On the prediction of clinical outcomes using Heart Rate Variability estimated from wearable devices



Davide Morelli

Linacre College

University of Oxford

Supervised by

Prof. David Clifton

A DPhil thesis submitted to the University of Oxford in accordance with the requirements of the degree of Doctor of Philosophy in the Department of Engineering Science.

Trinity Term 2023

Abstract

This thesis explores the use of Heart Rate Variability as a tool for predicting health outcomes, focusing on data derived from photoplethysmography (PPG) sensors in wrist-worn wearable devices such as smartwatches. These devices offer a unique opportunity for cost-effective, continuous, and unobtrusive monitoring of heart health. However, PPG data is susceptible to motion artefacts, challenging the reliability of Heart Rate Variability metrics derived from it.

A critical finding of this research is the unreliability of specific frequency-domain Heart Rate Variability features, such as the Sympathovagal Balance Index (SVI), due to low signal-to-noise ratio in certain frequency bands. Conversely, the thesis demonstrates that most HRV features, including Root Mean Square of Successive Differences between normal heartbeats (RMSSD) and Standard Deviation of Normal heartbeats (SDNN), can be reliably extracted under conditions of motion, such as during physical activity or recovery from exercise. This is achieved by employing accelerometry data from wearable devices to filter out unreliable PPG data.

The thesis also addresses the issue of missing data in Heart Rate Variability analysis, a consequence of motion artefacts and the energy-saving strategies of wearable devices. By exploring different interpolation methods and their effects on Heart Rate Variability features, this research identifies the best approaches for handling missing data. Particularly, it recommends operating on timestamp time-series over duration time-series, contradicting traditional Heart Rate Variability preprocessing practices. Quadratic interpolation in the time domain was identified as the most effective method, introducing minimal error across numerous Heart Rate Variability features, contrary to interpolation in the duration domain.

The research presented in this thesis evaluates Heart Rate Variability features derived from ultra-short measurement windows, demonstrating the feasibility of accurately estimating RMSSD and SDNN using 30-second and 1-minute time windows, respectively. This study, unique in assessing the effect of missing values on ultra-short Heart Rate Variability data, reveals that missing values significantly impact SDNN estimations while moderately affecting RMSSD. The analysis highlights that ultra-short inter-beat interval time series limit the assessment of very low frequency (VLF) components, increasing bias in SDNN estimates. This finding is particularly significant in light of the prevalent use of SDNN in commercial wearables, underscoring its importance for continuous heart health monitoring. The study notes that the shorter the measurement window and the greater the amount of

missing values, the larger the bias observed in SDNN.

A novel aspect of the thesis is the creation of an innovative mathematical model designed to estimate the impact of circadian rhythms on resting heart rate. This model stands out for its computational efficiency, making it particularly suitable for data obtained from wearable devices. It surpasses the single component cosinor model in accuracy, demonstrated by a lower root mean square error (RMSE) in predicting future heart rate values. Additionally, it retains the advantage of providing easily interpretable parameters, such as MESOR, Acrophase, and Amplitude, which are essential for assessing changes in heart activity.

The thesis demonstrates that Heart Rate data can accurately estimate SDNN24 (the Standard Deviation of NN intervals over 24 hours), with a difference of about 0.22 ± 11.47 (RMSE = 53.81 and $r^2 = 0.97$). This finding indicates that despite being fragmentary, 24-hour HR data from wrist-worn fitness devices is adequate for estimating SDNN24 and assessing health status, as evidenced by an F1 score of 0.97. The robustness of SDNN24 estimation against noisy data suggests that wrist-worn wearables are capable of reliably monitoring cardiovascular health on a continuous basis, thus facilitating early interventions in response to changes in Sinoatrial Node activity.

The final part of the thesis introduces an innovative approach to health outcome prediction, employing Heart Rate Variability data gathered during exercise alongside Electronic Health Record data. Employing Large Language Models to process EHR data and Convolutional AutoEncoders for Heart Rate Variability analysis, this approach reveals the untapped potential of exercise Heart Rate Variability data in health monitoring and prediction. Deep Learning models incorporating Heart Rate Variability data demonstrated enhanced predictive accuracy for cardiovascular diseases (CVD), coronary heart disease (CHD), and Angina, evidenced by higher Area Under the Curve (AUC) scores compared to models using only Electronic Health Records and demographic/behavioural data. The highest AUC scores achieved were 0.71 for CVD, 0.74 for CHD, and 0.73 for Angina.

In conclusion, this thesis contributes to the field of biomedical engineering by enhancing the understanding and application of HRV analysis in health outcome prediction using wearable device data. It offers insights for future work in continuous, unobtrusive health monitoring and underscores the need for further research in this rapidly evolving domain.

Acknowledgement

First and foremost, I would like to thank my supervisor, Prof. Clifton, for the guidance and support throughout this project. Prof. Clifton has been an invaluable resource, providing me with feedback, advice, and encouragement during the research and writing process. I am grateful for his expertise and dedication.

I would also like to thank my colleagues and friends at IBME, the University of Pisa, BioBeats and Huma for their insights and support. I have learned so much from them, and I am grateful for their willingness to share their knowledge and expertise with me.

