22HLT01 QUMPHY

D4

Report that at least 5 datasets using real and/or synthetic PPG data have been generated, that can be used to benchmark accuracy and uncertainty of supervised machine learning and deep learning models. This includes making the 5 reference problems and their respective 5 datasets available to the medical device and digital health communities.

Organisation name of the lead participant for the deliverable: THM

Due date of the deliverable: 31.07.2025 (M25)

Actual submission date of the deliverable: 31.07.2025 (M25)

Confidentiality Status: PU - Public, fully open

Deliverable Cover Sheet

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EURAMET. Neither the European Union nor the granting authority can be held responsible for them.

The project has received funding from the European Partnership on Metrology, co-financed from the European Union's Horizon Europe Research and Innovation Programme and by the Participating States.

European Partnership Co-funded by the European Union

METROLOGY PARTNERSHIP EURAMET

Contents

Summary

This report is part of the Qumphy project (22HLT01 Qumphy) that is funded by the European Union and is dedicated to the development of measures to quantify the uncertainties associated with Machine Learning algorithms applied to medical problems, in particular the analysis and processing of Photoplethysmography (PPG) signals. In this report, a list of six medical problems that are related to PPG signals and serve as Benchmark Problems is given. Suitable Benchmark datasets and their usage are described also.

1 Introduction

1.1 Photoplethysmography

Photoplethysmography (PPG) signals are rich in information and easy to measure passively without any physical or mental limitations of the subject. PPG is a widely used physiological sensing technigue. It consists of shining a light (often red, green or infrared) onto the skin and measuring the amount of light either reflected from, or transmitted through, a region of tissue.

This amount varies with each heartbeat and the PPG signals measure the fluctuations in blood volume which occur with each heartbeat, see Figure 1.

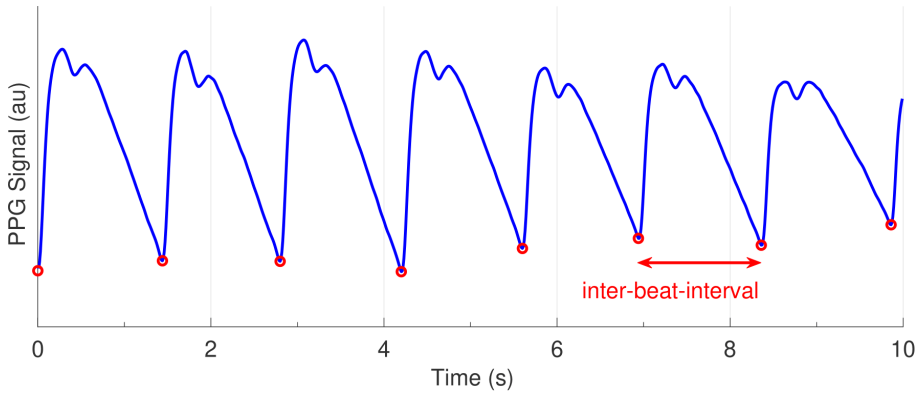

Figure 1: An exemplary photoplethysmogram (PPG) signal showing a pulse wave for each heartbeat. Pulse onsets, representing individual heartbeats, are shown as red circles. An inter-beat interval is labelled, corresponding to the time between consecutive heartbeats. Adapted from [1] by Peter H. Charlton - Own work, CC BY 4.0, https://commons.wikimedia.org/wiki/File:Photoplethysmogram_(PPG)_signal.png

Photoplethysmography (PPG) signals contain valuable information on the cardiovascular, respiratory, and autonomic nervous systems which is not yet routinely exploited. They are popular as they are easy to obtain non-invasively and PPG devices are cheap and widely available.

Until today however, an algorithmic evaluation of PPG signals to infer physiological parameters or detect diseases is crucial for saving patients' lives, but almost never used in clinical environments. One of the major reasons for this is the lack of trust in the output of any such algorithms.

Due to the vast amount of collected data, machine learning methodologies are essential for the extraction and evaluation of key features used for diagnosis. When applying machine learning in a medical context, however, confidence in the performance and predictions of the algorithms is particularly crucial since diagnostic mistakes can be fatal (false negative) or result in unnecessary anxiety and detrimental overtreatment (false positive). Hence an analysis of the uncertainty associated to ML algorithms and their predictions is indispensable to provide critical information about the quality and trustworthiness of the results produced.

1.2 The Qumphy project

The goal of the Qumphy project that is to satisfy these needs by developing an environment, i.e., a good practice guide including a software framework for independent assessment of accuracy and uncertainty of ML algorithms and benchmark cases to test and compare ML algorithms against, to increase trust in ML applications for PPG signals and lay a foundation towards standardisation of ML in healthcare. This project (22HLT01 QUMPHY) has received funding from the European Partnership on Metrology, co-financed from the European Union's Horizon Europe Research and Innovation Programme and by the Participating States. Funding for the UK partners was provided by Innovate UK under the Horizon Europe Guarantee Extension.

1.3 Aim of the report

In this report we report that more than 5 datasets using real and/or synthetic PPG data have been generated that can be used to benchmark the accuracy and uncertainty of supervised machine learning and deep learning models. This includes making six reference problems and their respective datasets available to the medical device and digital health communities.
The following six reference problems are chosen as benchmark problems:

1 Determine systolic and diastolic blood pressure

2 Detection of Atrial Fibrillation

3 Classification of Hypertension

4 Determine vascular age

5 Detection of Sleep Apnea

6 Determine respiratory rate

Our selection is based on the following aspects that we considered:

(i) How important is the problem to society?

(ii) Are there alternative methods not involving machine learning that are already established and sufficient?

(iii) How widely has the problem been tackled?

(iv) Are there sufficiently large, open datasets available?

(v) Is there a plausible physiological mechanism by which one could estimate the parameter from a PPG signal?

(vi) Is the data annotated?

(vii) How accurately can the ground truth be quantified?

In the following, we give a description of each individual benchmark problem together with their benchmark datasets and explain the possibilities to use them.

2 Benchmark I: Determine systolic and diastolic blood pressure

2.1 The problem

Blood pressure (BP) is one of the most commonly taken physiological measurements as it can be used to monitor cardiac health and predict adverse cardiac events, and is essential for the selection and monitoring of antihypertensive (BP lowering) treatments [2]. Blood pressure consists of two measurements, namely the peak systolic blood pressure (SBP) and the trough diastolic blood pressure (DBP) (see Figure 2). These are most commonly measured using an inflatable cuff, which has many limitations and which precludes continuous monitoring. Accurate blood pressure estimation from PPG signals is a regression machine learning problem and would have the advantage of generating a continuous readout of blood pressure, thus providing much extra, valuable information for the clinician.

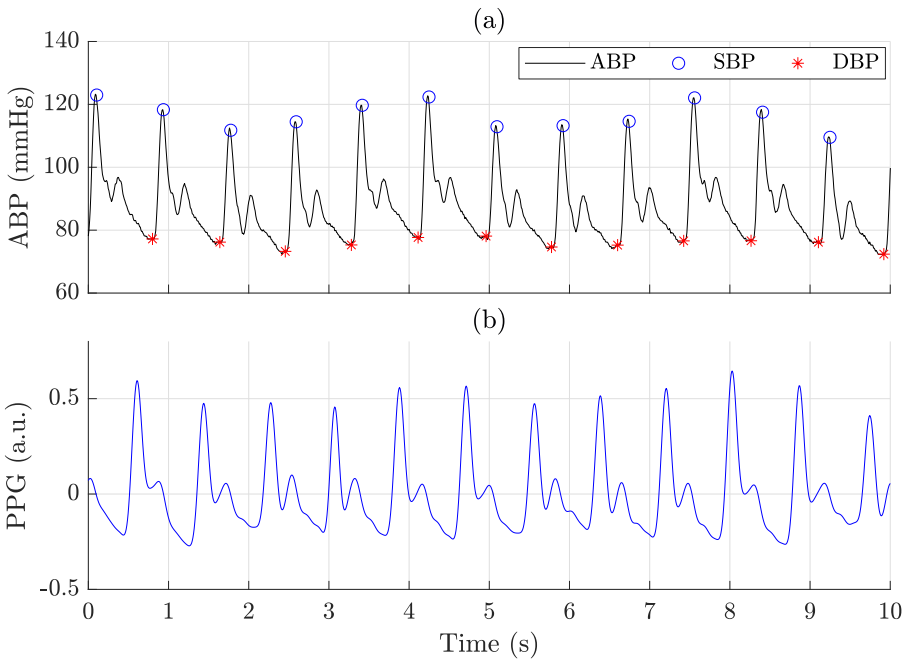

Figure 2: The example of (a) arterial blood pressure (ABP) segment with labeled fiducial points (systolic (SBP) and diastolic (DBP) blood pressure) and (b) PPG segment.

2.2 Potential datasets

This task requires datasets which include PPG segments together with (segment averaged) systolic and diastolic blood pressure measurements. There are quite a few such datasets either publicly available, available on request, or private. The main disadvantage of many of these, from a machine learning perspective, is that they are relatively small.

Another important factor to consider is the setting for collection of the data. The largest datasets are typically collected in clinical settings as it is relatively easy to collect data from patients in hospital. The disadvantage of such datasets is that these are sick patients who are undergoing medical interventions and are often on medication. One such dataset is MIMIC-III [3] which has data collected from over 40,000 patients who were in intensive care. We note that an even larger dataset, MIMIC-IV is now available [3]. The PPG signals in these datasets are generally collected from a pulse oximeter placed on a finger. Alternatively, data can also be collected in a community setting from generally healthy subjects, although of course some of these will have ongoing medical issues and be on medication. Such data is generally collected from a smartwatch on the wrist (or similar wearable device). One of the largest such datasets is Aurora BP [4] which contains data from over 1,000 participants and was collected with the aim of improving cardiovascular monitoring.

The length of signals available is another factor. The UK Biobank [5] has data available from over 200,000 subjects but the PPG signal is just a single beat, which is not ideal for machine learning. It is also not a free dataset. On the other hand, the signals in the MIMIC-III dataset range in length from minutes to several days, from which many shorter segments can be extracted.

We will consider in detail two datasets, namely Aurora BP [4], which is a large dataset collected in a community setting, and VitalDB [6], which is also a large dataset collected from surgery patients.

