Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Creating Clinically Labeled MIMIC-III PPG Datasets for CAD and CVD Research: A Reproducible Workflow

Raghunath Reddy

raghunathreddy@pace.ac.in

PACE Institute of Technology and Sciences

Gautam Kumar

Manipal University Jaipur

Veeranjaneyulu R

PACE Institute of Technology and Sciences

Research Article

Keywords: Photoplethysmography, Coronary artery disease, Cardiovascular disease, Biomedical engineering, MIMIC-III, Physiological waveform dataset, Signal quality control

Posted Date: June 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10055757/v1

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

Additional Declarations: No competing interests reported.

Creating Clinically Labeled MIMIC-III PPG Datasets for CAD and CVD Research: A Reproducible Workflow

Raghunath Reddy[1,*] Gautam Kumar[2,*] R Veeranjaneyulu [1]

[1]PACE Institute of Technology and Sciences, Ongole, AP, India

[2]Manipal University Jaipur, Jaipur, Rajasthan, India

[*]Corresponding authors: raghunathreddy@pace.ac.in; gautam.kumar@jaipur.manipal.edu

Abstract

Background: Public datasets containing photoplethysmography (PPG) signals with patient-level cardiovascular disease labels remain limited, particularly for coronary artery disease (CAD) research. This study describes a reproducible workflow for constructing clinically labeled CAD and broader cardiovascular disease (CVD) PPG datasets from the matched waveform and clinical components of MIMIC-III.

Methods: PPG channels were identified from waveform metadata, quality-screened, and segmented into six one-minute windows corresponding to minutes 3–8 of each accepted recording. Clinical information from PATIENTS, ADMISSIONS, and DIAGNOSES_ICD was used to derive demographics, admission-linked diagnosis events, ICD-9-based disease labels, and selected comorbidity indicators. Patient-level aggregation retained one selected waveform record per subject.

Results: The final strict CAD dataset contains 3,465 unique patients, including 1,625 CAD-positive subjects and 1,840 controls. The implemented cardiac/arterial

CVD dataset contains 5,456 unique patients, including 3,616 CVD-positive subjects and the same 1,840 controls. The released patient-level files combine quality-screened PPG waveform windows with age, sex, comorbidity flags, waveform metadata, and binary CAD/CVD labels.

Conclusions: The workflow provides a reusable and reproducible resource for cardiovascular PPG research. By documenting waveform extraction, signal-quality filtering, clinical linkage, ICD-9 labeling, and patient-level validation, the dataset framework supports future biomedical engineering studies using MIMIC-derived physiological waveforms.

Keywords: Photoplethysmography; Coronary artery disease; Cardiovascular disease; Biomedical engineering; MIMIC-III; Physiological waveform dataset; Signal quality control

1 Background

Cardiovascular diseases remain a leading cause of morbidity and mortality worldwide. Coronary artery disease (CAD), a major manifestation of cardiovascular disease (CVD), is strongly associated with myocardial infarction, ischemic heart disease, and adverse clinical outcomes. Scalable approaches for cardiovascular monitoring and risk assessment are therefore important for both clinical and translational research.

Photoplethysmography (PPG) is a non-invasive optical sensing modality widely available in pulse oximeters, bedside monitors, and wearable devices. Because PPG reflects peripheral blood-volume changes, its morphology contains information related to vascular compliance, arterial stiffness, autonomic regulation, and hemodynamic status. These properties have motivated increasing use of PPG in biomedical signal processing, cardiovascular assessment, blood pressure estimation, and wearable monitoring [1, 6, 13, 19].

Despite this interest, public PPG resources with clinically meaningful CAD/CVD labels are scarce. Many available datasets were designed for heart-rate estimation, stress recognition, activity monitoring, or cuff-less blood pressure prediction rather than patient-

level disease cohort construction. In addition, insufficient documentation of preprocessing, diagnosis-code logic, and patient-level separation can limit reproducibility and may introduce information leakage when waveform windows from the same subject are used across analysis partitions.

The MIMIC-III matched waveform and clinical databases provide an opportunity to construct clinically linked physiological datasets by combining bedside waveforms with electronic health records [9]. However, this requires careful waveform extraction, signal-quality assessment, diagnosis-code processing, demographic enrichment, and patient-level aggregation. Label assignment is particularly challenging when waveform timestamps or admission identifiers are not available in the extracted PPG rows.

This study presents a reproducible workflow for constructing patient-level CAD and CVD PPG datasets from MIMIC-III. The workflow extracts PPG waveform windows, applies signal-quality checks, integrates demographic and diagnosis information, assigns ICD-9-based subject-level labels, and produces leakage-controlled patient-level CSV files. The resulting datasets are intended as dataset-contribution resources for cardiovascular PPG research rather than as model-development benchmarks.

The main contributions are:

- A documented extraction-to-labeling workflow for constructing clinically linked PPG datasets from MIMIC-III matched waveform and clinical data.
- Strict CAD versus Non-CAD and cardiac/arterial CVD versus Non-CVD patient-level cohorts generated using ICD-9-based subject-level label definitions.
- Patient-level files combining PPG waveform windows with age, sex, and comorbidity indicators.
- Integrity checks for patient-level uniqueness, class counts, control sharing, and waveform-window availability.

2 Related Work

2.1 Public PPG Datasets

Early public PPG resources were frequently designed around heart-rate estimation during movement. The IEEE Signal Processing Cup dataset supported heart-rate monitoring during physical activity, while PPG-DaLiA provided wearable PPG and accelerometer recordings for heart-rate estimation in daily-life conditions [23, 18]. These datasets helped establish PPG as a practical signal source for wearable cardiovascular monitoring, but they were not intended for patient-level disease cohort construction.