Finally, I would like to thank my family for their love and support. They have always been my biggest champions, and I am so grateful for their sacrifices and encouragement.

I could not have completed this project without the help and support of all of these people. I am truly grateful for their contributions.

This Thesis is dedicated to my wife, Bianca, and my daughter, Anna. With all my love.

Relevant Publications

- **Morelli, D.**, Rossi, A., Bartoloni, L., Cairo, M., Clifton, D. SDNN24 estimation from semi-continuous HR measures. *Sensors*. MDPI (2021).

- **Morelli, D.**, Bartoloni, L., Rossi, A., Clifton, D. A computationally efficient algorithm to obtain an accurate and interpretable model of the effect of circadian rhythm on resting heart rate. *Physiological Measurement*. IOP Publishing (2019).

- **Morelli, D.**, Rossi, A., Cairo, M., Clifton, D. Analysis of the impact of interpolation methods of missing RR-intervals caused by motion artefacts on HRV features estimations. *Sensors*. MDPI (2019).

- **Morelli, D.**, Bartoloni, L., Colombo, M., Plans, D., Clifton, D. Profiling the propagation of error from PPG to HRV features in a wearable physiological-monitoring device. *Healthcare technology letters*. The Institution of Engineering and Technology (2018).

- Lu, L., Zhu, T., **Morelli, D.**, Creagh, A., Liu, Z., Yang, J., Liu, F., Zhang, Y., Clifton, D., Uncertainties in the Analysis of Heart Rate Variability: A Systematic Review, *IEEE Reviews in Biomedical Engineering* (2023).

- Rossi, A., Pedreschi, D., Clifton, D., **Morelli, D.**. Error estimation of ultra-short heart rate variability parameters: effect of missing data caused by motion artefacts, *Sensors*. MDPI (2020).

- Da Pozzo, E., Piccarducci, R., De Felice, M., Daniele, S., Tremolanti, C., Germelli, L., Rossi, A., Menicagli, D., Clifton, D., Sirbu, A., Priami, C., **Morelli, D.**, Martini, C. Keep a watchful eye on circadian rhythm and brain wellness, In Proceedings of the 45th FEBS Congress (2021).

- Rossi, A., Calogiuri, G., Maffi, S., Pedreschi, D., Clifton, D., **Morelli, D.** Physiological recovery among workers in long-distance sleddog races: A case study on female veterinarians in Finnmarksløpet, *Work* (2022).

- Bacciu, D., Bertoncini, G., **Morelli, D.** Topographic mapping for quality inspection and intelligent filtering of smart-bracelet data. *Neural Computing and Applications*. Springer (2022).

- Rossi, A., Da Pozzo, E., Menicagli, D., Tremolanti, C., Priami, C., Sîrbu, A., Clifton, D., Martini, C., **Morelli, D.** A public dataset of 24-H multi-levels psycho-physiological responses in young healthy adults. *Data*. MDPI (2020).

- Rossi, A., Da Pozzo, E., Menicagli, D., Tremolanti, C., Priami, C., Sirbu, A., Clifton, D., Martini, C., **Morelli, D.** Multilevel Monitoring of Activity and Sleep in Healthy People. *PhysioNet* (2020).

During my research I have also published other works not directly related to the subject of this thesis, including digital biomarkers [55, 54, 129]; mood and stress estimation from HRV data [8, 30]; epidemiological models [151, 160, 44, 53]; efficacy of stress reduction interventions driven by HRV data collected from smartphones [85, 105, 180, 176]; and experimental psychology [177, 181].