2.2.1 Aurora BP

The **AuroraBP** dataset [4] consists of PPG (and other) signals together with simultaneously collected auscultatory or oscillometric measurements of systolic and diastolic blood pressure. Measurements were made in a lab over a 24 hour period and some subjects also had ambulatory measurements. There were 1,125 particpants in the study (ages 21-85, average age 45.1 ± 11.3, 49.2% female, multiple hypertensive categories).

Fitzpatrick skin tone class is provided for 823 participants. However, the distribution of skin tones is highly imbalanced, with over 90% of classes 1, 2 or 3 and less than 1% of class 6 [7]. The PPG sensor was placed on the anterior surface of the arm (underside), in contrast to smartwatches which are worn on the posterier surface of the arm, and signals have been resampled to a frequency of 500 Hz.
The dataset contains 38,623 waveform records of varying durations. The average signal length was 19.73 s and the range was 9.1 s−87.1 s.
Requests for the data have to be made to the Data Access Committee and should include a research project description and details of the investigators.
The AURORA BP dataset has already been used in the following state-of-the-art studies [8, 9].

2.2.2 Vital DB

The PulseDB dataset [6] is a selection of high quality segments from two datasets, namely VitalDB (2,938 subjects) and MIMIC-III Waveform Database Matched Subset (2,423 subjects). The **PulseDB Vital** subset consists of 2,938 non-cardiac surgery patients (2,938 structured MAT files), containing raw and filtered synchronized physiological signals: (i) electrocardiogram (ECG), (ii) finger photoplethysmogram (PPG), and (iii) reference invasive blood pressure measurements. The PPG signals were collected from the fingertip, which poses a drawback since the majority of wearables utilize reflection PPG sensors positioned on the upper part of the wrist. However, the newest type of wearables—smart rings—can record PPG signals from the finger [10] and may still benefit from the results obtained from fingertip PPG recordings.
The PulseDB Vital subset includes patients during perioperative periods, which involve surgical operations (general, thoracic, urological, and gynecological surgery). The patients underwent routine or emergency operations at Seoul National University Hospital. The PulseDB Vital dataset provides information of subject age, height, weight, and the body-mass index (23.07 ± 3.53 kg/m^2). The average age of patients is 58.76 ± 15.02 years (54.73% male).
The PPG signals and continuous BP waveforms are sampled at a rate of 125 Hz. The signals are segmented into 10 s windows, providing a single average value of systolic and diastolic blood pressure for each segment. All 10 s segments related to PPG or BP signals with outliers or anomalies were removed. This means all aleatoric uncertainty was removed, and only epistemic uncertainty was left. The average duration of patient signals is 1.38 ± 1.18 h (range 10.08 s−9.16 h). Also, the PulseDB Vital data provides annotated fiducial points of ECGs (R peaks), PPGs (peaks and onsets) and BP measurements (peaks and onsets), which can be utilized to obtain beat-to-beat SBP/DBP sequences. In addition, version 2 of the Pulse DB Vital dataset provides the absolute time of each sample in the 10 s segment, and raw/filtered ECG and PPG signals with non-normalized absolute amplitudes. The PulseDB Vital dataset is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

2.3 Making the datasets available

2.3.1 Aurora BP

The Aurora BP dataset is not openly available. An application for access to the data should be made to the Data Access Committee. The process is described here.

2.3.2 VitalDB

The PulseDB dataset, which includes the VitalDB dataset, can be downloaded from here. After downloading, you can either proceed with the original instructions involving MATLAB and Python codes to generate the .npy files or implement a few adjustments to create lighter files tailored to the QUMPHY project's objectives. The adjustments are outlined in the D1 report of the QUMPHY project and [11] as well in the accompanying repository.

2.4 Dataset usage

In the following, an introduction to how to use these datasets for the problem of blood pressure estimation from PPG signals is given.

2.4.1 Aurora BP

Matlab code is provided which splits the Aurora BP dataset into 10 folds. The training, validation, test and calibration sets can then be defined by the user from these 10 folds as appropriate, with a split of 7/1/1/1 folds being suggested respectively (or 8/1/1 if a calibration set is not required). Using the folds for cross validation is clearly also an option. All the records for each subject are contained in a single fold. The folds have been stratified to give similar distributions of the following classes:

- Gender: Male/female.
- Blood pressure class: Three classes are defined in terms of systolic blood pressure (SBP) and diastolic blood pressure (DBP), and in accordance with the 2024 ESC Guidelines [2] by
  - Non-elevated: SBP < 120 mmHg and DBP < 70 mmHg
  - Elevated: $120 \leq$ SBP < 140 mmHg or $70 \leq$ DBP < 90 mmHg
  - Hypertensive: SBP ≥ 140 mmHg or DBP ≥ 90 mmHg

  Note that the Aurora BP variables *baseline_sbp* and *baseline_dbp* are used for the SBP and DBP values [12]. These provide a single value for each subject, which then allowed classification into one of the 3 classes.
- Cardiovascular disease: The Aurora BP metadata contains information about a range of self-reported cardiovascular diseases. Two classes are used, namely no cardiovascular disease and cardiovascular disease if at least one disease is reported.
- Body mass index (BMI): Three classes of BMI are used:
  - Healthy: BMI < 25 kg/m^2
  - Overweight: $25 \leq$ BMI < 30 kg/m^2
  - Obese: BMI ≥ 30 kg/m^2

2.4.2 VitalDB

The PulseDB dataset contains some metadata. The file `metadata.csv` comprises 1,494,474 rows and 20 columns, representing the combined MIMIC+VitalDB version of PulseDB. Key columns are as follows:

1. "source": Indicates the dataset origin (1 for the VitalDB dataset, 0 for the MIMC dataset).
2. "bmi", "weight", "height": These metrics are available only for the VitalDB dataset.
3. "dbp_avg" and "sbp_avg": These represent the average diastolic and systolic blood pressures respectively, which can be used as prediction targets. For convenience, we also provide ["dbp_avg","sbp_avg"] in the column "label" as prediction target.
4. "set": Differentiates the data into five distinct categories (see Table 4 of [6]):
   - set = 0 for train set
   - set = 1 for calibration based testing set
   - set = 2 for calibration-free testing set
   - set = 3 for AAMI calibration set
   - set = 4 for AAMI testing set
5. "set_calib", "set_calibfree," and "set_aami": Provides train/validation/calibration/test splits for the three scenarios (where X is "calib", "calibfree" or "aami"):
   - set_X=0 for training
   - set_X=1 for validation

- set_X=2 for calibration
- set_X=3 for testing

Usage example
To train a model on the VitalDB subset in the calibration free scenario, the following steps are required:

1. Load the file `metadata.csv`:

```
import pandas as pd
 #Load the CSV file
df = pd.read_csv('metadata.csv')
```

2. Select the desired indices based on the entries in the column "set_calibfree" with source=1 (VitalDB):

   - Train on entries with set_calibfree=0

     ```
     indices_train = df[(df['set_calibfree'] == 0) & (df['source'] == 1)].index
     ```

   - Validate on entries with set_calibfree=1 (e.g. for hyperparameter tuning and/or model selection)

     ```
     indices_val = df[(df['set_calibfree'] == 1) & (df['source'] == 1)].index
     ```

   - Calibrate the model uncertainty with set_calibfree=2 (e.g. for calibration or conformal prediction)

     ```
     indices_cal = df[(df['set_calibfree'] == 2) & (df['source'] == 1)].index
     ```

   - Test model performance with set_calibfree=3

     ```
     indices_test = df[(df['set_calibfree'] == 3) & (df['source'] == 1)].index
     ```

3. The selected row numbers for the subsets identified in the previous step can be used to extract the corresponding signals from the signals.npy file. The columns sbp_avg and dbp_avg can be used as prediction targets:

   ```
   import numpy as np
   # Load the signal file.npy
   signals = np.load('signals.npy')
   # Extract signals corresponding to indices_train
   signals_train = signals[indices_train]
   ```

   Note that if uncertainty quantification is not of interest, the validation and calibration sets can be combined by using

   ```
   set_calibfree=1 & set_calibfree=2
   ```

   to select validation set samples.

3 Benchmark II: Detection of atrial fibrillation

3.1 Problem

Atrial fibrillation (AF) is a major health and economic concern, reaching epidemic levels. More than 33 million people worldwide are diagnosed with AF, and this number is projected to double by 2050 [13]. Paroxysmal AF can be asymptomatic and difficult to detect, making early diagnosis crucial to prevent severe outcomes like ischemic brain stroke [14]. AF detection remains challenging due to its paroxysmal and sometimes brief episodes [15]. Opportunistic pulse palpation followed by a 12-lead ECG is a recommended screening method for individuals over 65 but is ineffective for asymptomatic cases, as it is performed only when symptoms are

reported. Therefore, developing technologies for screening high-risk populations in a convenient and affordable way is essential [16].

Traditional monitoring methods, such as ambulatory ECG monitors and cardiac event recorders, can be uncomfortable due to skin irritation and disruption to daily life, especially when long-term monitoring over several weeks is required to detect asymptomatic paroxysmal AF. Recent advances in wearable technology have introduced more convenient and cost-effective alternatives, such as PPG acquisition using a built-in camera of a smartphone [17], a web camera [18], earlobe sensor [19], or a smart wristband [20].

Finger- and wrist-based PPG face challenges from artifacts [21] caused by sensor displacement, forearm and hand motion, and poor contact, leading to missed episodes and false alarms. Additionally, irregular rhythms such as premature atrial contractions, bigeminy, atrial flutter, and tachycardia contribute to false positives [20, 22]. Therefore, further research is needed to determine whether machine learning algorithms trained on raw PPG data or those incorporating feature extraction and signal quality evaluation are more effective in reducing uncertainty in AF detection.

The development of reliable AF detectors using PPG is hindered by the limited availability of high-quality labeled datasets. The lack of guidelines for arrhythmia interpretation in PPG necessitates alternative annotation strategies such as the simultaneous acquisition of ECG for AF verification. As a result, existing PPG datasets are either too small or contain inaccurate labels due to reliance on automated annotations rather than expert verification.

3.2 Potential datasets

AF PPG datasets are categorized into wrist-based and fingertip-based datasets for training and testing purposes.