Other wearable datasets broadened the physiological context beyond heart-rate estimation. WESAD combined PPG with ECG, respiration, and other wearable signals for stress and affect recognition, whereas the Empatica E4 stress dataset similarly focused on PPG/EDA recordings for stress assessment [20, 3]. Activity-oriented and smartphone-based datasets, including the wireless-sensor activity dataset and BUT PPG, further contributed recordings for activity monitoring, signal-quality assessment, and heart-rate analysis [2, 15].

Recent datasets have emphasized more naturalistic and scalable wearable settings. WildPPG and GalaxyPPG provide longer or semi-naturalistic PPG recordings from consumer or wearable devices, while UTSA-PPG focuses on multi-channel PPG for HR/HRV studies [13, 16, 22]. Pulse-PPG extends this direction by using diverse wearable PPG data for foundation-model development across laboratory and field settings [19]. Together, these resources support wearable signal analysis, but they generally do not provide CAD/CVD labels derived from linked clinical diagnosis records.

2.2 Clinical and MIMIC-Derived PPG Resources

Clinical waveform repositories such as PhysioNet and MIMIC make it possible to connect physiological recordings with structured health-record data [7, 9]. VitalDB provides a complementary surgical vital-signs resource, and PulseDB combines MIMIC-III and Vi-

talDB signals to support benchmarking for cuff-less blood pressure estimation [11, 21]. These resources are clinically valuable, but their common downstream use has been parameter estimation rather than reusable CAD/CVD-labeled cohort release.

Several MIMIC-derived and clinical PPG studies have focused on cardiovascular or respiratory signal modeling. Kachuee et al. used physiological waveform data for cuff-less blood pressure estimation, and Ibtehaz and Rahman investigated PPG-to-arterial blood pressure waveform estimation. Pimentel et al. studied respiratory-rate estimation using pulse oximeter waveforms [10, 8, 17]. MIMIC-III-Ext-PPG is especially relevant because it provides a MIMIC-derived PPG benchmark for cardiovascular and respiratory signal analysis, although its objective differs from the present CAD/CVD cohort-construction framework [14].

Disease-oriented PPG work has also examined vascular risk and other clinical targets. Liang et al. used PPG for hypertension risk stratification, and Chan et al. evaluated smartphone PPG for atrial fibrillation screening [12, 4]. Hb-PPG provides multi-wavelength PPG data for non-invasive hemoglobin assessment, illustrating the use of PPG for clinically meaningful biomarker estimation beyond heart-rate monitoring [5]. These studies show the breadth of clinical PPG research, but publicly documented patient-level CAD/CVD PPG resources with explicit ICD-code logic and leakage-aware aggregation remain limited.

2.3 Dataset Gap

The main gap addressed by this study is the absence of a reusable PPG dataset framework that combines clinical CAD/CVD labels, transparent ICD-9 cohort definitions, and one-subject-level aggregation. This distinction is important because disease-labeled resources require different documentation and leakage controls than datasets designed primarily for signal reconstruction, heart-rate estimation, artifact detection, or cuff-less blood pressure modeling [24, 21].

Table 1 summarizes representative PPG datasets and highlights the distinction between parameter-estimation resources and the present disease-labeled dataset contribution. The

Table 1: Representative public PPG datasets and the proposed framework.

Dataset	Primary Application	Key Modalities	Clinical Labels
IEEE SPC [23]	Heart-rate estimation	PPG, accelerometer	No
PPG-DaLiA [18]	HR estimation in daily activities	PPG, accelerometer	No
WESAD [20]	Stress and affect recognition	PPG, ECG, respiration	No
WildPPG [13]	Real-world wearable HR monitoring	PPG, ECG, accelerometer, temperature	No
UTSA-PPG [22]	HR/HRV analysis	Multi-channel PPG, accelerometer	No
PulseDB [21]	Cuff-less BP estimation	ECG, PPG, ABP	Blood pressure
Hb-PPG [5]	Hemoglobin estimation	Multi-wavelength PPG	Hb/BP/Glucose
Proposed Dataset	**CAD/CVD research**	**PPG + clinical metadata**	**CAD/CVD labels**

proposed work specifically targets patient-level CAD and CVD cohort construction using MIMIC-III waveform and diagnosis data.

This work lowers the entry barrier for researchers interested in clinically linked PPG analysis. While MIMIC-III contains suitable waveform and clinical data, deriving CAD/CVD-labeled PPG cohorts requires non-trivial waveform extraction, quality filtering, clinical-table linkage, ICD-9 label construction, and patient-level aggregation. By providing a documented and reproducible workflow, this study makes such data easier to regenerate and reuse, supporting future research in an area where relatively few studies have exploited MIMIC-derived PPG signals for cardiovascular disease cohort construction.

3 Materials and Methods

3.1 Workflow Overview

The dataset was generated using the five-step workflow summarized in Figure 1: PPG waveform extraction, demographic linkage, diagnosis-event labeling, patient-level aggregation, and EDA-based validation. The released GitHub repository provides the code needed to implement each workflow step.