Contents

1 Introduction **1**

1.1 Heart Rate Variability 1

1.2 Research Shortcomings addressed in this thesis 21

1.3 Outline of the thesis 22

2 Error propagation from PPG to HRV **25**

2.1 Introduction 25

2.2 Experimental Design 26

2.3 Methods 27

2.4 Results 31

2.5 Conclusions 35

3 Effects of interpolation methods on HRV **37**

3.1 Introduction 37

3.2 Materials and Methods 39

3.3 Results and Discussions 45

3.4 Conclusions 51

4 Effect of missing data on HRV **53**

4.1 Introduction 54

4.2 Materials and Methods 54

4.3 Results 56

4.4 Discussion 60

4.5 Conclusions 63

5 Circadian rhythm estimation **65**

5.1 Introduction 66

5.2 Methods 70

5.3 Conclusion . . . 79

6 SDNN24 estimation from HR **81**

6.1 Introduction . . . 82

6.2 Methods . . . 83

6.3 Results . . . 89

6.4 Discussion . . . 92

6.5 Conclusion . . . 94

7 Health outcomes prediction **95**

7.1 Introduction . . . 95

7.2 Definition of the Health Outcomes of interest . . . 96

7.3 Dataset . . . 97

7.4 Electronic Health Records representation model . . . 99

7.5 HRV representation model . . . 111

7.6 Health Outcomes prediction from HRV . . . 118

7.7 Conclusions . . . 126

8 Conclusions and future work **127**

8.1 Limitations . . . 127

8.2 Future work . . . 128

8.3 Contributions of this Thesis and Conclusions . . . 131

A Hyper-parameter search **159**

A.1 BASIC . . . 159

A.2 EHR . . . 161

A.3 HRV . . . 162

A.4 BASIC+HRV . . . 164

A.5 BASIC+EHR . . . 165

A.6 BASIC+HRV+EHR . . . 167

A.7 HRV+EHR . . . 168

B ROC curves for 5 and 10 years prediction windows **171**

C Results for prediction window at 2 years **173**

D Experiment HW and SW stack **175**

E Relative Error on HRV features induced by interpolation methods **177**

F Survival Analysis for Health Outcomes **179**

Chapter 1

Introduction

1.1 Heart Rate Variability

Heart Rate Variability (HRV) is the study of the dynamics of heart activity. HRV is a popular and widely accepted clinical tool for the assessment of the activity of the autonomic nervous system [16], a complex and dynamic regulatory network, that plays a pivotal role in maintaining homeostasis and influencing physiological responses, making it a central focus in the realm of biomedical engineering research.

HRV has proven to predict adverse outcomes in clinical settings [131, 243, 28, 16], and is affected by, among others, cardiovascular disease [228, 87, 142, 5, 118, 90, 143, 59, 49, 216, 214, 226]; diabetes [11, 111, 87, 183, 122, 29, 162, 94]; hypertension [91, 212]; stress [71, 126, 19, 20, 84, 104, 13, 50, 233, 219, 196, 36]; and obesity [245].

Cardiovascular disease (CVD) is the leading cause of mortality in Europe, with almost 4 million deaths every year (45% of all deaths), more than 11 million new cases in Europe every year, 85 million people living with the disease, and with an estimated cost to the European Union economies as high as €210 billion a year [241]. Diabetes is a metabolic disease that causes heart conditions and kidney disease. In Europe, the prevalence of diabetes is estimated to be 8.5%, affecting 56 million people [220]. Hypertension affects 20% to 50% of the adult population in developed countries [15], is a known risk factor for CVD, increases the risk of heart attack, stroke, and other health problems. [91, 156, 80]. Psychological stress is not only the cause of decrease quality of life, but is also linked to enhanced risk of several health conditions, including cardiovascular disease, immunodeficiency, depression, and cancer [41, 227].

CVD screening aims at uncovering unsuspected early cardiovascular disease in patients with no symptoms, which could enable early treatment resulting in a delayed or even prevent the development of the disease [42]. A passive and continuous screening of heart activity could enable early detection, but the general population tests for CVD specific biomarkers (biological marker, i.e. a measurable physiological phenomenon, such as HRV, that is useful for the prediction or assessment of a medical condition [218]) only in absence of specific symptoms, or for periodic

health checks. The frequency of health checks in the general population can vary widely and is influenced by factors such as individual health status, age, risk factors, and healthcare system recommendations. In many countries, routine health check-ups are recommended annually or biennially for adults. In a study conducted in 2011 around cardiovascular risk prevention [45], it was found that less than 45% of the high risk population invited by the NHS to periodic health checks actually attended. Another study found that the uptake of health checks offered by the NHS in the general healthy population, without known health conditions, was less than 50%. We can therefore expect a large number of cases are not diagnosed until symptoms are present.

Most wearable devices, such as fitness trackers and smart watches, incorporate sensors that can passively measure cardiac activity during daily activities and during the night. Their adoption is projected to cover 25% of the US population in the next few decades [62]. Such a widespread adoption of wearable devices holds great promise for the development of systems that proactively monitor cardiac health and trigger early warnings when abnormal activity is detected. However, compared to measures taken using clinical grade devices, wearable devices are affected by considerable noise, and as will be discussed in section 1.1.5, the error propagates to Heart Rate Variability in ways that make this data unreliable and problematic to use in prediction models. In Chapters 2, 3, and 4 I study how error propagates from noise in the duration of the individual heartbeats to HRV features, and the effect of interpolation techniques commonly used to compensate for missing heartbeats.

Most research has focused on HRV obtained from 5 minutes of recording or less (*short* and *ultra-short* HRV features). As I will explain in section 1.1.2, this mostly captures the activity of the autonomic nervous system. However, as explained in section 1.1.1, the activity of the Sinoatrial node, the pacemaker of our heart, is not visible at such high frequencies, but can be estimated from *ultra-low* frequencies, that require up to 24 hours. In Chapter 6 I demonstrate how it is possible to estimate Sinoatrial node activity from fitness trackers, and in Chapter 5 I accurately model the effect of circadian rhythm on heart dynamics.

Because HRV data is traditionally extracted from data collected during hospital visits, most research focuses on HRV at rest. As discussed in sections 1.1.3, there is an increasing amount of evidence that the heart dynamics during physical exercise has prognostic relevance, especially for CVD. In Chapter 7 I also discuss how HRV during physical activity, the kind of data that can routinely be collected by wearable devices, can predict a number of health conditions, and design an experiment to validate my hypothesis.

1.1.1 Physiology of Heart Rate Variability

Heart Rate Variability is a physiological phenomenon of fluctuation between adjacent inter-beat intervals (IBI) [140, 203]. The heart beats average (HR) and HRV are regulated by two main physiological mechanisms: the Autonomic Nervous System and the Sinoatrial Node.