Wrist-based datasets:

DeepBeat is the preferred dataset for training due to its large scale and existing code support, despite some label noise. The dataset comprises over 3336185 25-second PPG segments sampled at 32 Hz from 167 individuals. PPG signals were collected using a wrist-based wearable device in various conditions, including before cardioversion, during an exercise stress test, and in daily life. The dataset was restructured in [11] to ensure an equal AF/non-AF ratio across training, validation, calibration, and test sets eliminating redundancy from the original data split and providing a more reliable representation for model evaluation. The final distribution includes the following splitting pattern:

- Training set: 40253 AF samples from 44 subjects and 65489 non-AF samples from 54 subjects.
- Validation set: 5749 AF samples from 20 subjects and 9343 non-AF samples from 16 subjects.
- Calibration set 5753 AF samples from 20 subjects and 9392 non-AF samples from 15 subjects.
- Test set: 5746 AF samples from 19 subjects and 9,343 non-AF samples from 17 subjects.

The code to generate this final distribution is described in deliverable D1 and [11].

The **TriggersAF** dataset [23] includes data from patients diagnosed with paroxysmal AF, recruited from inpatient and outpatient wards of the Cardiology Department at Vilnius University Hospital Santaros Klinikos. Prior to participation, all patients provided written informed consent in accordance with the ethical principles of the Declaration of Helsinki, and the study received approval from the Vilnius Regional Bioethics Committee (Reference Number 158200-18/7-1052-557). The dataset comprises 133 subjects (48 female, 85 male), with an average age of 57.9±11.6 years and a BMI of 28.4±4.9. There are no cases of permanent AF, with 45 subjects experiencing AF and 88 classified as non-AF. The average duration of recordings is 6.9±1.1 days, capturing a total of 307 AF episodes, with an individual average of 10±19 episodes. The AF burden is

approximately 6.8±26%. ECG and PPG signals are sampled at 500 Hz and 100 Hz, respectively. Given its quality, TriggersAF is being evaluated for use as either a training set (whole samples) or a high-quality test set. An example of synchronized ECG and PPG signals from the TriggersAF dataset with atrial fibrillation as well as premature beats and tachycardia are shown in figures 3 and 4, respectively.

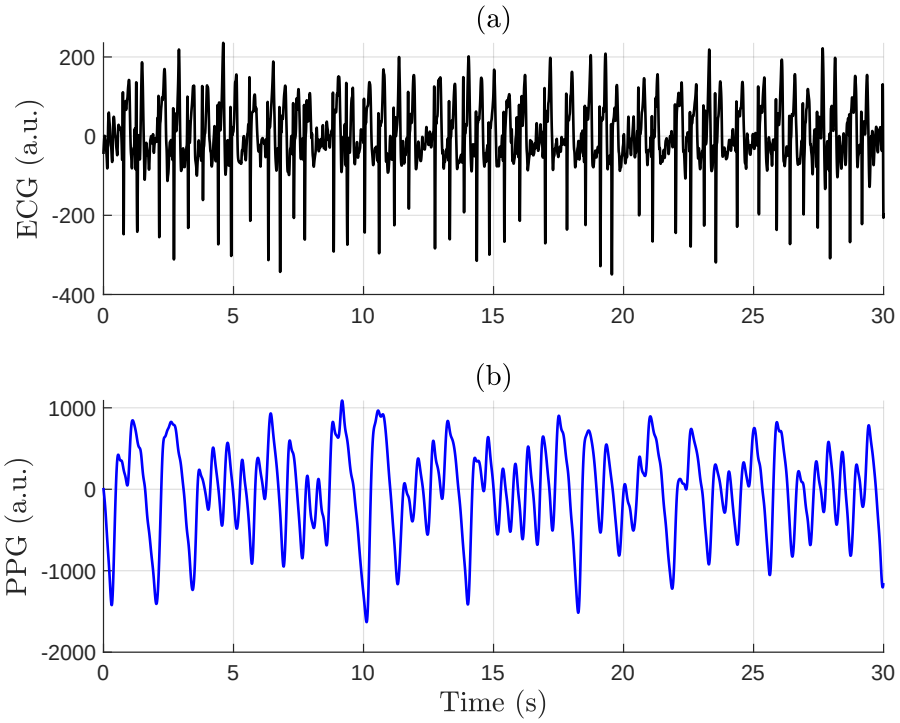

Figure 3: The example of (a) ECG segment and (b) PPG segment with atrial fibrillation from TriggersAF dataset.

The **UMMC Simband** dataset [24, 25] consists of 37 patients (28 male, 9 female) aged between 50 and 91 years, all diagnosed with cardiac arrhythmia. Participants wore the Samsung Simband 2 smartwatch, and their ECG was taken using a 7-lead Holter monitor. The data were preprocessed and segmented into 30-second windows with no overlap. The ECG signals were sampled at 128 Hz, while PPG signals were downsampled to 50 Hz. The signals were labeled into five categories: 0 (normal sinus rhythm), 1 (atrial fibrillation), 2 (premature atrial contractions/premature ventricular contractions), 3 (uncertain if NSR or PAC/PVC), and 5 (noisy PPG). For this study, only the PPG signals were used, and they were labeled as AF (1) or normal (0, 2, 3). Labels 5 and NaN (no ECG reference) were excluded from the analysis. The Simband dataset was discussed as a test set, providing additional validation for wrist-based AF detection models.

Fingertip-based datasets:

MIMIC-III-Ext-PPG is a new PPG-based benchmark dataset that has been created during the course of the QUMPHY project. It is a large-scale dataset based on the MIMIC III Waveform Database capturing almost five million 30s PPG waveform segments from more than 6000 patients. Rhythm annotations were inferred from heart rhythm chart events in the corresponding MIMIC-III clinical database [26]. It covers 26 different rhythm types, including atrial fibrillation and atrial flutter, which allows to use the dataset as a benchmark dataset for robust atrial fibrillation detection, using all other rhythm classes as negative examples. It is particularly suited for training due to its size and diverse rhythm types. A potential second MIMIC-III-based dataset [27] with manual AF annotations was used for technical validation of the label quality of the MIMIC-III-Ext-PPG dataset and is therefore not considered as a benchmark dataset.
Liu2022 The way of rhythm annotations were derived in MIMIC-III-Ext-PPG necessitates a second fingertip-

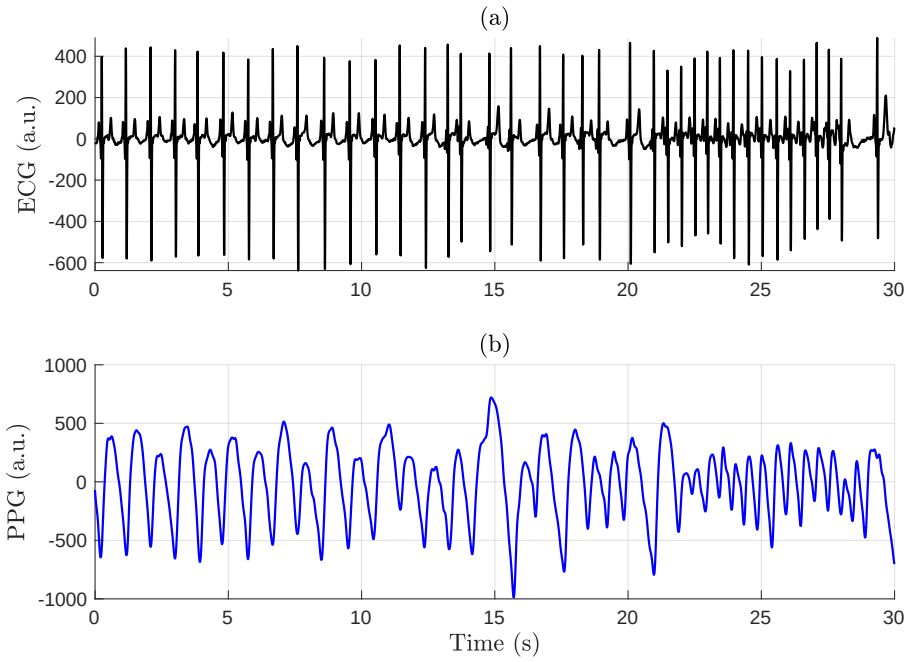

Figure 4: The example of (a) ECG segment and (b) PPG segment with premature beats and tachycardia from TriggersAF dataset.

based dataset for external validation. Here, we leverage the dataset prepared by Liu et al [28], which provides 91 10s PPG segments from 91 patients annotated according to five different rhythm types (sinus rhythm, premature ventricular contraction, premature atrial contraction, ventricular tachycardia, supraventricular tachycardia, atrial fibrillation). The dataset is not derived from MIMIC-III and therefore ideally suited as an external validation dataset for MIMIC-III-Ext-PPG.

3.3 Dataset selection

We recommend to use one large training set for wrist based measurements and one for fingertip measurements together with an additional test set in both situations. Concerning wrist based measurements, we chose DeepBeat as training set and TriggersAF as test set. For the fingertip measurements, we take MIMIC PPG as training set and the dataset from Liu as external test set.

3.4 Making the datasets available

The qumphy version of DeepBeat has already been described in the D1 report and [11]. Triggers AF is available on zenodo [23]. The MIMIC-III-Ext-PPG dataset is currently under review for publication on Physionet along with a corresponding dataset descriptor that is supposed to be submitted Scientific Data. The dataset will become publicly available once the review process is finished. The Liu2022 dataset is publicly available from an associated github repository [29]. The dataset is supposed to be used in its entirety as a test set.

3.5 Dataset usage

Wrist-based datasets:
In the previous report D1 of the qumphy project and [11], it was already described how to train models on DeepBeat with the following splitting pattern:
Training set: 40253 AF samples from 44 subjects and 65489 non-AF samples from 54 subjects.
Validation set: 5749 AF samples from 20 subjects and 9343 non-AF samples from 16 subjects.

Calibration set: 5753 AF samples from 20 subjects and 9392 non-AF samples from 15 subjects.
Test set: 5746 AF samples from 19 subjects and 9,343 non-AF samples from 17 subjects. This is supposed to be complemented by TriggersAF as external test set.
Fingertip-based datasets:
MIMIC-III-Ext-PPG comes with 13 predefined stratified splits that can be used for structured benchmarking. In particular, the final fold is supposed to be used as internal test set. The Liu2022 dataset will be used as external test set to assess the generalizability of models trained on MIMIC-III-Ext-PPG.

4 Benchmark III: Classification of hypertension

4.1 Problem

Elevated blood pressure and hypertension has a very high prevalence, i.e. around 30% of people in Europe are affected. It is one of the major risk factors for many diseases, in particular cardiovascular diseases. This includes Hypertension-mediated organ damage and hypertension is associated with cardiovascular and non-cardiovascular outcomes such as stroke, cognitive impairment, heart failure, atrial fibrillation or diabetes [2].