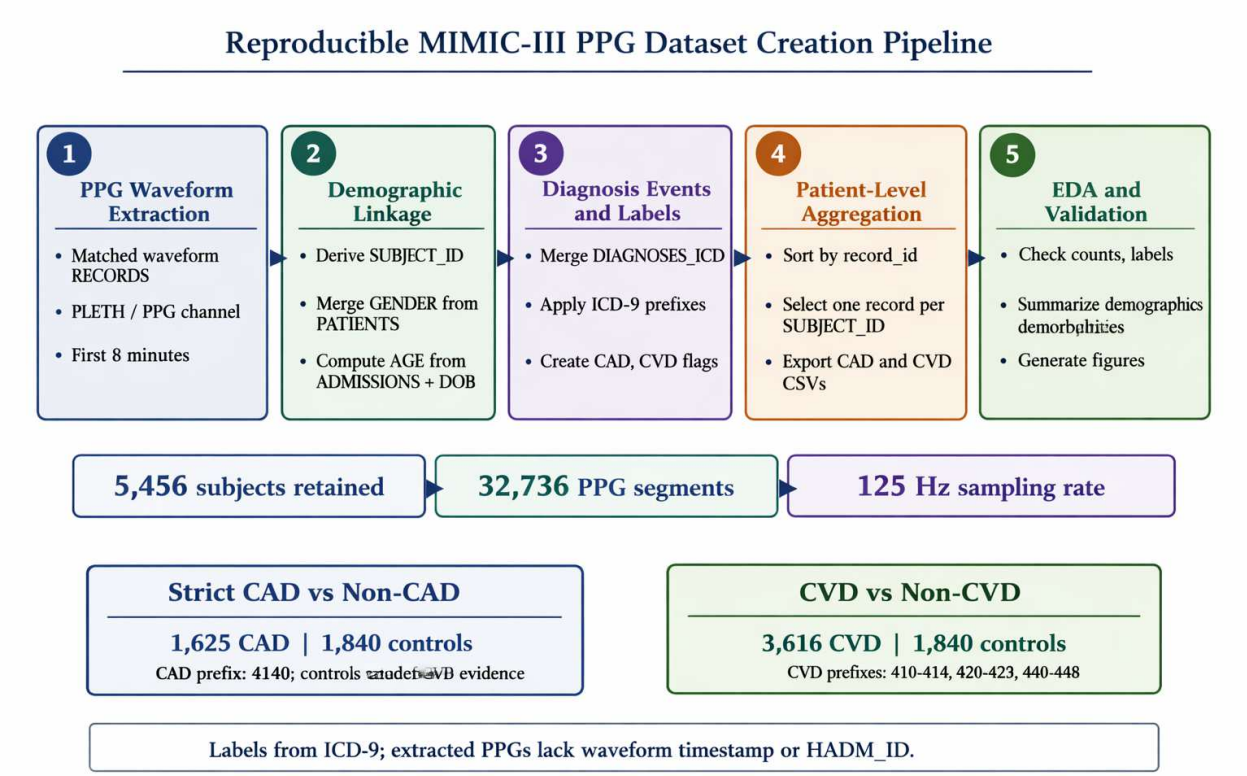


Figure 1: Five-step workflow for constructing patient-level CAD and CVD PPG datasets.

1. **Step 1: PPG waveform extraction.** Candidate PPG waveforms were extracted from MIMIC-III matched waveform records, quality checked, and stored as six one-minute windows.

2. **Step 2: Demographic linkage.** Waveform rows were linked to PATIENTS and ADMISSIONS using SUBJECT_ID to add gender and age.

3. **Step 3: Diagnosis events and labels.** DIAGNOSES_ICD was linked with ADMISSIONS to construct diagnosis events, apply ICD-9 prefix rules, and generate CAD, CVD, and comorbidity flags.

4. **Step 4: Patient-level aggregation.** Labeled waveform rows were reduced to one selected waveform record per SUBJECT_ID and exported as final CAD and CVD CSV files.

5. **Step 5: EDA and validation.** The final patient-level files were used to regenerate cohort summaries, validation checks, and manuscript figures.

3.2 Data Sources and Linkage Keys

The workflow used MIMIC-III matched waveform records and three MIMIC-III clinical tables. Table 2 summarizes the source files, fields, and linkage roles used for dataset construction.

Table 2: Data sources and linkage fields used in the dataset construction workflow.

Source	**Fields used**	**Role in workflow**
Matched waveform records	`patient_id`, waveform group, `record_id`, channel names, sampling frequency	PPG channel detection, waveform extraction, and waveform metadata.
PATIENTS	`SUBJECT_ID`, DOB, GENDER	Gender linkage and date-of-birth input for age calculation.
ADMISSIONS	`SUBJECT_ID`, `HADM_ID`, ADMITTIME	First-admission age calculation and admission-time linkage for diagnosis events.
DIAGNOSES_ICD	`SUBJECT_ID`, `HADM_ID`, `SEQ_NUM`, `ICD9_CODE`	CAD, CVD, and comorbidity label construction using ICD-9 prefix rules.

The waveform extraction output stored the waveform patient identifier as a string such as `p010023`. In Step 2, SUBJECT_ID was derived by removing the leading character and converting the remaining value to an integer. This SUBJECT_ID became the common key for merging waveform rows with clinical information.

3.3 Step 1: PPG Waveform Extraction

Waveform extraction was performed using the MIMIC-III matched waveform database (`mimic3wdb-matched/1.0`). The RECORDS index file was used to enumerate waveform patient folders and guide automated extraction. Waveform records were accessed using WFDB APIs [7].

For each candidate waveform record, the workflow searched the waveform header for the first channel whose name contained `PLETH` or `PPG`. If no such channel was found, the record was skipped. For records with a PPG channel, the workflow read the first eight minutes of

samples. Non-finite samples were removed, and records with fewer than eight minutes of finite samples were excluded.

The first two minutes were discarded, and minutes 3–8 were retained as a six-minute PPG segment. The quality checks were then applied to this retained segment. A record was excluded if the retained segment had standard deviation below 0.01 or peak-to-peak amplitude below 0.05. No separate outlier-fraction or 3σ artifact-rejection rule was implemented in the final extraction workflow.