Autonomic Nervous System

The Autonomic Nervous System (ANS) controls the activity of many organs and bodily functions, including heart rate, pupil dilation, digestion, breathing rate, and others. ANS is composed of two antagonistic sub-systems, the sympathetic nervous system (SNS) and the Parasympathetic Nervous Systems (PNS) [139].

Sympathetic Nervous System (SNS) is responsible for the *fight or flight* response, which is activated in the presence of danger or stressful situations. The primary role of SNS activation is to suspend non-essential bodily functions, particularly diverting energy away from digestion to the muscles, thus facilitating physical activities. This prepares the body for potential threats by providing additional energy through increased heart rate and breathing rate.

Parasympathetic Nervous System (PSN) is called the *rest and digest* system, and is predominant when the subject is at rest, sleeping, in a relaxed state, or digesting. The goal of PNS activation is to direct the resources of the body to building up strength by increasing salivation and digestion, cellular regeneration, lowering heart rate. PNS has nerves that start both at the sacral spinal chord and at the cranium. In particular the vagus nerve, an important mediator of PNS activity, starts from the cranium and arrives at the diaphragm. The breathing cycle stimulates the vagus nerve, leading to cyclic PNS inhibition during the inhaling phase and PNS elicitation in the exhaling phase [139].

ANS affects the activity of the heart and of many other organs in the body. Both SNS and PNS have nerves from the spinal column to stomach; duodenum; colon; ileum; spleen; gallbladder; liver; pancreatic head; appendix; kidneys; and ureters.

Because of how many organs are affected by ANS, being able to measure its activity, and to assess the balance between SNS and PNS, can provide useful insight on the overall health of the subject.

Since ANS also affects heart dynamics, its activity can be assessed by studying HRV. SNS makes the heart beat faster, PNS slower [139, 136, 189].

Sinoatrial Node

The second mechanism that regulates HR and HRV is the *coupled-clock* system which drives the automatism of the Sinoatrial Node (SAN), the pacemaker cells [229, 189].

Exploring the frequency analysis of the Sinoatrial Node gains particular significance as it addresses a notable gap in HRV studies, where the focus predominantly revolves around autonomic nervous system activity, underscoring the need to comprehensively assess Sinoatrial Node dynamics for a more nuanced understanding of cardiac regulation. Rosenberg et Al. [189] analysed the relative contribution of ANS and SAN to canine HRV: by performing ANS-blockade they could observe HRV spectrum when affected by activity of SAN only, without any influence from ANS, and compare it against HRV spectrum when affected by both SAN and ANS, allowing them to identify the

contributions of ANS and SAN. SAN gives the HRV spectrum the characteristic shape between pink (spectral density inversely proportional to f) and Brownian noise (spectral density inversely proportional to f^2, the square of the frequency). ANS contribution to the ultra low frequencies is extremely limited but dominates the very low, low, and high frequencies of HRV spectrum.

Most of the HRV literature focuses on analyses from measurements of 5 minutes or less, therefore restricted to low and high frequency bands, that are dominated by ANS. However, the activity of the pacemaker of the heart is obviously of great interest when evaluating cardiac health. Because SAN activity dominates over ANS activity only in ultra low frequencies, to appreciate the activity of SAN it is necessary to measure cardiaac activity for long periods. Measuring the duration of individual heartbeats, called Inter-Beat Invervals (IBI) for 24 hours allows to capture the effect of circadian modulation on the heart. Inter-Beat Intervals are also called NN intervals, representing the temporal duration between the peak of one QRS complex (explained in section 1.1.2) to the peak of the next QRS complex on an electrocardiogram. One of the most informative tests for cardiac health involves using a Holter device (an ECG chest strap) for 24 hours. From this data the Standard Deviation of IBI over 24 hours (SDNN24) is calculated, and regarded as a predictor of cardiac health [203, 108, 72, 101, 86]. In Chapter 6 I will study how SDNN24 can be estimated from data that can be easily collected from commercial grade fitness trackers, potentially enabling SAN activity monitoring via passively collected wearable data.

1.1.2 Measures of Heart Rate Variability

History of HRV analysis

Fragments of writings on pulse rate date back to ancient Greek and Roman physicians [16]. Herophilus first wrote about heartbeat in the third century AD. Archigens in the first century AD observed that pulse rate is not regular. Rufus of Ephesus in the second century noted that pulse rate is the effect of contractions of the heart.

Galenus in the second century AD extensively wrote about pulse rate and how it can be used in medical practice to diagnose diseases. His work influenced practitioners for centuries.

Floyer in the beginning of the eighteenth century invented *the physician pulse watch*, a device that could measure time, and allowed physicians to accurately measure pulse rate and breathing rate [65]. In 1733 this innovation allowed Hales to measure and describe changes in duration of heartbeats throughout the respiratory cycle [99]. Ludwig in 1847 first measured Respiratory Sinus Arrhythmia [133], a physiological phenomenon by which breathing induces noticeable changes in the instantaneous frequency of heartbeats [16].

