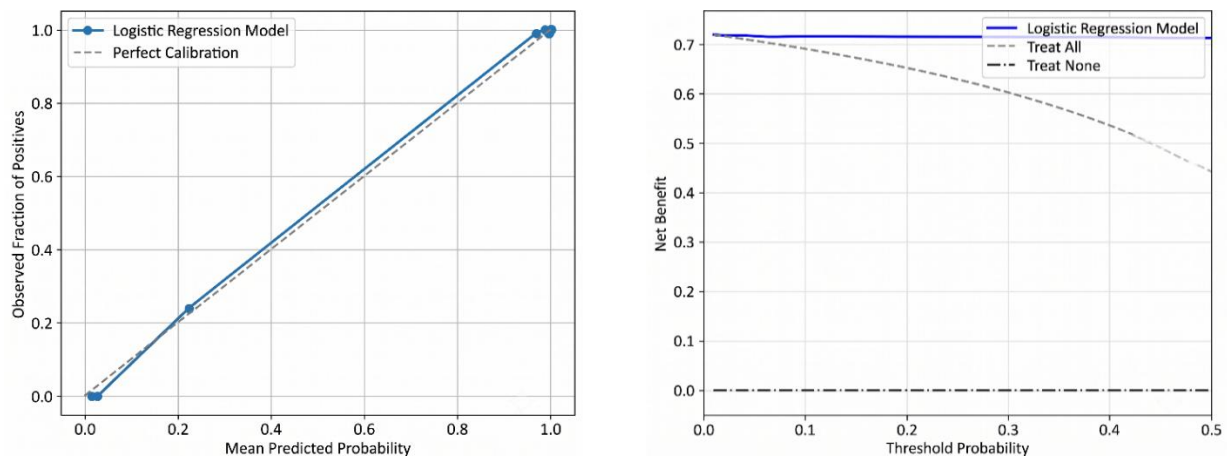


Figure 6(a). Calibration curve, **Figure 6(b).** Decision curve analysis for the Validation of the Cardiovascular Risk Calculator

Statistical Analysis of the Database

The statistical graphs were generated using seaborn and matplotlib conducted on the CVDdataset to obtain the reports statistically, which provides an ultimate understanding of the illness.

Figure 7 illustrates the mortality patterns among male and female patients concerning age. The joint plot visualizes the associations between mortality, age, and sex. The x-axis represents patient age across the entire population, ranging from 18 to 84 years, whereas the y-axis distinguishes sex, with males coded as 1 and females as 0. The color gradient in the plot reflects the mortality rate, and distinct shades are used to differentiate the outcomes. CVD-positive cases were more prevalent in both men and women than CVD-negative cases. Males aged between 40 and 78 years had a greater incidence of CVD than females in the same age range. A notable spike occurred between the ages of 60 and 65, indicating that this group was the most affected.