Each accepted six-minute segment was split into six one-minute waveform columns:

`ppg_min_3`, `ppg_min_4`, `ppg_min_5`, `ppg_min_6`, `ppg_min_7`, `ppg_min_8`

The executed extraction and quality-control workflow retained 5,456 accepted PPG records. Each accepted record contained six one-minute waveform windows, giving 32,736 retained one-minute windows before task-specific cohort construction. Intermediate attempted-record counts and per-filter rejection counts were not retained in the final output files, so only final retained counts are reported here. Figure 2 illustrates an accepted waveform and a rejected waveform; the rejected example shows low variability and fails the variability criterion.

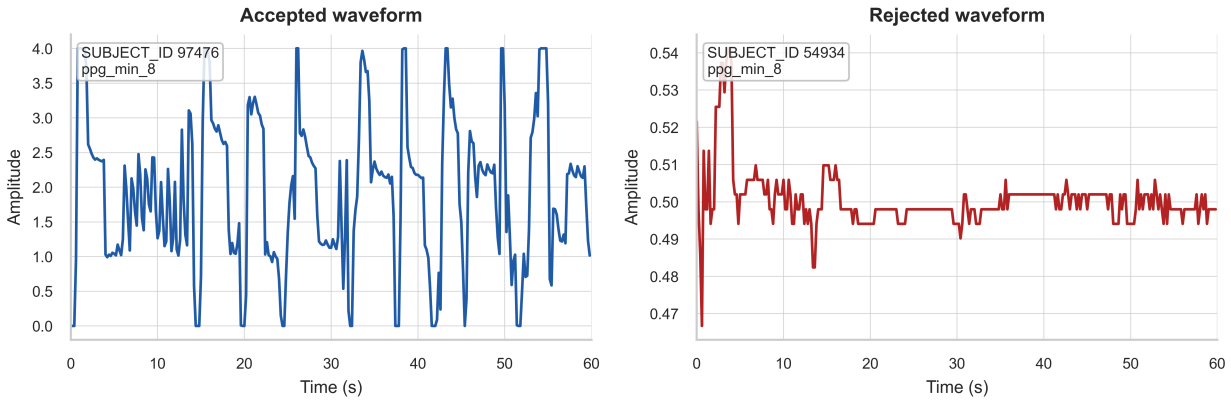


Figure 2: Representative accepted and rejected PPG waveforms.

3.4 Step 2: Demographic Linkage

Step 2 added demographic information to the extracted waveform rows using SUBJECT_ID. Gender was merged directly from PATIENTS using SUBJECT_ID. Age was computed by first selecting the earliest ADMISSIONS row for each SUBJECT_ID, then merging this first-admission time with DOB from PATIENTS. Age was calculated from first admission year and birth year.

Negative ages were set to missing. De-identified elderly ages greater than 300 years were mapped to 90 years during dataset preparation. The final released CSV files retain the age field generated during dataset preparation. For manuscript summaries and EDA plots, ages greater than 89 years were grouped as 90 years, following MIMIC-III de-identification conventions. Gender was encoded as Male/Female, and no unknown gender category was present in the final released CAD or CVD CSV files.

3.5 Step 3: Diagnosis Events, CAD/CVD Labels, and Comorbidities

Step 3 converted the raw diagnosis table into a subject-level source of cardiovascular evidence. Each row in DIAGNOSES_ICD represents one diagnosis assigned during a hospital admission and includes the patient identifier, admission identifier, diagnosis order, and ICD-9 code. The ADMISSIONS table was used to add the admission time for each diagnosis. The two tables were linked using SUBJECT_ID and HADM_ID. Diagnosis rows were used for label construction only when the patient identifier, ICD-9 code, and admission time were available.

The ICD-9 prefix rules used for label construction are shown in Table 3. Prefix matching was performed on the ICD9_CODE strings used in the workflow. The strict CAD rule used prefix `4140`. The broader CVD rule used prefixes `410`–`414`, `420`–`429`, and `440`–`448`. Comorbidity flags used prefixes `250`, `272`, and `278`.

Table 3: ICD-9 prefixes for CAD, CVD, and comorbidity labels.

Variable	ICD-9 Prefix	Meaning
Strict CAD	4140	Coronary atherosclerosis / coronary artery disease family
CVD	410–414	Ischemic heart disease, including myocardial infarction, angina, and coronary atherosclerosis
CVD	420–429	Other heart disease, including cardiomyopathy, conduction disorders, dysrhythmias, and heart failure
CVD	440–448	Diseases of arteries, arterioles, and capillaries
Diabetes	250	Diabetes mellitus
High cholesterol	272	Disorders of lipoid metabolism, high cholesterol, or hyperlipidemia
Obesity	278	Overweight, obesity, and hyperalimentation disorders

Strict CAD and Non-CAD labels. The workflow first identified the earliest CAD diagnosis time for each SUBJECT_ID using ICD-9 prefix 4140. A subject was labeled CAD-positive when CAD evidence was present. A subject was labeled strict Non-CAD only when no broader CVD evidence was present for that subject. Subjects that satisfied neither condition were treated as unresolved and excluded from the strict CAD dataset.

CVD and Non-CVD labels. The workflow identified the earliest broader CVD diagnosis time for each SUBJECT_ID using the CVD prefixes in Table 3. A subject was labeled CVD-positive when any broader CVD evidence was present. Subjects without such evidence were labeled Non-CVD. Unlike the strict CAD dataset, the CVD dataset was exhaustive and did not include an unresolved class.

Subject-level label assignment. The final extracted PPG rows contained waveform record identifiers but did not contain a usable waveform timestamp or HADM_ID. Therefore, the workflow could not directly match each PPG waveform to the exact hospital admission in which a diagnosis was recorded. CAD and CVD labels were assigned at the SUBJECT_ID level using the available ICD-9 diagnosis evidence for each patient. This makes the cohort construction reproducible, but it also means that a selected PPG waveform may not have been recorded during the same admission as the diagnosis used for labeling.