In the late nineteenth century electrocardiography (ECG) was invented. In 1870 Muirhead first reported to have observed the electrical activity of the heart by attaching wires to the patient [99], in 1887 Waller created a machine that could record ECG in realtime [238]. In 1885 Einthoven named the characteristic ECG waveform features with the P, Q, R, S, and T letters that are still in use today (see figure 1.1), and developed mathematical equations that allowed to remove instrumental error from ECG devices [92]. His work was rewarded with the Nobel Prize in

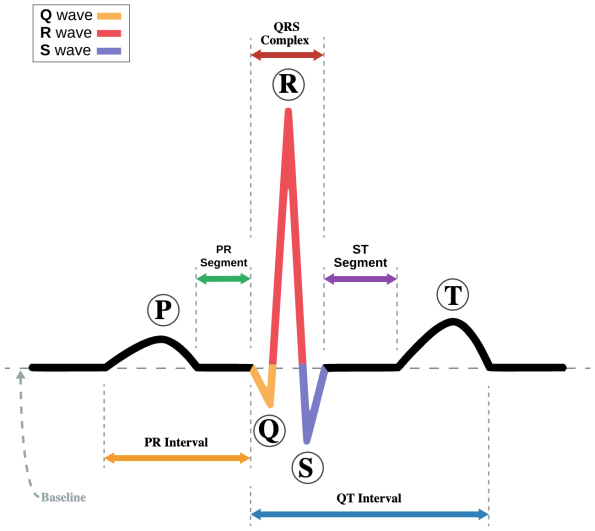


Figure 1.1: Schematic diagram of the QRS complex, preceded by the P wave, and followed by the T wave, in normal sinus rhythm for a human heart as seen on ECG. Credit: Anthony Atkielski, wikipedia.org.

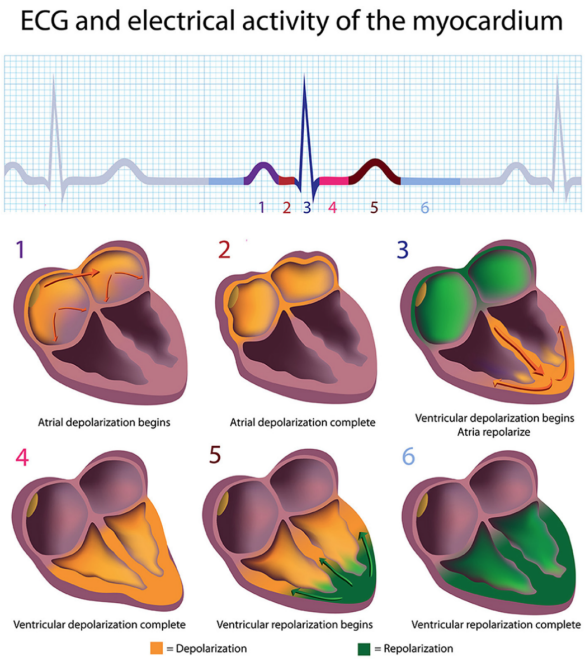


Figure 1.2: The heart depolarisation cycle. Credit: Alila Sao Mai, Shutterstock.com.

Medicine in 1924.

ECG devices allowed precise inter-beat-intervals measurements. This enabled researchers to analyse HRV. As a consequence, in the 1960s, time domain HRV features started to be defined to quantify cardiovascular activity [16]. In the 1970s power spectrum analysis started to be investigated and this allowed frequency domain HRV features to be defined [16]. In the 1990s non-linear techniques, such as Sample Entropy or Detrended Fluctuation Analysis, were used to quantify the chaotic behaviour of HRV [16].

Electrocardiography and Photoplethysmography

In this section I will discuss the differences between two of the most widely used techniquest to measure heart activity: Electrocardiography and Photoplethysmography.

Electrocardiography (ECG) ECG is a process of measurement of the heart electrical activity by attaching electrodes to the skin of the patients in specific regions. The electrodes measure the electrical discharge that follows the depolarisation of the heart muscles during each cardiac cycle.

As shown in figure 1.2, at each heartbeat a specific sequence of muscle contraction happens, starting from the SAN cells, propagating to the atrium, then the ventricles. This pattern creates an electrical activity that, when measured by an ECG device, results in the characteristic ECG tracing.

As discussed in section 1.1.2 Einthoven observed and named the ECG waveform features that make the QRS complex, preceded and followed by the P and T waves, respectively, as shown in figure 1.1.

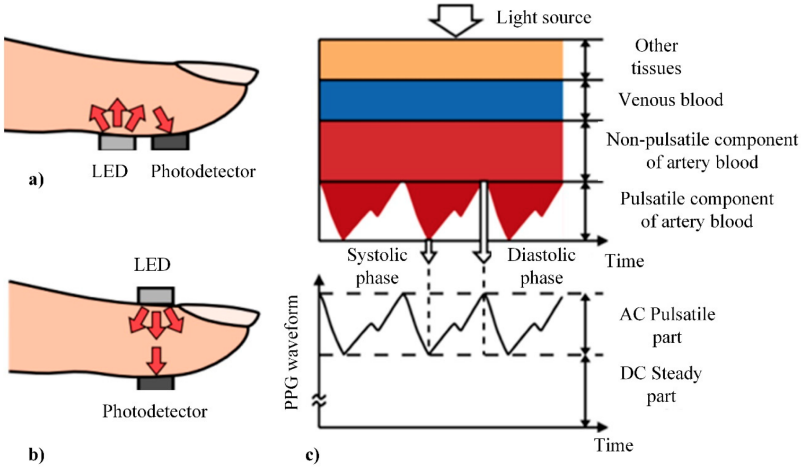


Figure 1.3: Mechanism of action of photoplethysmography. Figure (a) shows a reflective pulse oximeter. Figure (b) shows a transmitting pulse oximeter. Figure (c) shows how is light absorbed and how it results in the characteristic PPG trace. Credit: Dzedzickis et Al. [56].