4.2 Potential datasets

4.2.1 AURORABP

As presented in Section 2.2.1, the **AuroraBP** dataset [4] consists of PPG, ECG and BP signals for two separate cohorts. What differentiates the two cohorts is the technique used for blood pressure measurement (auscultatory or oscillometric) and the presence or absence of ambulatory measurements. The study comprised of 1,125 participants aged 21-85, of which 49.2% were female.

4.3 Making the datasets available

The AURORABP dataset is available upon request and approval from the data regulatory committee, as described in Section 2.3.1.
The MatLab code that has been used to split the data into different classes and to preprocess the data will be made publicly available, to allow for reproducibility.

4.4 Dataset usage

Multiple records were present for each subject, and they are contained in a single fold. The folds have been stratified to give similar distributions of the following classes:

- Gender: Male/female.
- Blood pressure class: Three classes are defined in terms of systolic blood pressure (SBP) and diastolic blood pressure (DBP), and in accordance with the 2024 ESC Guidelines [2] by
  - Non-elevated: SBP $<$ 120 mmHg and DBP $<$ 70 mmHg
  - Elevated: 120 $\leq$ SBP $<$ 140 mmHg or 70 $\leq$ DBP $<$ 90 mmHg
  - Hypertensive: SBP $\geq$ 140 mmHg or DBP $\geq$ 90 mmHg

  Note that the Aurora BP variables *baseline_sbp* and *baseline_dbp* are used for the SBP and DBP values [12]. These provide a single value for each subject, which then allowed classification into one of the 3 classes.
- Cardiovascular disease: The Aurora BP metadata contains information about a range of self-reported cardiovascular diseases. Two classes are used, namely no cardiovascular disease and cardiovascular disease if at least one condition is reported.

- Body mass index (BMI): Three classes of BMI are used:
  - Healthy: BMI < 25 kg/m^2
  - Overweight: $25 \leq$ BMI < 30 kg/m^2
  - Obese: BMI ≥ 30 kg/m^2

Table 1: Population characteristics for the BP classification and regression cohort - HW: Healthy Weight, OW: Overweight, OB: Obese. Age, weight and height are expressed as mean ± standard deviation.

Total: 1100: 557M, 543F (49.4%F)						
	Non Elevated **213 (19.4%)** 73M / 140 F (65.7%F)		**Elevated** **645 (58.6%)** 337M / 308F (47.8%F)		**Hypertensive** **241 (21.9%)** 146M / 95F (39.4%F)	
CVD	**no cvd** 172 (76.5%)	**cvd** 53 (23.5%)	**no cvd** 374 (59.0%)	**cvd** 260 (41.0%)	**no cvd** 77 (32.0%)	**cvd** 164 (68%)
Gender (M/F)	51 / 114 (69.1% F)	22 / 26 (54.2% F)	189 / 191 (50.2% F)	148 / 117 (44.2% F)	53 / 24 (31.2% F)	93 / 71 (43.3% F)
BMI	HW 90 (54.6%) OW 46 (27.9%) OB 29 (17.6%)	HW 12 (25.0%) OW 14 (29.2%) OB 22 (45.8%)	HW 128 (33.7%) OW 143 (37.6%) OB 109 (28.7%)	HW 39 (14.7%) OW 94 (35.4%) OB 132 (49.8%)	HW 18 (23.4%) OW 27 (35.1%) OB 32 (41.5%)	HW 21 (12.8%) OW 52 (31.7%) OB 90 (54.9%) (1 NaN)
Age	38.2±10.28	45.81±11.65	43.06±10.31	48.46±10.39	46.85±12.47	51.35±10.58
Height (m)	1.68±0.09	1.71±0.11	1.71±0.10	1.72±0.10	1.75±0.09	1.71±0.10
Weight (Kg)	73.56±18.38	93.18±34.08	82.67±19.80	93.16±23.60	91.34±22.37	93.36±21.84

This large collection of data allows you to perform several versions of classification tasks between non-elevated, elevated and hypertensive patients and to analyse the influence of many risk factors. Matlab code that has been used to create stratified 10-fold splits for these tasks will be made publicly available.

5 Benchmark IV: Classification / regression vascular age

5.1 Problem

Vascular Ageing (VA) is a complex process that involves the gradual deterioration of arterial structure and function over time, negatively impacting organ function [30]. The gold-standard measurement for VA is carotid-femoral pulse wave velocity, which requires trained personnel and is not routinely clinically available [31]. In healthy ageing, chronological and vascular age typically correspond [32].
Early detection of premature VA is critical for the timely identification and treatment of cardiovascular disease (CVD), which remains a leading global health burden.
Non-invasive signals from photoplethysmography (PPG) or tonometry can help assess vascular age, by analysing the shape of the pulse wave, which changes with age due to arterial stiffening. PPG is an optical method used in clinical and wearable devices to measure pulse waves at sites like the wrist and finger

[33]. Arterial tonometry, mainly used clinically, measures pressure from superficial arteries such as the radial or carotid [34]. By comparing signal-based estimates of vascular age to a person's chronological age, we hypothesised we could identify early VA in community based settings.

5.2 Potential datasets

5.2.1 AURORABP

As presented in Section 2.2.1, the **AuroraBP** dataset [4] consists of PPG, ECG and BP signals for two separate cohorts. What differentiates the two cohorts is the technique used for blood pressure measurement (auscultatory or oscillometric) and the presence or absence of ambulatory measurements. The study comprised of 1,125 participants aged 21-85, of which 49.2% were female.

5.2.2 Pulse Wave Database (PWDB)

Another potential dataset is the Pulse Wave Database [35], a dataset of simulated waveforms that comprises of 4,374 healthy male subjects divided into six 10-year age groups, spanning from 25 to 75 years of age (25, 35, 45, 55, 65, 75). It contains single-cycle PPG signals, among others, at various locations in the body and is free to download and use.

5.3 Dataset selection

The AURORABP dataset was used for this task. Participants were separated into "suitable" and "unsuitable". Participants who were marked as "suitable" satisfied **all** of the following conditions: no cvd, no hypertension, and no obesity.
Participants were labeled as **hypertensive** if SBP was over 140 mmHg or DBP was greater than 90 mmHg, in agreement with the 2024 ESC guidelines [36]. They were labeled as **obese** if their BMI was ≥ 30, while they were assigned to the "CVD" (cardiovascular disease) if their records contained at least one of the following:

- high blood pressure (not highly reliable according to published study)
- coronary artery disease
- diabetes
- arrythmia
- previous heart attack
- previous stroke
- heart failure
- aortic stenosis
- valvular heart disease
- "other cv diseases" (not better specified)
- cvd meds (not better specified)

Participants were then separated into 10-year age groups:

- < 30
- 30-40
- 40-50
- 50-60

- 60-70
- 70+,

and an overview of the population composition is presented in Table 2.
Note: separation by gender was not accounted for because the gender distribution was uneven across ages, even though it might impact ppg waveforms. The limitations mentioned in the BP section remain valid also for the Age regression section.

Table 2: Population characteristics for the age regression cohort - HW: Healthy Weight, OW: Overweight, OB: Obese. Age, weight, height, SBP and DBP are expressed as mean ± standard deviation. SBP and SBP values referred here are the "baseline_sbp" and "baseline_dbp" values used to divide the population into BP classes.

Total: 546						
Age group	**<30** 57 (10.4%)	**30-39** 205 (37.5%)	**40-49** 146 (26.7%)	**50-59** 118 (21.7%)	**60-69** 16 (3.0%)	**70+** 4 (0.7%)
Height (m)	1.70±0.10	1.71±0.10	1.71±0.10	1.70±0.09	1.69±0.09	1.72±0.10
Weight (Kg)	70.41±15.21	82.49±21.44	81.59±19.65	79.42±18.65	70.32±9.28	75.40±16.39
SBP	114.16±10.55	117.34±10.53	118.38±10.11	118.90±10.49	125.49±11.66	131.66±5.68
DBP	67.03±8.80	71.82±7.94	73.62±8.22	74.25±6.86	72.63±7.46	67.01±4.46

5.4 Making the datasets available

The AURORABP dataset is available upon request and approval from the data regulatory committee, as described in Section 2.3.1.
The MatLab code that has been used to split the data into different classes and to preprocess the data will be made publicly available, to allow for reproducibility.
The PWDB dataset is free to download on Zenodo [37].

5.4.1 Dataset usage

The population was split into 10 separate folds, balanced for age and sex, that can then be used for train, validation, calibration and test.

6 Benchmark V: Detection of sleep apnea

6.1 The problem

Obstructive sleep apnea (OSA) is a prevalent sleep disorder characterized by repeated episodes of partial or complete upper airway obstruction during sleep, leading to disrupted breathing, oxygen desaturation, and fragmented sleep [38, 39]. Its clinical significance is significant, as untreated OSA is associated with a range of comorbidities, including hypertension, cardiovascular disease, stroke, type 2 diabetes, and neurocognitive impairments such as memory deficits and impaired executive function [40, 41]. In addition, OSA contributes to daytime fatigue, increasing the risk of car accidents and workplace errors [42]. The chronic physiological stress from recurrent apneas and hypopneas triggers systemic inflammation and oxidative stress, further exacerbating metabolic and cardiovascular complications. The condition's impact on quality of life, coupled with its economic burden due to healthcare costs and lost productivity, underscores the need for timely diagnosis

and effective management [43].

Detecting OSA poses significant challenges, primarily due to underdiagnosis and limited access to diagnostic tools. Polysomnography (PSG), the gold standard for diagnosis, is resource-intensive, requiring overnight monitoring in a sleep laboratory with specialized equipment and trained personnel [44]. This makes it costly and inaccessible for many patients, particularly in underserved or rural areas [45]. Home sleep apnea testing (HSAT) offers a more accessible alternative, but its sensitivity and specificity vary, and it may miss milder cases or misclassify complex sleep disorders [46]. Moreover, many individuals with OSA remain asymptomatic or attribute symptoms like snoring or daytime sleepiness to other causes, delaying presentation to healthcare providers. Public awareness of OSA is low, and screening tools, such as questionnaires (e.g., STOP-Bang), while useful, lack precision and can lead to over- or under-referral for testing [47, 48]. These detection barriers highlight the need for scalable, cost-effective diagnostic solutions, such as wearable devices and AI-driven screening tools, to improve early identification and treatment of OSA.