Comorbidity flags. Comorbidity flags were computed from the same diagnosis-event

table at SUBJECT_ID level. A subject was flagged as diabetic when any ICD9_CODE began with `250`, as having high cholesterol/hyperlipidemia when any code began with `272`, and as obese when any code began with `278`. These flags were merged back into the labeled PPG table using SUBJECT_ID. Missing comorbidity flags were filled as 0.

3.6 Step 4: Patient-Level Aggregation and Final Outputs

Step 4 converted labeled waveform rows into patient-level CAD and CVD datasets. The final patient-level CAD and CVD datasets contained no missing AGE or GENDER values. The preprocessing code nevertheless included median/mode fallback handling for these fields to support reproducible reruns if missing values occur.

Each final CAD and CVD CSV file contains one retained row per SUBJECT_ID. A deterministic safeguard was retained in the aggregation code to sort by record_id and keep the first row if duplicate subject rows occur during reruns.

3.7 Step 5: EDA and Dataset Validation

Step 5 regenerated the descriptive statistics, validation checks, and manuscript figures from the final patient-level CSV files. Figures 3–5 summarize class balance, age distribution, and sex/comorbidity prevalence in the final cohorts. The validation checks included:

1. **Patient-level uniqueness:** The final CAD and CVD CSV files were checked to ensure that each SUBJECT_ID appeared only once.

2. **Binary-label validation:** CAD, CVD, and comorbidity fields were checked for valid 0/1 values.

3. **Case-control consistency:** CAD and CVD case-control counts and the shared control cohort were checked.

4. **Waveform-window completeness:** retained rows were checked for the six expected one-minute PPG windows.

5. **EDA outputs:** missingness summaries, demographic summaries, comorbidity summaries, class-balance plots, and accepted/rejected waveform examples were regenerated from the final datasets.

These procedures were implemented to reduce information leakage, document signal-quality behavior, and improve reproducibility. Summary results from validation procedures are provided in Section 4, and the detailed reproducibility checklist is provided as supplementary material.

3.8 Ethics, Data Access, and Reproducibility

This study used de-identified publicly available data from PhysioNet MIMIC-III. Access to MIMIC-III waveforms and clinical records is granted after completion of required credentialing and data-use agreements. No new human participant data were collected. The associated repository provides the code used to implement the extraction, preprocessing, labeling, patient-level aggregation, and EDA workflow. All data processing was performed using Python 3.10.x with `pandas`, `numpy`, `scipy`, and `wfdb`. Complete environment specifications are provided in the repository requirements file. A practical execution environment requires sufficient local storage for MIMIC waveform files; the workflow was designed with resume-safe extraction to support large-scale processing.

4 Dataset Description

The final patient-level files combine identifiers, demographic variables, ICD-9-derived labels, comorbidity indicators, waveform metadata, and six retained PPG windows per subject. Table 4 summarizes the main variable groups provided for downstream reuse.

Table 4: Main variable groups in the released datasets.

Variable group	Variables	Description
Identifiers	`subject_id`, `record_id`	Patient identifier and selected waveform record identifier used for provenance and patient-level aggregation.
Demographics	`age`, `gender`	Patient-level age and binary sex/gender field derived from MIMIC-III clinical tables.
Disease labels	`is_cad`, `is_cvd`	Binary subject-level CAD and CVD labels generated from ICD-9 prefix definitions.
Comorbidities	`is_diabetic`, `has_high_cholesterol`, `is_obese`	Binary comorbidity indicators derived from ICD-9 diagnosis prefixes.
PPG windows	`ppg_min_3`–`ppg_min_8`	Six retained one-minute PPG waveform windows extracted after discarding the first two minutes.
Waveform metadata	waveform group, sampling frequency, source record fields	Metadata supporting waveform provenance, quality-control review, and reproducible downstream analysis.

4.1 Strict CAD Cohort

The final strict CAD dataset included 3,465 unique patients comprising 1,625 CAD-positive subjects and 1,840 Non-CAD control subjects (Table 5).

Table 5: Strict CAD cohort characteristics.

Variable	CAD+ (n=1,625)	Non-CAD (n=1,840)
Demographics		
Age (years)	70.8 ± 12.6	55.3 ± 17.1
[95% CI]	[70.2–71.4]	[54.5–56.1]
Male sex, n (%)	1,031 (63.5%)	919 (50.0%)
Diabetes, n (%)	724 (44.6%)	370 (20.1%)
High cholesterol, n (%)	1,025 (63.1%)	446 (24.2%)
Obesity, n (%)	149 (9.2%)	123 (6.7%)
Waveform Characteristics		
Available PPG windows per patient	6	6
Sampling frequency (Hz)	125	125

The CAD cohort shows significantly higher mean age in CAD-positive subjects (70.8 ± 12.6 years [95% CI: 70.2–71.4]) than in controls (55.3 ± 17.1 years [95% CI: 54.5–56.1]; Welch's t-test $t(3357) = 30.6$, $p < 0.0001$, Cohen's $d = 1.02$). The large effect size (Cohen's

$d > 0.8$) indicates substantial age separation between groups, which is expected for CAD-positive status associated with advanced age. Age statistics in this table use the EDA reporting variable in which ages greater than 89 years are grouped as 90 years. The released patient-level files include sex and comorbidity variables in addition to the raw age field. Diabetes, high cholesterol, and obesity were more prevalent among CAD-positive subjects than controls, consistent with known cardiovascular risk-factor patterns.

