

Supplementary File 1: Reproducibility Checklist

Raghunath Reddy, Heena Kauser, Venkata Sumanth Rayasam

This supplementary checklist summarizes the key methodological decisions used to create the patient-level MIMIC-III PPG CAD and CVD datasets.

- ☐ **Data source:** MIMIC-III matched waveform database and clinical tables accessed through PhysioNet.
- ☐ **Credentialing:** Users rerunning the workflow require authorized MIMIC-III and PhysioNet access and must follow the applicable data-use agreement.
- ☐ **PPG channel identification:** Metadata-based search for channels containing “PLETH” or “PPG”.
- ☐ **Initial discard:** First 2 minutes of each recording removed.
- ☐ **Analysis window:** Minutes 3–8 retained as six 1-minute windows.
- ☐ **Quality filtering:** PPG-channel detection, minimum duration, finite-sample retention, variability threshold, and amplitude threshold.
- ☐ **Clinical inputs:** PATIENTS, ADMISSIONS, and DIAGNOSES_ICD tables used for demographics and diagnosis-derived labels.
- ☐ **CAD labeling:** Strict CAD label generated using ICD-9 prefix 4140.
- ☐ **CVD labeling:** Implemented CVD cohort generated using ICD-9 prefixes 410–414, 420–429, and 440–448.
- ☐ **Comorbidities:** Diabetes, high cholesterol/hyperlipidemia, and obesity derived from ICD-9 prefixes 250, 272, and 278.
- ☐ **Temporal linkage limitation:** Labels are assigned at subject level because extracted PPG rows do not contain waveform timestamps or HADM_ID values.
- ☐ **Patient aggregation:** One selected waveform record retained per subject after deterministic sorting by record identifier.
- ☐ **Validation:** Patient-level uniqueness, duplicate-subject handling, binary-label validity, case-control separation, control-cohort consistency, waveform-window availability, and reported cohort counts checked from the final CSV files.
- ☐ **EDA outputs:** Descriptive statistics and figures generated from the final patient-level datasets using the EDA notebook.