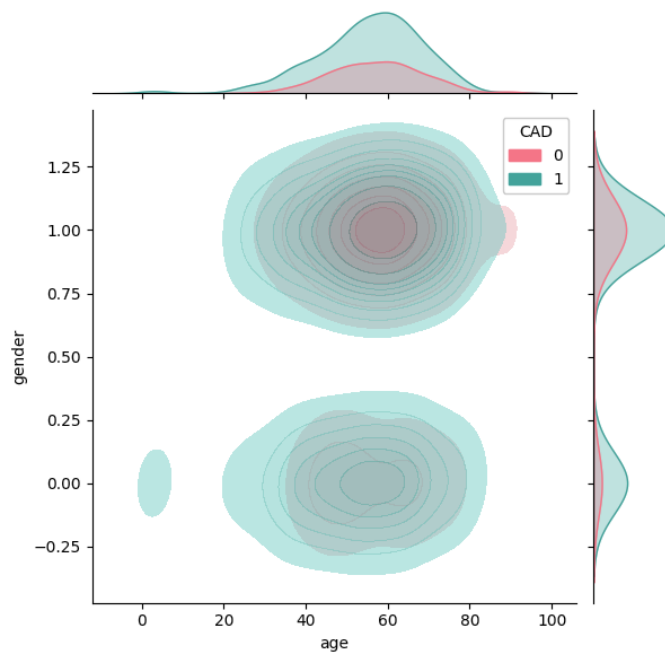


Figure7. Sex and age specific CVD mortality ratios in the dataset

Figure 8 illustrates the major risk factors, namely, hypertension, diabetes, cholesterol (LDL), and BMI, which are most closely related to CVD. In Figure 8(a), the joint plot depicts the relationship between hypertension and CVD with reference to the patient's age. The hypertension label indicated high risk in patients aged between 48 and 70 years, including all positive CVD cases included. Figure 8(b) shows the relationship between diabetes and CVD concerning patient age. The diabetes label strongly affected the age group ranging from 55 to 67 years, with all positive cases. Figure 8(c) shows the relationship between low-density lipoprotein (LDL) cholesterol and CVD with age. The cholesterol label affected almost all age groups in a scattered manner from 15 to 90 years, with all positive cases and some cases indicating increased cholesterol levels and a high risk for individuals. Additionally, Figure 8(d) shows the relationship between body mass index (BMI) and CVD with respect to age. A BMI label below 19 or over 25 is considered unhealthy. The BMI score was abnormal in the 55–60-year-old age group, with positive cases.

Functioning of the Database and Website

The CardioDB database was developed in a web-based system in which the clinical reports of 730 patients with CVD were stored. The web database stored in the high-configuration server is accessible via the link <https://cardiodb.coenpt.in/>. To protect against external security threats and ensure data privacy, the website is encrypted and accessible only to authorized users. The CardioDB system currently employs strict, role-based login authentication, and upon approval, unique, secure credentials are issued to the requesting party. The database web pages contain a unique patient_id through which patient records are retrieved throughout the database. All these data were stored in MariaDB database version 10.4.17 in phpMyAdmin version 5.2.2, an open-source software tool that is portable for web database applications.

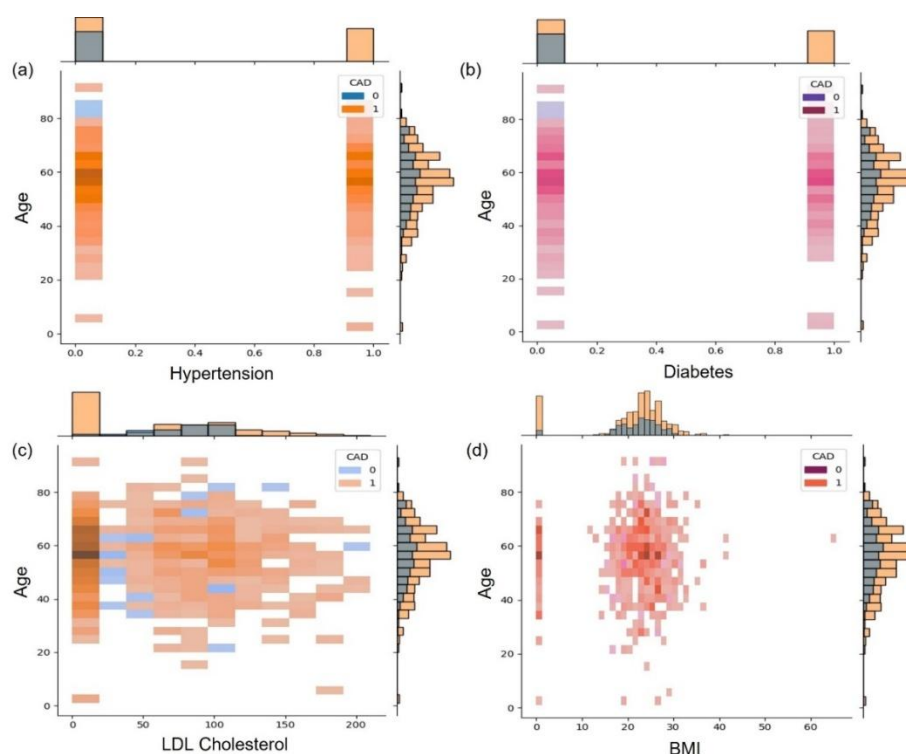


Figure 8. Major CVD risk factors according to age group in our dataset

Figure 9 shows the complete working scenario of our database. It has two components: (a) a client and (b) a server. In the workbench, the MySQL database was connected to the server with the required authentication. The client side was enabled with an interface for true visualization and provided esteemed services from which the data could be retrieved. The PHP connection pages used `get()`, `select()`, `getId()` methods at the server site, whereas the `register()` and `view()` methods are used at the client.