4.2 Implemented CVD Cohort

The implemented CVD dataset included 5,456 unique patients comprising 3,616 CVD-positive subjects and 1,840 Non-CVD control subjects (Table 6).

Table 6: Implemented CVD cohort characteristics.

Variable	**CVD+ (n=3,616)**	**Non-CVD (n=1,840)**
Demographics		
Age (years)	68.4 ± 14.8	55.3 ± 17.1
[95% CI]	[67.9–68.9]	[54.5–56.1]
Male sex, n (%)	2,071 (57.3%)	919 (50.0%)
Diabetes, n (%)	1,338 (37.0%)	370 (20.1%)
High cholesterol, n (%)	1,725 (47.7%)	446 (24.2%)
Obesity, n (%)	347 (9.6%)	123 (6.7%)
Waveform Characteristics		
Available PPG windows per patient	6	6
Sampling frequency (Hz)	125	125

Compared with the strict CAD cohort, the implemented CVD cohort demonstrated larger population coverage (5,456 vs. 3,465 patients) and class imbalance toward cardiovascular disease cases (66.3% CVD-positive). The CVD-positive group was significantly older (68.4±14.8 years [95% CI: 67.9–68.9]) than controls (55.3±17.1 years [95% CI: 54.5–56.1]; Welch's t-test $t(3264) = 27.8$, $p < 0.0001$, Cohen's $d = 0.84$). The large effect size reflects clinically meaningful age differences between CVD-positive and control groups. Diabetes, high cholesterol, and obesity were also more common in the CVD-positive cohort, reflecting the implemented cardiac and arterial phenotype integrated from ICD-9 diagnosis codes.

4.3 Signal Characteristics

The generated datasets contain six one-minute waveform windows per accepted recording, with fixed 125 Hz sampling and 60-second segment duration. Across the two released task-specific patient-level files, 8,921 rows represent 5,456 unique patients, and the combined file structure preserves the waveform windows, demographic fields, and labels needed for downstream analysis.

4.3.1 EDA Visualizations

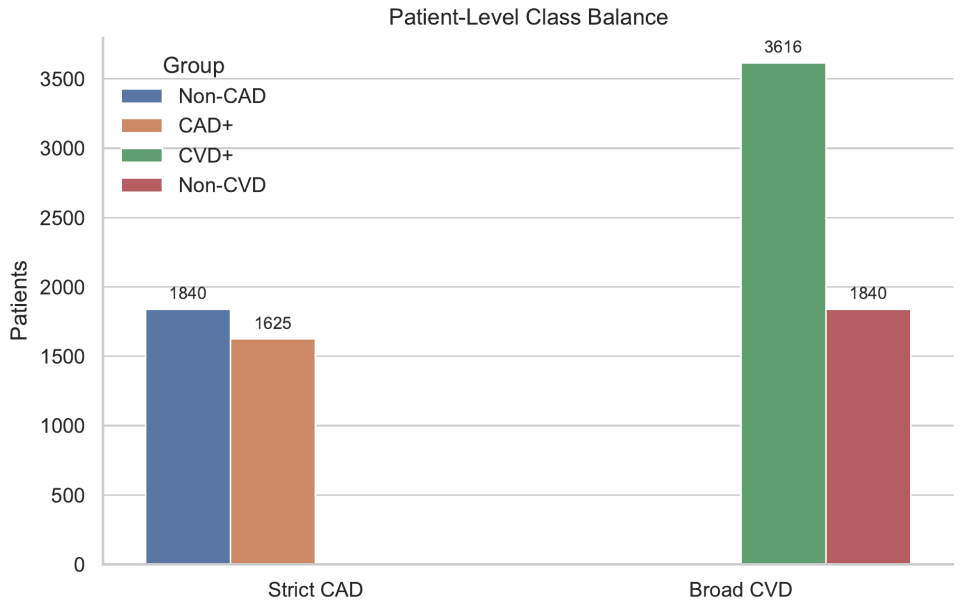


Figure 3: Patient-level class balance for the strict CAD and implemented CVD datasets.

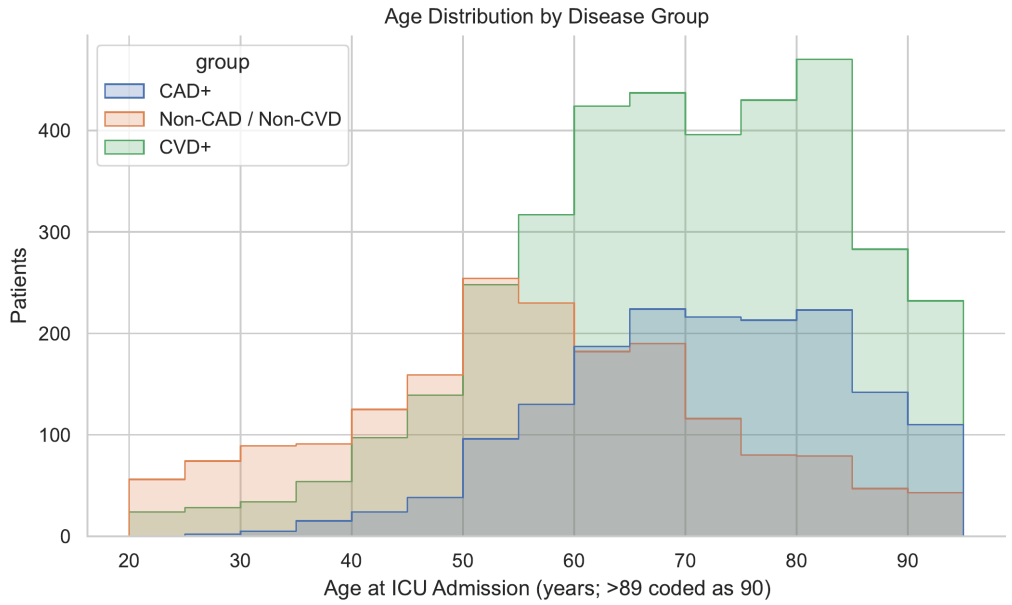


Figure 4: Age distributions for CAD-positive, CVD-positive, and shared control groups.