The P wave signifies atrial depolarisation, reflecting the contraction of the atria as an electrical impulse spreads through them. Following the P wave, the QRS complex represents ventricular depolarisation, signifying the contraction of the ventricles. The R wave is the first positive deflection in the QRS complex, indicating the peak of ventricular depolarisation. The S wave, a negative deflection, follows the R wave, marking the completion of ventricular depolarisation. Finally, the T wave represents ventricular repolarisation, illustrating the recovery phase when the heart chambers prepare for the next cycle. These ECG labels are crucial in diagnosing cardiac abnormalities, offering a comprehensive insight into the coordinated electrical events that drive the cardiac cycle. Deviations from the expected PQRST complex are used in the diagnosis of, among others, cardiac arrhythmias [47]; ventricular hypertrophy [148]; and myocardial infarction [237].

Photoplethysmography (PPG) With every cardiac cycle, the heart pumps blood in the arteries to every part of the body, creating a wave of blood that goes through the vascular system. The blood absorbs light, it is therefore possible to measure the difference over time of the quantity of blood in a specific point of the body by measuring the amount of light reflected by (or going through) the skin.

Figure 1.3 shows the mechanism of PPG, with photons emitted by a LED and detected by a sensor placed nearby. Veinous blood and tissues in the skin constantly absorb a portion of the emitted light, that will result in a DC component of the obtained PPG trace. The variable component of the PPG trace is caused by the pulsatile part of the artery blood absorbing part of the light.

Figure 1.4 shows the fiducial points of the PPG waveform. The distance in time between the fiducial points yield important information about the cardiovascular health [141, 146]. The systolic peak is effect of the systolic ejection; the diacrotic notch and diastolic peak are the effect of wave reflection from the vascular system [7]. For example, the timing between systolic peak, diacrotic notch, and diastolic peak, are influenced by arterial stiffness

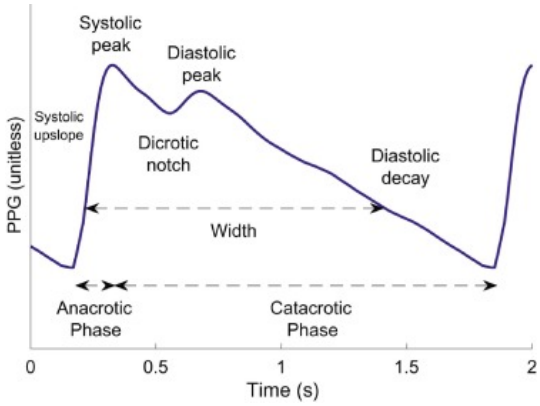


Figure 1.4: Fiducial points of the PPG waveform. Credit: Mejía-Mejía et Al. [141].

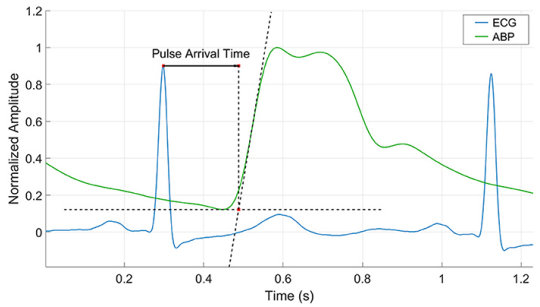


Figure 1.5: Comparison of an ECG trace and its corresponding Ambulatory Blood Pressure (ABP) trace. Credit: [60].

[146].

Figure 1.5 shows an ECG trace and an Ambulatory Blood Pressure (ABP) trace measured simultaneously. ABP measures the pulsatile component of artery blood, similarly to PPG. The heart contraction is visible in both ECG and ABP (and PPG), with a delay in the ABP (PPG). This delay is called Pulse Arrival Time (or also Pulse Transmit Time) and is caused by the difference in time required by electric signals to reach the ECG sensors, compared to the time necessary to the blood to travel from the heart to the point in the body where the ABP (PPG) sensor is placed.

Most wearable devices today have a sensor dedicated to tracking heart activity. Some wearable devices have an ECG sensor, but it usually requires active explicit actions from the user to activate the measurement, which greatly reduces the possibility to track users for extended time, during the night, with frequent measurements that allow to spot deterioration of health status early on. Most wearable devices have optical PPG sensors that allow heart activity to be collected passively, without requiring active interaction with the user, which enables tracking for extended periods, during day and night, with frequent measurements, which allows the early detection of health deterioration.

HRV analysis

HRV analysis attempts to quantify the complex dynamics of heart activity with numerical values, to find correlations with health outcomes, to be able to use HRV as a descriptor and predictor of health and wellbeing.

HRV analysis is usually done by applying a series of formulae to the sequence of the inter-beat-intervals, called HRV features, that are commonly separated in three groups: time domain, frequency domain, and non-linear features.