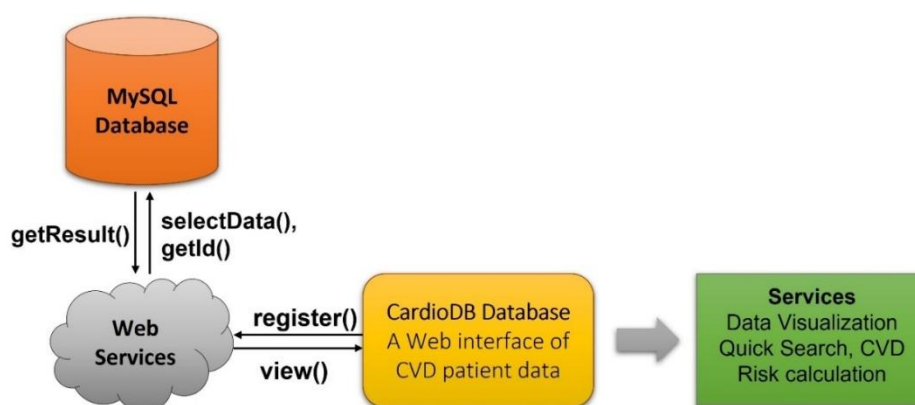


Figure 9. Working interface of a CVD patient's clinical database

Discussion

This large cardiovascular database comprises 1,001 patient records and includes a wide range of features, from patient demographics to diagnosis reports, collected over a two-year period explained in Tables 1 and 2. Figure 7 shows that the mortality rate was higher in male patients than in female patients, likely due to risk factors such as disease symptoms, diabetes, hypertension, smoking, and alcohol consumption. Maintaining and accessing real clinical data is quite challenging because of the complexity and lack of advancements at the sources, namely hospitals, clinics, and healthcare sectors. This method of designing a database is ideal for both short-term utilization

and long-term storage, facilitating access to datasets with important features, and can be useful for decision-making diagnosis in the future. However, this type of database is not very fruitful for research because of its vast number of heterogeneous features, which is one of its limitations. The CardioDB database is a comprehensive primary database of patients with CVD, containing complete information from the date of admission to the date of discharge. With this vast diagnostic insight, clinical professionals, researchers, and scientists will be better able to understand cases and improve diagnostic systems. In CardioDB, cardiology test reports, clinical examinations, etiology and comorbidity reports, and ECG, ECHO, and angiography results are presented in two-dimensional table formats on our web portal portrayed in Figure 4. There are few real-world studies on the current trends in CVD, epidemiological patterns, and sex differences in the clinical practice of patients with coronary artery disease (CAD) [Error! Bookmark not defined., Error! Bookmark not defined.]. The 5-year risk analysis calculates the risk of patients in the future presented in Figure 5. Unlike molecular databases (e.g., MorCVD) or meta-catalogs (e.g., IndiaBioDB), our system is purpose-built for granular, patient-level clinical data management [Error! Bookmark not defined., Error! Bookmark not defined., Error! Bookmark not defined.]. While aligning with large-scale registries like the Malaysian NCVD-PCI Registry, our database offers superior multidimensional granularity by integrating over 100 variables per record, including specific ECG, ECHO, and angiographic findings. Furthermore, unlike public repositories, our system enforces strict, role-based authentication to ensure ethical compliance when handling sensitive patient data. While the risk calculator provides immediate decision support for new patients at the source institution, its generalizability to geographically distinct populations is limited [Error! Bookmark not defined., 11, 12]. Direct deployment elsewhere risks miscalibration due to regional variations in genetics and lifestyle, necessitating future multi-center external validation and model recalibration. Beyond geographical constraints, this study has several inherent methodological and systemic limitations. The retrospective design introduces potential selection and documentation bias. Furthermore, the linear logistic regression model may fail to capture complex, non-linear interactions as effectively as advanced machine learning. Finally, long-term system scalability, interoperability, and refined access controls require ongoing optimization.

Conclusion

This study introduces CardioDB, an integrated relational database purpose-built for the structured management of multidimensional CVD patient data. By capturing more than 100 variables per patient including demographics, laboratory panels, ECG, and ECHO findings, the system enables comprehensive clinical data consolidation. The accompanying logistic regression-based risk calculator showed robust internal predictive performance, providing a practical tool for early risk stratification. However, the single-center, retrospective design of this study necessitates future multi-center external validation to confirm the model's generalizability across diverse populations. CardioDB thus represents a meaningful advance in population-specific CVD risk prediction, with ongoing work focused on enhancing system interoperability and reinforcing role-based security to support broader clinical adoption.

List of Abbreviations: Cardiovascular Disease (CVD), Coronary Artery Disease (CAD), Percutaneous Coronary Intervention (PCI), Host–Pathogen Interactions (HPI), Institutional Ethical Committee (IEC), Entity Relationship Diagram (ER-Diagram), Hyper Text Markup Language (HTML), Hypertext Preprocessor (PHP), Java Script (JS), Body Mass Index (BMI), Low Density Lipoprotein (LDL), Thyroid Stimulating Hormone (TSH), Etiologic, Electrocardiogram (ECG), Echocardiogram (ECHO), Angiography (ANGIO), Weighted Assisted Scoring (WAS), New York Heart Association (NYHA).