An apnea event is characterized as a lack of airflow for at least 10 s with an associated oxygen desaturation of at least 3-4% and/or arousals. A hypopnea event is characterized as a 50% or greater reduction in airflow for at least 10 s with an associated oxygen desaturation of at least 3-4% and/or arousal. Episodes of apneas and hypopneas in the airflow, PPG, heart rate (HR) and SpO2 signals are illustrated in the Figure 5.

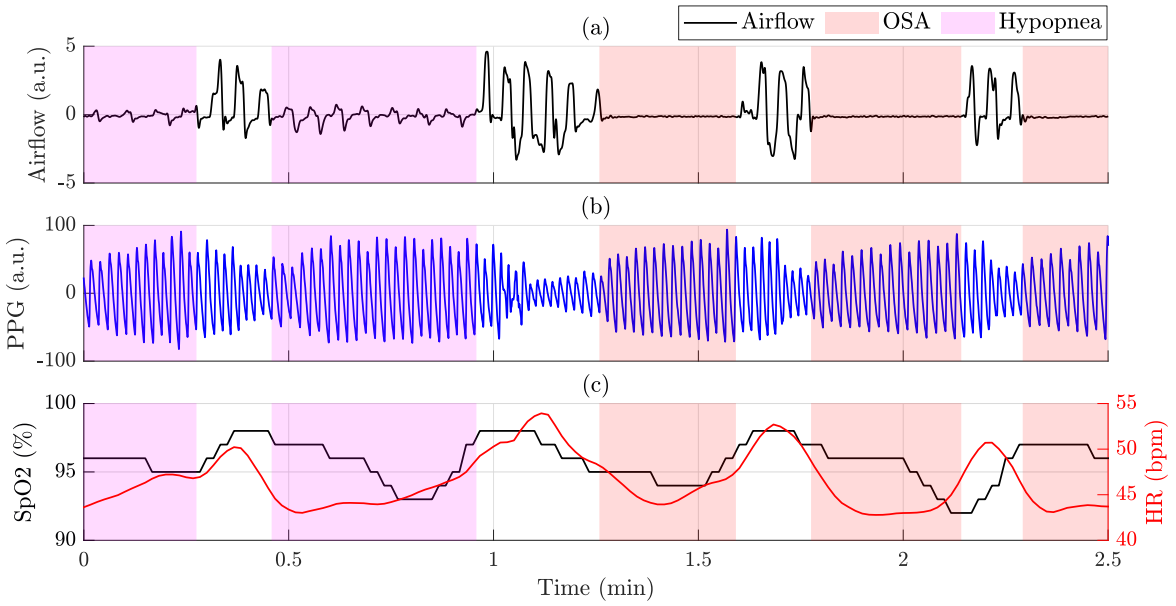

Figure 5: The changes of signals during sleep apnea episodes: (a) respiratory airflow, (b) finger PPG signal, (c) arterial blood oxygen saturation, SpO2, and heart rate, HR.

6.2 Potential datasets

6.2.1 OSASUD

The open-access **OSASUD** polysomnography (PSG) dataset [49] involves 30 patients (36.67% of female) admitted to the stroke unit at the Clinical Neurology Unit of Udine University Hospital between 2019-2020 due to suspected cerebrovascular events, such as ischemic and hemorrhagic stroke, or transient ischemic attack. The dataset provides information of subject age (68.97 $\pm$ 11.22 years), the body-mass index (29.17 $\pm$ 5.74 kg/m^2), the apnea-hypopnea index (25.43 $\pm$ 21.31 counts/h), and signal duration (8.91 $\pm$ 1.47 h). Patients of age <18 years, those unable to comply with standard requirements for PSG monitoring, individuals with severe aphasia impacting understanding or consent, and those at high risk of alcohol or drug withdrawal syndrome were excluded from the study. Notably, conditions such as obesity, diabetes mellitus, atrial fibrillation and other cardiac disease did not serve as exclusion criteria.

The dataset includes PSG signals such as finger photoplethysmography (PPG), electrocardiography (ECG), oxygen saturation (SpO2), respiratory rate, perfusion index, heart rate, nasal airflow, snoring, 3-axis accelerometer, abdominal and thoracic movements. The sampling rate of PPG and ECG signals is 80 Hz. While respiratory rate and SpO2 time series are sampled at 1 Hz. In addition, SpO2 values <50% or >100% were identified as artifacts and marked as null. This method was similarly applied to heart rate measurements <20 or >200, as well as respiratory rate values <5 or >40.

Recorded PSG data underwent thorough analysis using Embla RemLogic software, version 3.4.1.2371 (Natus Medical Inc., Pleasanton, CA, USA), which facilitates signal processing, detailed examination, and annotation procedure. The OSASUD data were annotated by a trained sleep medicine physician based on the sleep scoring guidelines established by the American Academy of Sleep Medicine, identifying occurrences of central, obstructive, mixed apnea, and hypopnea events (1 s granularity). The OSASUD dataset has already been used in the following state-of-the-art studies [50–53].

6.2.2 MESA

The **MESA** dataset [54] comprises data from 2055 patients (53.63% of female), aged between 54 and 95 years. It contains a total of 16 300 hours of fully annotated overnight PSG recordings collected during a sleep study (2010-2012), funded by the National Heart, Lung, and Blood Institute. The average age of subjects is 69.37 $\pm$ 9.12 years. Participants included in the study had not used treatments for sleep apnea, such as continuous positive airway pressure or oxygen devices, for more than a month prior to the study, or only used it less than once a week. The dataset represents a cohort, featuring participants from a variety of racial/ethnic groups, including White, African American, Hispanic, and Chinese American individuals.

PSG recordings were obtained at patient home using the Compumedics Somte monitoring system (Compumedics Ltd., Abbotsville, Australia). The dataset includes signals such as PPG, ECG, electrooculography-EOG, electroencephalography-EEG, electromyography-EMG, SpO2, nasal airflow, snoring, abdominal and thoracic movements. The sampling rate of PPG and ECG signals is 256 Hz. PPG signals were recorded from the finger using the Nonin 8000 sensor. In addition, time labels and durations of apnea and hypopnea episodes in respiratory flow signal are annotated. However, this database is not open-access and is only available on request. The MESA dataset has already been used in the following state-of-the-art studies [55, 56].

6.3 Making the datasets available

The OSASUD is fully open access, whereas the MESA is only available on request. The original OSASUD dataset can be downloaded from here. The MESA dataset is not freely available, and an application for access to the data should be made to the National Sleep Research Resource. This can be done here. The Matlab codes to extract 10 folds/splits from OSASUD and MESA datasets will be available at QUMPHY repository very soon. While the OSASUD is freely available dataset, *.mat and *.pkl derived files are additionally provided.

6.4 Dataset usage

In the following, an introduction to how to use these datasets for the problem of sleep apnea detection from PPG signals is given.

6.4.1 OSASUD

The reduced OSASUD dataset file (*OSASUD_initial*), the Matlab code (*Data_Segmentation_OSASUD*) which splits this OSASUD file into 10 folds, and these folds will be published very soon. The training, validation, calibration, and test sets can be defined by the user from these 10 folds as appropriate, with a split of 7/1/1/1 folds being suggested respectively (or 8/1/1 or 7/2/1 if a calibration set is not required). Using the folds for cross-validation is clearly also an option. All the records for each subject are contained in a single fold. These folds have been stratified to give similar distributions of apnea-hypopnea index (AHI) values.

The OSASUD consists of 30 subjects, so each fold consists of nocturnal raw PPG signals and estimated features from 3 subjects. Segment duration - 1 min (raw PPG segment - 4800 samples, PPG feature sequence - 60 samples).

Data_Segmentation_OSASUD code processes:

1. *OSASUD_initial* file in order to obtain *OSASUD_segments* file with 1-min raw PPG intervals and PPG feature sequences for model training/validation/calibration/testing.
2. *OSASUD_segments* file in order to obtain *OSASUD_PPG_feature_folds*, *OSASUD_PPG_segment_folds*, *OSASUD_Apnea_label_folds*, and *OSASUD_RRate_label_folds* files with 10 folds (10 groups of 3 subjects) stratified according to AHI labels.

OSASUD_initial file includes data of the table, whose columns are:

1. Subject ID;
2. SpO2 time series (sampling rate - 1 Hz);
3. Apnea event type (APNEA-OBSTRUCTIVE, HYPOPNEA, APNEA-CENTRAL, APNEA-MIXED, ALL APNEAS);
4. Anomaly (1 - if there is any apnea event, 0 - otherwise);
5. PPG signal (sampling rate - 80 Hz);
6. Respiratory rate (sampling rate - 1 Hz).

OSASUD_segments file includes estimated PPG features, raw PPG segments, and labels divided for 1-min 13,829 PPG intervals:

1. *X_cell_features* includes 4 features (1 - pulse interval, 2 - peak-to-peak amplitude, 3 - area-related feature, and 4 - SpO2) sampled at 1 Hz (segment duration - 60 samples) for 30 subjects.
2. *X_cell_ppg_segments* includes raw 1-min PPG segments sampled at 80 Hz (segment duration - 4800 samples) for 30 subjects.
3. *Y_cell* includes 6 category labels for 30 subjects:
   (a) APNEA-OBSTRUCTIVE events;
   (b) HYPOPNEA events;
   (c) APNEA-OBSTRUCTIVE & HYPOPNEA events;
   (d) APNEA-CENTRAL events;
   (e) APNEA-MIXED events;

(f) ALL APNEAS.

4. *Y_cell_rr* includes mean respiratory rate labels for 30 subjects.

OSASUD_PPG_feature_folds file includes 10 folds (10 groups of 3 subjects) of estimated PPG features (1 - pulse interval, 2 - peak-to-peak amplitude, 3 - area-related feature, and 4 - SpO2). PPG feature sequence duration - 60 samples (1-min).
OSASUD_PPG_segment_folds file includes 10 folds (10 groups of 3 subjects) of raw PPG segments. PPG segment duration - 4800 samples (1-min).
OSASUD_Apnea_label_folds file includes 10 folds (10 groups of 3 subjects) of apnea labels (1 - APNEA-OBSTRUCTIVE events, 2 - HYPOPNEA events, 3 - APNEA-OBSTRUCTIVE & HYPOPNEA events, 4 - APNEA-CENTRAL events, 5 - APNEA-MIXED events, 6 - ALL APNEAS) for 1-min PPG signals.
OSASUD_RRate_label_folds file includes 10 folds (10 groups of 3 subjects) of respiratory rate labels for 1-min PPG signals.