Time domain analysis usually contains various statistical variables of the duration time series. The frequency domain analysis investigates the power spectrum of RR-intervals time series in order to assess the cardiac autonomic balance (i.e., sympathetic and parasympathetic nervous systems activity). Non-linear HRV features try to capture the non-periodic behaviour of the HRV and the complexity that exists inside the inter-beat-interval dynamics.

Time Domain features

Heart Rate Heart Rate (HR) is a measure of the frequency of heartbeats, typically expressed in beats per minute (bpm). HR is determined by counting the number of heartbeats that occur within a predetermined time interval. HR serves as a vital indicator of cardiac function and overall heart health. It is subject to a variety of influences, including sex, age, stress levels, fitness level, autonomic nervous system activity, dietary habits, and various medical conditions.

HR holds a ubiquitous presence in the realm of heart activity assessment, being well-recognised by the general populace and incorporated into a wide array of wearable devices. To conserve battery life, measurements are often taken over intervals shorter than a minute. Taking this principle to the extreme in some cases is defined the notion of instantaneous HR, calculated as $60/\mathrm{NN}_i$, where NN_i represents the duration of the i^{th} heartbeat, measured as the distance between two consecutive normal peaks in the QRS complex in an ECG trace. It's essential to note that taking the average of instantaneous HR values is not equivalent to computing the HR as a mean of instantaneous heart rates ($\mathrm{HR} \neq \frac{1}{n}\sum_i^n \frac{60}{\mathrm{NN}_i}$). This discrepancy arises from the fact that the average of reciprocal values in a dataset is generally not equal to the reciprocal of the mean of the original values.

To illustrate this, consider a short NN sequence comprising intervals of 0.5, 0.5, 1.5, and 1.5 seconds. These NN values correspond to instantaneous heart rates of 120 bpm, 120 bpm, 40 bpm, and 40 bpm, respectively. We count 4 heartbeats in 4 seconds, which corresponds to an HR of 60 bpm. However, when averaging the instantaneous HR directly (120, 120, 40, and 40), we obtain 80 bpm.

AverageNN The average inter-beat-interval duration [215, 200, 203]. AverageNN is shown in Equation (1.1).

$$\mathrm{AverageNN} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{NN}_i \tag{1.1}$$

AverageNN is the reciprocal of HR multiplied by 60: AverageNN = $\frac{60}{HR}$

RMSSD The root mean square of the successive RR-intervals (distance between two consecutive R peaks in an ECG trace) differences (Equation (1.2)) represents the strength of the autonomic nervous system (specifically the parasympathetic branch) at a given time [97, 215, 200, 203].

$$\text{RMSSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n-1}(\text{NN}_{i+1} - \text{NN}_i)^2} \tag{1.2}$$

Because the first step in RMSSD calculation is to compute successive differences, any trend and low frequencies present in the data are removed. RMSSD therefore reflects the NN variance that is sensitive to high-frequency (HF) heart period fluctuations such as in the respiratory frequency range and is used as an index of vagal cardiac control [204]. It is considered the most useful marker for the parasympathetic nervous system activation [136, 252]. RMSSD is widely used by professional athletes in order to monitor cardiac activity and consequently their health status. In particular, it has been found that acute decreases in RMSSD have been linked to intense endurance training [178], resistance training [31], and competition [58]. Therefore, low RMSSD reflects acute fatigue from high intensity physical effort. As discussed in section 1.1.3 low RMSSD is also associated with higher risk of cardiovascular disease. In the last decades, strong association between reduced RMSSD and depression has been reported[106, 1].

SDNN The Standard Deviation of NN intervals (Equation (1.3)). It reflects the cyclic components responsible for variability in the RR-intervals time series. The SDNN is the "gold standard" for medical stratification of both morbidity and mortality [203, 14, 97, 215, 200, 136]. SDNN is the standard deviation of the inter-beat-interval sequence, and is therefore affected both by slow and fast changes in HRV. However, slow changes, in particular a trend in the time series, will affect SDNN the most. I will discuss in Chapter 6 how SDNN can be used to estimate SAN activity, and how this metric predicts cardiovascular health.

$$\text{SDNN} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{NN}_i - \overline{\text{NN}})^2} \tag{1.3}$$

where $\overline{\text{NN}}$ is the mean NN duration in the time series.

SDNN, and RMSSD, i.e. the Root Mean Squared of Successive Differences (inter-beat intervals), are two of the most widely used HRV parameters to predict many health issues [203]. It was been found that SDNN24 (SDNN of 24 hours of inter-beat intervals) values lower than 50 milliseconds (ms), between 50 and 100 ms and higher than 100 ms indicates people with unhealthy, compromised health and healthy status, respectively [108]. In particular, Keiger et al. [108] demonstrated that people with SDNN24 higher than 100 ms have a 5.3 times lower risk of mortality at follow-up compared to those with SDNN24 lower than 50 ms. Moreover, it was found that SDNN24 values lower than 80 ms are predictive of cardiac events [102]. I will discuss SDNN24 in Chapter 6.