Author Contributions: P.K.N. defined the research strategy and final review of the manuscript. S.S. curated clinical cardiovascular datasets and maintained the data quality. T.K.B. developed the database schema, implemented the relational backend and constructed the clinical. All authors have read and approved the final manuscript.

Funding: Financial support from ICMR, Govt. of India, Sanction No. ISRM/12(129)/2020, Project Id: 2020-5341 is highly acknowledged. ICMR – Indian Council of Medical Research [Grant No. ISRM/12(129)/2020].

Ethical Clearance Reference Number: T/EMF/CTVS/19/32 on dated: 10.02.2020.

Data Availability Statement: CardioDB integrates the patient data from Hospital and curated clinical datasets that contain protected health information. The database schema and clinical information images will publicly available at <https://cardiodb.coenpt.in/>.

Conflict of interest: All the authors declare that they have no conflicts of interest.

References

1. Lee CY, Hairi NN, Wan Ahmad WA, Ismail O, Liew HB, Zambahari R, et al. Are there gender differences in coronary artery disease? The Malaysian National Cardiovascular Disease Database-Percutaneous Coronary Intervention (NCVD-PCI) Registry. *PLoS One*. 2013;8(8):e72382. <https://doi.org/10.1371/journal.pone.0072382>
2. Choi EK. Cardiovascular research using the Korean national health information database. *Korean Circ J*. 2020;50(9):754-772. <https://doi.org/10.4070/kcj.2020.0176>
3. Badawy MAEMD, Naing L, Johar S, Ong S, Rahman HA, Tengah DSNAP, et al. Evaluation of cardiovascular diseases risk calculators for CVDs prevention and management: scoping review. *BMC Public Health*. 2022;22(1):1742. <https://doi.org/10.1186/s12889-022-13944-w>
4. Akhter M, Bhuiyan MNZ, Siddique S, Aleem NT, Jahan M, Sultana SN. Clinical evaluation and semen analysis in male infertility: a study on 100 cases. *Saudi Int J Obstet Gynecol*. 2022;5(2):43-51. <https://doi.org/10.36348/sijog.2022.v05i02.004>
5. Berger J. Mean platelet volume and prevalence of peripheral artery disease. *Circulation*. 2009;120(18 Suppl):S39-S40. https://doi.org/10.1161/circ.120.suppl_18.a39-b
6. Maas A, Appelman Y. Gender differences in coronary heart disease. *Neth Heart J*. 2010;18(12):598-603. <https://doi.org/10.1007/s12471-010-0841-y>
7. Durai R. Colonic pseudo-obstruction. *Singapore Med J*. 2009;50(3):237-244.
8. Agarwal A, Agrawal P, Sharma A, Kumar V, Mugdal C, Dhall A, Raghava GP. A repository of web-based bioinformatics resources developed in India. *bioRxiv*. 2020. <https://doi.org/10.1101/2020.01.21.855627>
9. Sengupta D, Arora P, Pant S, Naik PK. Design of dimensional model for clinical data storage and analysis. *Appl Med Inform*. 2013;32(2):47-53.
10. Yang C, Street WN, Robinson JG. 10-year CVD risk prediction and minimization via InverseClassification. In: *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*; 2012 Jan 28-30; Miami, FL. New York: ACM; 2012. p. 603-610. <https://doi.org/10.1145/2110363.2110431>
11. Novotný P, Wild J. The relational modelling of hierarchical data in biodiversity databases. *Database (Oxford)*. 2024;2024:baae107. <https://doi.org/10.1093/database/baae107>
12. Mesbah I, Habermann B, Rico F. MechanoProDB: a web-based database for exploring the mechanical properties of proteins. *Database (Oxford)*. 2024;2024:baae047. <https://doi.org/10.1093/database/baae047>
13. Tripathy K, Singh B, Singh N, Rai V, Misra G, Singh NK. A database of wild rice germplasm of *Oryza rufipogon* species complex from different agro-climatic zones of India. *Database (Oxford)*. 2018;2018:bay058. <https://doi.org/10.1093/database/bay058>
14. Blansit BD. The basics of relational databases using MySQL. *J Electron Resour Med Libr*. 2006;3(3):135-148. https://doi.org/10.1300/J382v03n03_08