Note 1: we suggest using '3 - APNEA-OBSTRUCTIVE & HYPOPNEA events' label for detecting apnea segments.
Note 2: instead of using raw PPG segments, we suggest using estimated (a) 3 PPG features (pulse interval, peak-to-peak amplitude, area-related feature) and (b) SpO2 time series as model inputs, separately, and then compare the (a) and (b) results.
Note 3: all data - 30 subjects, 13,829 1-min segments.

6.4.2 MESA

The Matlab code (*Data_Segmentation_MESA*) which splits the MESA dataset into 10 folds will be published very soon. The training, validation, calibration, and test sets can be defined by the user from these 10 folds as appropriate, with a split of 7/1/1/1 folds being suggested respectively (or 8/1/1 or 7/2/1 if a calibration set is not required). Using the folds for cross-validation is clearly also an option. All the records for each subject are contained in a single fold. These folds have been stratified to give similar distributions of AHI values.

The original MESA consists of 2055 subjects. For signal quality and class balance issues, we used only 160 subjects with 'PSG Study Quality Grade = 7' (highest quality) and 'AHI > 25'. Thus, each fold consists of nocturnal raw PPG signals and estimated features from 16 subjects. Segment duration - 1 min (raw PPG segment - 15360 samples, PPG feature sequence - 60 samples).

Data_Segmentation_MESA code processes:

1. 2055 *.edf PPG signals in order to obtain *MESA_segments* file with 1-min raw PPG intervals and PPG feature sequences for model training/validation/calibration/testing, and *MESA_metadata* file with subject AHI values. For signal quality and class balance issues, *MESA_segments* and *MESA_metadata* files include information of only 160 subjects with 'PSG Study Quality Grade = 7' (highest quality) and 'AHI > 25'.

2. *MESA_segments* file in order to obtain *MESA_PPG_feature_folds*, *MESA_PPG_segment_folds*, and *MESA_Apnea_label_folds* files with 10 folds (10 groups of 16 subjects) stratified according to AHI labels.

MESA_segments file includes estimated PPG features, raw PPG segments, and labels divided for 1-min PPG intervals:

1. *X_cell_features* includes 4 features (1 - pulse interval, 2 - peak-to-peak amplitude, 3 - area-related feature, and 4 - SpO2) sampled at 1 Hz (segment duration - 60 samples) for 160 subjects.

2. *X_cell_ppg_segments* includes raw 1-min PPG segments sampled at 256 Hz (segment duration - 15360 samples) for 160 subjects.

3. *Y_cell* includes 3 category labels for 160 subjects:

(a) APNEA-OBSTRUCTIVE events;

(b) HYPOPNEA events;

(c) APNEA-OBSTRUCTIVE & HYPOPNEA events.

MESA_PPG_feature_folds file includes 10 folds (10 groups of 16 subjects) of estimated PPG features (1 - pulse interval, 2 - peak-to-peak amplitude, 3 - area-related feature, and 4 - SpO2). PPG feature sequence duration - 60 samples (1-min).
MESA_PPG_segment_folds file includes 10 folds (10 groups of 16 subjects) of raw PPG segments. PPG segment duration - 15360 samples (1-min).
MESA_Apnea_label_folds file includes 10 folds (10 groups of 16 subjects) of apnea labels (1 - APNEA-OBSTRUCTIVE events, 2 - HYPOPNEA events, 3 - APNEA-OBSTRUCTIVE & HYPOPNEA events) for 1-min PPG signals.

7 Benchmark VI: Regression respiratory rate

7.1 Problem

Respiratory rate (RR) gives much information of the physiological state [57, 58] and is the most sensitive vital sign marker that indicates clinical deterioration [59–62]. RR is therefore a highly informative indicator for the patient's health states and continuous monitoring of RR is desired.
On the other hand, it is known that physiological mechanisms cause ECG and PPG signals to be modulated by respiration. Three types of modulation can be observed, as seen in Figure 6:

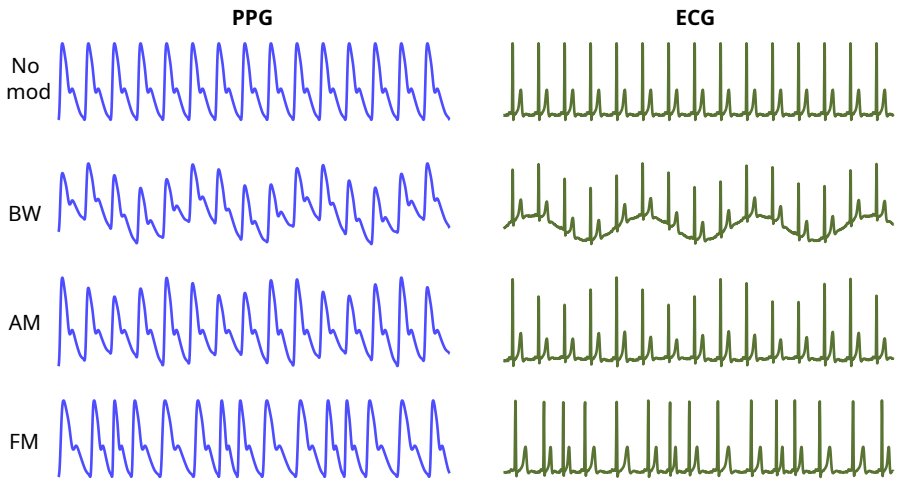

Figure 6: Idealised respiratory modulations of the PPG (left) and ECG (right). From top: no modulation, baseline wander (BW), amplitude modulation (AM), and frequency modulation (FM). Adapted from [63] and [64] by Peter H. Charlton - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=104501363

Baseline wander (BW), amplitude modulation (AM) and frequency modulation (FM) [65, 66].

This motivates to try to gain insight on the respiratory rate from PPG-signals. Recently, approaches to estimate RR have been incorporated into consumer wearables which measure the PPG, and therefore it is important to investigate the performance of this approach, and to optimise it.

7.2 Potential datasets

Besides the three datasets which are considered in Qumphy and are described below, i.e. MIMIC-III-Ext-PPG, MIMIC Perform Large and OSASUD, there are many other datasets containing PPG signals and respiratory signals or values for respiratory rate. Examples are the Vortal dataset [58], BIDMC [67] that is hosted on PhysioNet [68], and Capnobase.

7.3 Data selection

MIMIC-III-Ext-PPG, MIMIC Perform Large and OSASUD are our preferred datasets for RR estimation based on the same assessment dataset rules as for the other benchmark problems.

7.3.1 MIMIC-III-Ext-PPG

The MIMIC-III-Ext-PPG dataset was introduced in Sec. 3.2. For the subset of samples were a respiration (RESP) channel was present, we extracted also respiratory rates leveraging best practices from the literature [69]. This resulted in 4.3M 30s PPG waveform segments with respiratory rate annotation from more than 4600 patients.

7.3.2 MIMIC Perform Large

MIMIC Perform Large is a newly created dataset which is an extension of the MIMIC Perform datasets that are extracted from the MIMIC III Waveform Database. On zenodo, a training and a test set are provided:

- MIMIC PERform Large Training Dataset: 5,248 Recordings from 681 patients during routine clinical care, all of whom are adults.
- MIMIC PERform Large Testing Dataset: 1,257 Recordings from 167 patients during routine clinical care, all of whom are adults.

In both datasets, recordings are 32 seconds long, with reference respiratory rates (denoted rr) derived from the imp signals using the algorithm proposed in [69].

7.3.3 OSASUD

As pointed out before in the Sleep apnea benchmark problem, the OSASUD consists of 30 subjects, so each fold consists of nocturnal raw PPG signals and estimated features from 3 subjects. Segment duration - 1 min (raw PPG segment - 4800 samples, PPG feature sequence - 60 samples)

7.4 Making the datasets available

MIMIC-III-Ext-PPG will be released as a dataset on Physionet along with a corresponding dataset descriptor to be submitted to Scientific data.
MIMIC Perform Large can be downloaded via zenodo where a training and a test set are provided. Code to create 10 stratified folds with respect to respiration rate and the 10 folds can be downloaded from our repository very soon.
The original OSASUD dataset can be downloaded from here. Code to create 10 stratified folds with respect to respiration rate and the 10 folds can be downloaded from our repository very soon.

7.5 Data usage

All three datasets, MIMIC-III-Ext-PPG, MIMIC PERform Large and OSASUD are provided together with splits into multiple folds. Thus there are many different possibilities for the use of the datasets: For every dataset, training, validation, calibration and test sets can be taken from the provided folds such that models could be calculated from MIMIC-III-Ext-PPG or MIMIC PERform Large or OSASUD respectively. The other two datasets could be used as external test sets.
Alternatively, the original training and test set for MIMIC PERform Large could be used.

8 Conclusion

Photoplethysmographic signals are inexpensive and easy to collect and contain valuable information on the cardiovascular, respiratory, and autonomic nervous systems which is not yet routinely exploited. The report helps to enable the scientific and medical engineering community to create methods such that this information can be deduced from the signals. This is done by providing six exemplary benchmark problems of high clinical interest together with suitable benchmark datasets that are freely available or on demand, and explanations how to use these datasets to solve the Benchmark problems. We hope to support therefore the scientific progress in the field of PPG analysis by our report.

References

[1] P. H. Charlton, P. Kyriacou, J. Mant, and J. Alastruey.
"Acquiring Wearable Photoplethysmography Data in Daily Life: The PPG Diary Pilot Study".
In: *Engineering Proceedings* 2.1 (2020), p. 80. DOI: https://doi.org/10.3390/ecsa-7-08233.

[2] J. W. McEnvoy et al.
"2024 ESC Guidelines for the management of elevated blood pressure and hypertension".
In: *European Heart Journal* 45 (2024), pp. 3912–4018.

[3] *Medical Information Mart for Intensive Care*. 2024.
URL: https://mimic.mit.edu/ (visited on 04/24/2025).

[4] R. J. Mieloszyk et al. "A Comparison of Wearable Tonometry, Photoplethysmography, and Electrocardiography for Cuffless Measurement of Blood Pressure in an Ambulatory Setting".
In: *IEEE Journal of Biomedical and Health Informatics* 26 (2022), pp. 2864–2875.

[5] *UK Biobank*. 2025. URL: https://www.ukbiobank.ac.uk/ (visited on 04/24/2025).