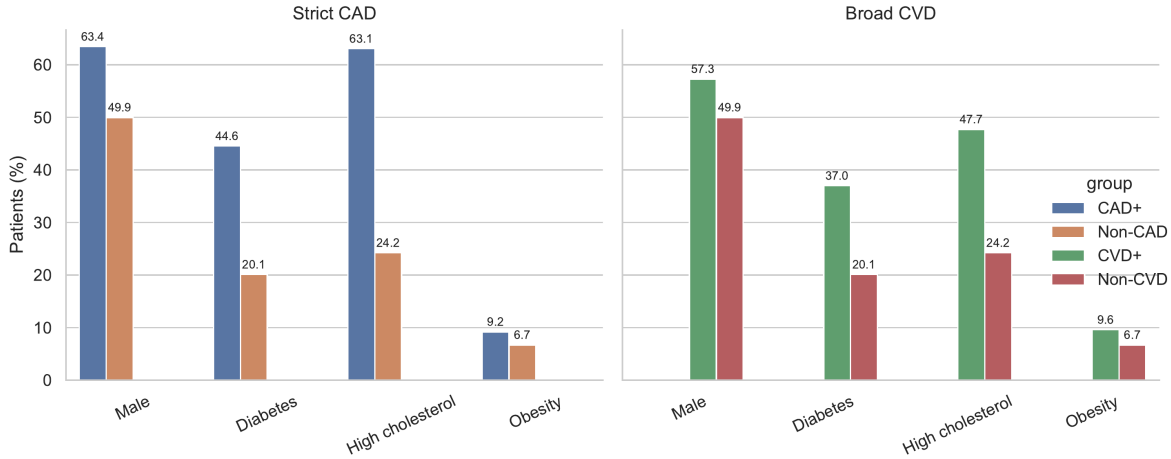


Figure 5: Sex and comorbidity prevalence by task and disease group.

5 Discussion

This study provides a patient-level PPG dataset resource for CAD and cardiac/arterial CVD research using the matched waveform and clinical components of MIMIC-III. The main contribution is not a predictive model, but a documented cohort-construction workflow that combines PPG waveform windows, demographic variables, comorbidity indicators, and ICD-

9-based cardiovascular labels in release-ready CSV files.

The two released tasks provide complementary cohort definitions. The strict CAD dataset emphasizes specificity by retaining subjects with ICD-9 evidence of coronary atherosclerosis and excluding unresolved cardiovascular cases from the CAD control group, whereas the CVD dataset uses broader cardiac and arterial ICD-9 prefixes to provide greater cardiovascular phenotype coverage. Both final task-specific files are organized at patient level with one retained waveform record per SUBJECT_ID, which supports leakage-aware analysis by preventing multiple waveform windows from the same patient from appearing as independent subjects, but does not capture within-patient variability across multiple waveform records.

The EDA results show clinically plausible age and comorbidity differences between disease-positive and control groups. These differences are expected in retrospective cardiovascular cohorts and should be considered during downstream modeling or benchmarking. Users of the dataset should therefore report patient-level splitting, evaluate possible confounding by age and sex, and avoid interpreting model performance as disease-specific signal evidence without appropriate adjustment.

These findings should be interpreted in the context of the dataset design. The resource favors transparent cohort construction and leakage-aware patient organization over exhaustive temporal phenotyping. Future extensions could improve temporal specificity and broaden utility by adding timestamp-resolved waveform linkage, clinician-reviewed phenotypes, external validation, and additional physiological modalities.

6 Limitations

This study has several limitations. First, the datasets were derived from a single-center retrospective critical-care population, which may limit generalizability to broader outpatient or community settings. Second, cardiovascular labels were generated using ICD-9 coding and admission linkage procedures, which may introduce retrospective coding-related uncertainty. Third, the extracted PPG rows did not include explicit waveform timestamps or HADM_ID

values, preventing direct verification that each selected waveform was temporally aligned with the admission in which a diagnosis was coded. Finally, waveform quality-control filtering may preferentially retain cleaner physiological recordings and thereby introduce selection bias.

7 Conclusion

This study presents patient-level CAD and cardiac/arterial CVD PPG datasets derived from the matched waveform and clinical components of MIMIC-III. The released files combine quality-screened PPG windows, demographics, comorbidity indicators, and ICD-9-based labels in one-record-per-subject CSV files. By documenting the extraction, labeling, aggregation, and EDA workflow, the dataset provides a reproducible resource for cardiovascular PPG signal-processing research.

Declarations

Ethics approval and consent to participate

This study used the publicly available de-identified MIMIC-III database. Access to MIMIC-III requires completion of the required training and data-use agreement through PhysioNet. No new human participant data were collected.

Consent for publication

Not applicable.