[6] W. Wang, P. Mohseni, K. L. Kilgore, and L. Najafizadeh. "PulseDB: A large, cleaned dataset based on MIMIC-III and VitalDB for benchmarking cuff-less blood pressure estimation methods".
In: *Frontiers in Digital Health* 4 (2023).

[7] P. J. Aston. "Does Skin Tone Affect Machine Learning Classification Accuracy Applied to Photoplethysmography Signals?" In: *Computing in Cardiology* 51 (2024), p. 038.

[8] A. Cisnal, Y. Li, B. Fuchs, M. Ejtehadi, R. Riener, and D. Paez-Granados. "Robust Feature Selection for BP Estimation in Multiple Populations: Towards Cuffless Ambulatory BP Monitoring".
In: *IEEE Journal of Biomedical and Health Informatics* 28.10 (10/2024), pp. 5768–5779.
ISSN: 2168-2208. DOI: 10.1109/JBHI.2024.3411693.
URL: https://ieeexplore.ieee.org/document/10552318 (visited on 05/12/2025).

[9] Z.-D. Liu, Y. Li, Y.-T. Zhang, J. Zeng, Z.-X. Chen, J.-K. Liu, and F. Miao.
"HGCTNet: Handcrafted Feature-Guided CNN and Transformer Network for Wearable Cuffless Blood Pressure Measurement". eng.
In: *IEEE Journal of Biomedical and Health Informatics* 28.7 (07/2024), pp. 3882–3894.
ISSN: 2168-2208. DOI: 10.1109/JBHI.2024.3395445.

[10] J. Kim, S.-A. Chang, and S. W. Park. "First-in-Human Study for Evaluating the Accuracy of Smart Ring Based Cuffless Blood Pressure Measurement".
In: *Journal of Korean Medical Science* 39.2 (12/2023), e18. ISSN: 1011-8934.
DOI: 10.3346/jkms.2024.39.e18.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10789523/ (visited on 05/12/2025).

[11] M. Moulaeifard, L. Coquelin, M. Rinkevičius, A. Sološenko, O. Pfeffer, C. Bench, N. Hegemann, S. Vardanega, M. Nandi, J. Alastruey, C. Heissa, V. Marozas, A. Thompson, P. J. Aston, P. H. Charlton, and N. Strodthoff. "Machine-learning for Photoplethysmography Analysis: Benchmarking Feature, Image, and Signal-Based Approaches". In: *ArXiv* arXiv:2502.19949 (2025).

[12] In: (). URL: https://github.com/microsoft/aurorabp-sample-data/tree/main/docs#synthetic-measurements.

[13] B. P. Krijthe, A. Kunst, E. J. Benjamin, G. Y. Lip, O. H. Franco, A. Hofman, J. C. Witteman, B. H. Stricker, and J. Heeringa. "Projections on the number of individuals with atrial fibrillation in the European Union, from 2000 to 2060". In: *European heart journal* 34.35 (2013), pp. 2746–2751.

[14] S. S. Martin, A. W. Aday, Z. I. Almarzooq, C. A. Anderson, P. Arora, C. L. Avery, C. M. Baker-Smith, B. Barone Gibbs, A. Z. Beaton, A. K. Boehme, et al. "2024 heart disease and stroke statistics: a report of US and global data from the American Heart Association". In: *Circulation* 149.8 (2024), e347–e913.

[15] I. C. Van Gelder, M. Rienstra, K. V. Bunting, R. Casado-Arroyo, V. Caso, H. J. Crijns, T. J. De Potter, J. Dwight, L. Guasti, T. Hanke, et al. "2024 ESC Guidelines for the management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS) Developed by the task force for the management of atrial fibrillation of the European Society of Cardiology (ESC), with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC. Endorsed by the European Stroke Organisation (ESO)". In: *European Heart Journal* (2024), ehae176.

[16] N. R. Jones, C. J. Taylor, F. R. Hobbs, L. Bowman, and B. Casadei. "Screening for atrial fibrillation: a call for evidence". In: *European heart journal* 41.10 (2020), pp. 1075–1085.

[17] J. Lee, B. A. Reyes, D. D. McManus, O. Maitas, and K. H. Chon. "Atrial fibrillation detection using an iPhone 4S". In: *IEEE Transactions on Biomedical Engineering* 60.1 (2012), pp. 203–206.

[18] J.-P. Couderc, S. Kyal, L. K. Mestha, B. Xu, D. R. Peterson, X. Xia, and B. Hall. "Detection of atrial fibrillation using contactless facial video monitoring". In: *Heart rhythm* 12.1 (2015), pp. 195–201.

[19] T. Conroy, J. H. Guzman, B. Hall, G. Tsouri, and J.-P. Couderc. "Detection of atrial fibrillation using an earlobe photoplethysmographic sensor". In: *Physiological measurement* 38.10 (2017), p. 1906.

[20] A. Sološenko, A. Petrėnas, B. Paliakaitė, L. Sörnmo, and V. Marozas. "Detection of atrial fibrillation using a wrist-worn device". In: *Physiological measurement* 40.2 (2019), p. 025003.

[21] B. Paliakaitė, A. Petrėnas, A. Sološenko, and V. Marozas. "Modeling of artifacts in the wrist photoplethysmogram: Application to the detection of life-threatening arrhythmias". In: *Biomedical Signal Processing and Control* 66 (2021), p. 102421.

[22] A. Petrėnas, V. Marozas, and L. Sörnmo. "Low-complexity detection of atrial fibrillation in continuous long-term monitoring". In: *Computers in biology and medicine* 65 (2015), pp. 184–191.

[23] J. Bacevičius, V. Pluščiauskaitė, Ž. Abramikas, I. Badaras, M. Butkuvienė, S. Daukantas, E. Dvinelis, M. Gudauskas, E. Jukna, M. Kiseliūtė, R. Kundelis, J. Marinskienė, B. Paliakaitė, A. Petrėnas, M. Petrylaitė, A. Pilkienė, G. Pudinskaitė, V. Radavičius, A. Rapalis, ..., and V. Marozas. "Long-term electrocardiogram and wrist-based photoplethysmogram recordings with annotated atrial fibrillation episodes [Data set]". In: *Zenodo* (2024). DOI: doi.org/10.5281/zenodo.11242869. URL: https://zenodo.org/records/11242869.

[24] S. Bashar, D. Han, S. Hajeb-Mohammadalipour, E. Ding, C. Whitcomb, D. D. McManus, and K. H. Chon. "Atrial Fibrillation Detection from Wrist Photoplethysmography Signals Using Smartwatches." In: *Scientific Report* 9 (2019), p. 15054. DOI: 10.1038/s41598-019-49092-2.

[25] D. Han, S. K. Bashar, F. Mohagheghian, E. Ding, C. Whitcomb, D. D. McManus, and K. H. Chon. "Premature Atrial and Ventricular Contraction Detection Using Photoplethysmographic Data from a Smartwatch". In: *Sensors* 20.19 (2020), p. 5683. DOI: 10.3390/s20195683.

[26] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3.1 (2016), pp. 1–9.

[27] S. K. Bashar, E. Ding, A. J. Walkey, D. D. McManus, and K. H. Chon.
"Noise Detection in Electrocardiogram Signals for Intensive Care Unit Patients".
In: *IEEE Access* 7 (2019), pp. 88357–88368. DOI: 10.1109/ACCESS.2019.2926199.

[28] Z. Liu, B. Zhou, Z. Jiang, X. Chen, Y. Li, M. Tang, and F. Miao. "Multiclass Arrhythmia Detection and Classification From Photoplethysmography Signals Using a Deep Convolutional Neural Network".
In: *Journal of the American Heart Association* 11.7 (2022). DOI: 10.1161/JAHA.121.023555.

[29] Z. Liu. *PPGArrhythmiaDetection*. https://github.com/zdzdliu/PPGArrhythmiaDetection.
Multiclass Arrhythmia Detection and Classification from Photoplethysmography Signals Using a Deep Convolutional Neural Network. 2022.

[30] R. E. Climie, J. Alastruey, C. C. Mayer, A. Schwarz, A. Laucyte-Cibulskiene, J. Voicehovska, E. Bianchini, R.-M. Bruno, P. H. Charlton, A. Grillo, et al.
"Vascular ageing: moving from bench towards bedside".
In: *European journal of preventive cardiology* 30.11 (2023), pp. 1101–1117.

[31] A. Reshetnik, C. Gohlisch, M. Tölle, W. Zidek, and M. Van Der Giet.
"Oscillometric assessment of arterial stiffness in everyday clinical practice".
In: *Hypertension Research* 40.2 (2017), pp. 140–145.

[32] M. R. Hamczyk, R. M. Nevado, A. Barettino, V. Fuster, and V. Andrés.
"Biological versus chronological aging: JACC focus seminar".
In: *Journal of the American College of Cardiology* 75.8 (2020), pp. 919–930.

[33] P. H. Charlton, P. A. Kyriacou, J. Mant, V. Marozas, P. Chowienczyk, and J. Alastruey.
"Wearable Photoplethysmography for Cardiovascular Monitoring".
In: *Proceedings of the IEEE* 110.3 (2022), pp. 355–381. DOI: 10.1109/JPROC.2022.3149785.

[34] P. Salvi, A. Grillo, and G. Parati. "Noninvasive estimation of central blood pressure and analysis of pulse waves by applanation tonometry". In: *Hypertension Research* 38.10 (2015), pp. 646–648.

[35] P. H. Charlton, J. Mariscal Harana, S. Vennin, Y. Li, P. Chowienczyk, and J. Alastruey.
"Modeling arterial pulse waves in healthy aging: a database for in silico evaluation of hemodynamics and pulse wave indexes".
In: *American Journal of Physiology-Heart and Circulatory Physiology* 317.5 (2019), H1062–H1085.

[36] J. W. McEvoy, C. P. McCarthy, R. M. Bruno, S. Brouwers, M. D. Canavan, C. Ceconi, R. M. Christodorescu, S. S. Daskalopoulou, C. J. Ferro, E. Gerdts, H. Hanssen, J. Harris, L. Lauder, R. J. McManus, G. J. Molloy, K. Rahimi, V. Regitz-Zagrosek, G. P. Rossi, E. C. Sandset, B. Scheenaerts, J. A. Staessen, I. Uchmanowicz, M. Volterrani, R. M. Touyz, and E. S. D. Group.
"2024 ESC Guidelines for the management of elevated blood pressure and hypertension: Developed by the task force on the management of elevated blood pressure and hypertension of the European Society of Cardiology (ESC) and endorsed by the European Society of Endocrinology (ESE) and the European Stroke Organisation (ESO)". In: *European Heart Journal* 45.38 (08/2024), pp. 3912–4018.
ISSN: 0195-668X. DOI: 10.1093/eurheartj/ehae178.
eprint: https://academic.oup.com/eurheartj/article-pdf/45/38/3912/59633218/ehae178.pdf.
URL: https://doi.org/10.1093/eurheartj/ehae178.

[37] P. H. Charlton, J. M. Harana, S. Vennin, Y. Li, P. Chowienczyk, and J. Alastruey.
Pulse Wave Database (PWDB): A database of arterial pulse waves representative of healthy adults.
Version 0.2.0 Revised Submission to AJP Heart Circ. Zenodo, 04/2019.
DOI: 10.5281/zenodo.3275625. URL: https://doi.org/10.5281/zenodo.3275625.

[38] A. Abbasi, S. S. Gupta, N. Sabharwal, V. Meghrajani, S. Sharma, S. Kamholz, and Y. Kupfer.
"A comprehensive review of obstructive sleep apnea". In: *Sleep Science* (2021).
DOI: 10.5935/1984-0063.20200056.

[39] V. Kapur, D. Auckley, S. Chowdhuri, D. C. Kuhlmann, R. Mehra, K. Ramar, and C. Harrod.
"Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea: An American Academy of Sleep Medicine Clinical Practice Guideline." In: *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine* (2017).
DOI: 10.5664/jcsm.6506.

[40] Y. Yeghiazarians, H. Jneid, J. R. Tietjens, S. Redline, D. L. Brown, N. El-Sherif, R. Mehra, B. Bozkurt, C. E. Ndumele, and V. K. Somers. "Obstructive Sleep Apnea and Cardiovascular Disease: A Scientific Statement From the American Heart Association." In: *Circulation* (2021). DOI: 10.1161/cir.0000000000000988.

[41] O. M. Bubu, A. G. Andrade, O. Q. Umasabor-Bubu, M. Hogan, A. D. Turner, M. J. d. Leon, G. Ogedegbe, I. Ayappa, G. J.-L. G, M. L. Jackson, A. W. Varga, and R. S. Osorio. "Obstructive sleep apnea, cognition and Alzheimer's disease: A systematic review integrating three decades of multidisciplinary research." In: *Sleep Medicine Reviews* (2020). DOI: 10.1016/j.smrv.2019.101250.

[42] M. H. Smolensky, L. D. Milia, M. M. Ohayon, and P. Philip. "Sleep disorders, medical conditions, and road accident risk". In: *Accident Analysis & Prevention* (2011). DOI: 10.1016/j.aap.2009.12.004.

[43] P. B. z. Nieden. "The economic burden of (obstructive) sleep apnea. Costs and implications for Germany based on the results of an international systematic review". In: *Journal of public health* (2024). DOI: 10.1007/s10389-024-02269-0.

[44] L. C. Markun and A. Sampat. "Clinician-Focused Overview and Developments in Polysomnography." In: *Current sleep medicine reports* (2020). DOI: 10.1007/s40675-020-00197-5.

[45] M. Hirshkowitz. "Polysomnography Challenges." In: *Sleep medicine clinics* (2016). DOI: 10.1016/j.jsmc.2016.07.002.

[46] I. Rosen, D. Kirsch, R. Chervin, K. Carden, K. Ramar, R. Aurora, D. Kristo, R. Malhotra, J. L. Martin, E. Olson, C. Rosen, and J. Rowley. "Clinical Use of a Home Sleep Apnea Test: An American Academy of Sleep Medicine Position Statement." In: *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine* (2017). DOI: 10.5664/jcsm.6774.

[47] J. Costa, A. Rebelo-Marques, J. Machado, B. Valentim, C. S. d. A. V. Ferreira, J. D. O. Gonçalves, J. M. R. Gama, M. d. F. L. Teixeira, and J. Moita. "STOP-Bang and NoSAS questionnaires as a screening tool for OSA: which one is the best choice?" In: *Revista da Associação Médica Brasileira* (2020). DOI: 10.1590/1806-9282.66.9.1203.

[48] K. A. Evans, T. Yap, and B. Turner. "Screening Commercial Vehicle Drivers for Obstructive Sleep Apnea: Tools, Barriers, and Recommendations". In: *Workplace health & safety* (2017). DOI: 10.1177/2165079917692597.

[49] A. Bernardini, A. Brunello, G. L. Gigli, A. Montanari, and N. Saccomanno. "OSASUD: A dataset of stroke unit recordings for the detection of Obstructive Sleep Apnea Syndrome". In: *Scientific Data* 9.1 (2022), p. 177.

[50] A. Bernardini, A. Brunello, G. L. Gigli, A. Montanari, and N. Saccomanno. "AIOSA: An approach to the automatic identification of obstructive sleep apnea events based on deep learning". In: *Artificial Intelligence in Medicine* 118 (2021), p. 102133.

[51] M. A. Almarshad, S. Al-Ahmadi, M. S. Islam, A. S. BaHammam, and A. Soudani. "Adoption of transformer neural network to improve the diagnostic performance of oximetry for obstructive sleep apnea". In: *Sensors* 23.18 (2023), p. 7924.

[52] Y. Ji, D. Chen, Y. Zuo, T. Gao, and Y. Tang. "Accurate apnea and hypopnea localization in PSG with multi-scale object detection via dual-modal feature learning". In: *Biomedical Signal Processing and Control* 89 (2024), p. 105717.

[53] P. Kulkarni et al. "Obstructive sleep apnea syndrome identification using CNN-LSTM hybrid model". In: *Journal of Electrical Systems* 20.2 (2024), pp. 2386–2394.

[54] X. Chen, R. Wang, P. Zee, P. L. Lutsey, S. Javaheri, C. Alcántara, C. L. Jackson, M. A. Williams, and S. Redline. "Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA)". In: *Sleep* 38.6 (2015), pp. 877–888.

[55] T. Choksatchawathi et al. “ApSense: data-driven algorithm in PPG-based sleep apnea sensing”. In: *IEEE Internet of Things Journal* 11.20 (2024), pp. 33915–33926.

[56] A. S. Alarcón, N. M. Madrid, R. Seepold, and J. A. Ortega. “Obstructive sleep apnea event detection using explainable deep learning models for a portable monitor”. In: *Frontiers in Neuroscience* 17 (2023), p. 1155900.

[57] S. R. Braun. “Respiratory Rate and Pattern”. In: *Clinical Methods: The History, Physical, and Laboratory Examinations*. Ed. by H. Walker, W. Hall, and J. Hurst. Third edition. Boston: Butterworths, 1990. Chap. 43.

[58] P. H. Charlton, T. Bonnici, L. Tarassenko, D. A. Clifton, R. Beale, and P. J. Watkinson. “An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram”. In: *Physiological Measurement* 37.4 (03/2016), p. 610. DOI: 10.1088/0967-3334/37/4/610. URL: https://dx.doi.org/10.1088/0967-3334/37/4/610.

[59] R. Schein, N. Hazday, M. Pena, R. B.H., and C. Sprung. “Clinical antecedents to in-hospital cardiopulmonary arrest”. In: *Chest* 98.6 (1990), pp. 1388–92. DOI: 10.1378/chest.98.6.1388.

[60] D. Goldhill, S. White, and A. Sumner. “Physiological values and procedures in the 24 h before ICU admission from the ward.” In: *Anaesthesia* 54.6 (1999), pp. 529–34. DOI: 10.1046/j.1365-2044.1999.00837.x.

[61] S. Ridley. “The recognition and early management of critical illness.” In: *Ann R Coll Surg Engl.* 87.5 (2005), pp. 315–22. DOI: 10.1308/003588405X60669.

[62] M. Cretikos, R. Bellomo, K. Hillman, J. Chen, S. Finfer, and A. Flabouris. “Respiratory rate: the neglected vital sign.” In: *Med J Aust* 188.11 (2008), pp. 657–9. DOI: 10.5694/j.1326-5377.2008.tb01825.x.

[63] P. S. Addison, J. N. Watson, M. L. Mestek, and M. R. S. “Developing an algorithm for pulse oximetry derived respiratory rate (RR_{oxi}): a healthy volunteer study”. In: *J Clin Monit Comput* 26 (2012), pp. 45–51. DOI: 10.1007/s10877-011-9332-y.

[64] M. Pimentel, P. Charlton, and D. Clifton. “Probabilistic estimation of respiratory rate from wearable sensors.” In: *Wearable Electronics Sensors*. Ed. by S. Mukhopadhyay. Vol. 15. Cham: Springer, 2015, pp. 241–62.

[65] R. Bailon, L. Sornmo, and P. Laguna. “ECG-derived respiratory frequency estimation.” In: *Advanced Methods and Tools for ECG Data Analysis*. London: Artech Houes, 2006. Chap. 8, pp. 215–44.

[66] D. Meredith, D. Clifton, P. Charlton, J. Brooks, and L. Pugh C.W. Tarassenko. “Photoplethysmographic derivation of respiratory rate: a review of relevant physiology”. In: *J Med Eng Technol.* 36.1 (2012), pp. 1–7. DOI: 10.3109/03091902.2011.638965.

[67] M. A. F. Pimentel, A. E. W. Johnson, P. H. Charlton, D. Birrenkott, P. J. Watkinson, L. Tarassenko, and D. A. Clifton. “Toward a Robust Estimation of Respiratory Rate From Pulse Oximeters”. In: *IEEE Transactions on Biomedical Engineering* 64.8 (2017), pp. 1914–1923. DOI: 10.1109/TBME.2016.2613124.

[68] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. “PhysioBank, PhysioToolkit, and PhysioNet”. In: *Circulation* 101.23 (2000), e215–e220. DOI: 10.1161/01.CIR.101.23.e215. eprint: https://www.ahajournals.org/doi/pdf/10.1161/01.CIR.101.23.e215. URL: https://www.ahajournals.org/doi/abs/10.1161/01.CIR.101.23.e215.

[69] P. H. Charlton, T. Bonnici, L. Tarassenko, D. A. Clifton, R. Beale, P. J. Watkinson, and J. Alastruey. “An impedance pneumography signal quality index: Design, assessment and application to respiratory rate monitoring”. In: *Biomedical signal processing and control* 65 (2021), p. 102339.