Availability of data and materials

The datasets described in this study were derived from MIMIC-III waveform and clinical records available through PhysioNet after credentialing and completion of the required data-use agreement. Processing code, EDA code, notebook documentation, and dependency information are available at `https://github.com/raghunathreddy-hub/CreateMimicPpg`. The workflow was implemented in Python and is intended to be rerun locally by users

with authorized MIMIC-III access. Raw MIMIC-III waveform and clinical files are not redistributed with this article or repository. Users who wish to rerun the extraction workflow must obtain authorized access to the original MIMIC-III records through PhysioNet and generate the derived outputs locally. Any sharing or redistribution of generated MIMIC-derived data should follow the applicable PhysioNet and MIMIC data-use terms.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

Raghunath Reddy conceptualized the study, developed the workflow, performed data curation and formal analysis, generated visualizations, and drafted and revised the manuscript. Heena Kauser supervised the study, contributed to methodology and project administration, and reviewed and edited the manuscript. Venkata Sumanth Rayasam contributed to software validation, data curation, visualization, and manuscript revision. All authors read and approved the final manuscript.

AI-assisted writing disclosure

During preparation of this work, the authors used AI-assisted writing tools to support language editing, organization, and consistency checking. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors acknowledge PhysioNet and the contributors of the MIMIC-III database for providing publicly accessible physiological waveform and clinical data resources that enabled this research.

References

[1] John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3):R1–R39, 2007.

[2] Giorgio Biagetti, Paolo Crippa, Laura Falaschetti, Leonardo Saraceni, Andrea Tiranti, and Claudio Turchetti. Dataset from ppg wireless sensor for activity monitoring. *Data in Brief*, 29:105044, 2020.

[3] Sara Campanella, Ayham Altaleb, Alberto Belli, Paola Pierleoni, and Lorenzo Palma. Ppg and eda dataset collected with empatica e4 for stress assessment. *Data in Brief*, 53:110102, 2024.

[4] Pak Hei Chan, Chun Kit Wong, Yee Chung Poh, Lin Pun, Wallace W. C. Leung, Yu Fung Wong, M. M. Y. Wong, D. W. S. Chu, and C. W. Siu. Diagnostic performance of a smartphone-based photoplethysmographic application for atrial fibrillation screening in a primary care setting. *Journal of the American Heart Association*, 5(7), 2016.

[5] Liangqiu Chen, Shiyong Li, Lei Liu, Guanfeng Huang, Yan Zhuang, Zhencheng Chen, Yongbo Liang, and Mohamed Elgendi. A four-wavelength photoplethysmography dataset for non-invasive hemoglobin assessment. *Scientific Data*, 13:564, 2026.

[6] Mohamed Elgendi. On the analysis of fingertip photoplethysmogram signals. *Current Cardiology Reviews*, 8(1):14–25, 2012.

[7] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, C.-K. Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[8] Nabil Ibtehaz and M. Saifur Rahman. Ppg2abp: Translating photoplethysmogram signals to arterial blood pressure waveforms using deep learning. *IEEE Journal of Biomedical and Health Informatics*, 24(7):1924–1934, 2020.

[9] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[10] Mohammad Kachuee, Mohammad Mahdi Kiani, Hoda Mohammadzade, and M. Shabany. Cuffless blood pressure estimation algorithms for continuous health-care monitoring. In *IEEE EMBC*, pages 859–862, 2017.

[11] Hanjun Lee et al. Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9:279, 2022.

[12] Yongbo Liang, Zhencheng Chen, Rabab Ward, and Mohamed Elgendi. Hypertension assessment using photoplethysmography: A risk stratification approach. *Journal of Clinical Medicine*, 7(10):360, 2018.

[13] Manuel Meier, Berken Utku Demirel, and Christian Holz. Wildppg: A real-world ppg dataset of long continuous recordings. *NeurIPS Datasets and Benchmarks Track*, 2024.

[14] Mohammad Moulaeifard, Marie Kutscher, Philip J. Aston, Peter H. Charlton, and Nils Strodthoff. Mimic-iii-ext-ppg, a ppg-based benchmark dataset for cardiovascular and respiratory signal analysis. *Scientific Data*, 13:668, 2026.

[15] Andrea Nemcova, Eniko Vargova, Radovan Smisek, Lucie Marsanova, Lukas Smital, and Martin Vitek. Brno university of technology smartphone ppg database (but ppg): Annotated dataset for ppg quality assessment and heart rate estimation. *BioMed Research International*, 2021:3453007, 2021.

[16] Sangjun Park, Dejiang Zheng, and Uichin Lee. A ppg signal dataset collected in semi-naturalistic settings using galaxy watch. *Scientific Data*, 12:892, 2025.

[17] Marco A. F. Pimentel, Alistair E. W. Johnson, Peter H. Charlton, Drew Birrenkott, Peter J. Watkinson, Lionel Tarassenko, and David A. Clifton. Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Transactions on Biomedical Engineering*, 64(8):1914–1923, 2017.

[18] Attila Reiss, Inken Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019.

[19] Mithun Saha, Maxwell A. Xu, Wanting Mao, Sameer Neupane, James M. Rehg, and Santosh Kumar. Pulse-ppg: An open-source field-trained ppg foundation model for wearable applications across lab and field settings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3), 2025.

[20] Philip Schmidt, Attila Reiss, Robert Duerichen, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 400–408, 2018.

[21] Weinan Wang, Pedram Mohseni, Kevin L. Kilgore, and Laleh Najafizadeh. Pulsedb: A large, cleaned dataset based on mimic-iii and vitaldb for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health*, 4:1090854, 2023.

[22] Jingye Xu, Yuntong Zhang, Wei Wang, Mimi Xie, and Dakai Zhu. A comprehensive ppg-based dataset for hr/hrv studies. *arXiv preprint arXiv:2505.18165*, 2025.

[23] Zhilin Zhang. Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction. In *IEEE Signal Processing Cup*, 2015.

[24] Yuwei Zheng et al. Tiny-ppg: A lightweight deep neural network for real-time detection of motion artifacts in photoplethysmogram signals on edge devices. *arXiv preprint arXiv:2305.03308*, 2023.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.zip