SDNNI SDNN Index (SDNNI) is the mean of the Standard Deviations of all the NN intervals, measured in segments of 5 minutes, for 24 hours. Even if SDNNI is measured over 24 hours, because the standard deviation is calculated on segments of 5 minutes, only components of HRV that are faster than 5 minutes are captured by SDNNI, unlike SDNN24, where most of the measured energy comes from frequencies lower than 5 minutes.

pNN50 The ratio between the number of pairs of successive inter-beat-intervals that differ by more than 50 ms and the total number of inter-beat-intervals (Equation (1.4)) [97, 215, 200, 203].

$$\text{pNN50} = \frac{\sum_{i=1}^{n} \mathbb{1}_{[0.05,\infty)}(|\text{NN}_{i+1} - \text{N}_i|)}{n} \tag{1.4}$$

where $\mathbb{1}_{[0.05,\infty)}(x)$ is the indicator function that is 1 if $x > 0.05$ and 0 otherwise.

Frequency Domain features

These features are calculated using Power Spectral Density (PSD) estimation, which describes the distribution of spectral power into frequency components composing that signal. Fourier techniques are often used to extract frequency domain features. However, Fourier analysis requires data to be evenly sampled, but the time-series of heart beats duration does not meet this requisite, because each event (heartbeat) will happen at slightly different interval with respect to its predecessors. For this reason, the Lomb–Scargle periodogram for PSD estimation was found to be the most appropriate method to analyze inter-beat-interval data [39].

Ultra-Low-Frequency (ULF) (i.e., power in frequencies lower than 0.003 Hz). In this range we find components that have period with duration between 5 minutes and 24 hours. This means that this frequency band captures periodic phenomena like circadian rhythm, and other slow changes in HRV. As will be discussed at length in Chapter 6, SDNN24 correlates with ULF, because the standard deviation captures the energy of the whole HRV spectrum over 24 hours, and in human HRV ultra-low frequencies dominate the other frequency bands, as shown in figure 6.1.;

Very-Low-Frequency (VLF) (i.e., frequencies between 0.003 and 0.04 Hz). In this range we find components that have period with duration between 5 minutes and 25 seconds. SDNNI correlates with VLF;

Low-Frequency (LF) (i.e., frequencies between 0.04 and 0.15 Hz). In this range we find components that have period with duration between 25 and 6 seconds. Primarily PNS, but also SNS, have an influence LF. LF correlates with SDNN, when measured over 5 minutes;

High-Frequency (HF) (i.e., frequencies between 0.15 and 0.4 Hz). In this range we find components that have period with duration between 6 and 2 seconds. This range captures mostly the activity of PNS. Respiratory Sinus

Arrhythmia is the physiological phenomenon with the strongest influence in this range. HF correlates with RMSSD, pNN50. Reduced resting HF is linked to altered activation of prefrontal cortex and consequently to a low cognitive functions that was found to be associated with stress, panic, anxiety or worry [96];

Total Power (Power in the LF and HF frequency ranges) were obtained by the sum of the power in the relevant frequency range in the spectrum. Total Power represents the overall energy expressed by the HRV spectrum, between 0.04 and 0.4 Hz.

Symphatovagal Balance Index (SVI) defined as the LF:HF ratio. The definition of the boundaries of the LF and HF bands varies slightly between authors, being a heuristic definition [179, 14, 215]. Because HF are dominated by the Parasympathetic Nervous System only, and LF is influenced by both the Sympathetic Nervous System and the Parasympathetic Nervous System, SVI is used to estimate the relative strength of the Sympathetic compared to the Parasympathetic Nervous System. High SVI values indicate Sympathetic activation. For this reason it's often used to estimate stress (Sympathetic activation, Parasympathetic suppression), and low SVI value indicate resting state (where Parasympathetic dominates over Sympathetic). However, the effective utility of SVI to estimate Autonomic Nervous System balance has been challenged [203], and as I will show in section 2 SVI is very sensitive to noise in the data.

Non-linear Time Domain features

Poincaré plot is a type of recurrence plot used to quantify self-similarity in processes. A Poincaré plot is a graph of RR intervals (RR_n) against the previous ones (RR_{n-1}). As shown in figure 1.6, from this scatter plot, it is possible to quantitatively analyze the variance of two consecutive RR-intervals by fitting an ellipse to the plotted shape. $SD1$ is the standard deviation of Poincaré plot perpendicular to the line-of-identity, while $SD2$ is the standard deviation of the Poincaré plot along the line-of-identity.

Entropy measures: Approximate Entropy; Sample Entropy Several measures of entropy have been applied to Inter-Beat Intervals: approximate entropy (ApEn) [173, 174], sample entropy (SampEn), fuzzy entropy (FuzzEn) [185], and Shannon entropy (ShanEn) [182]. Entropy in information theory was introduced by Shannon in 1948, and measures the amount of novelty in a communicated message, quantifying how surprising vs predictable the message is. Completely random messages will have maximum entropy, and completely predictable messages will have zero entropy. Shannon entropy is defined as $H(X) = -\sum_{x \in X} p(x) \log p(x)$ where x are the elements of a set X and $p(x)$ is their probability. Elements with very low or with very high probability will add very little entropy, because either $p(x)$ or $\log p(x)$ will be close to zero. A set with elements predictably frequent or rare will therefore have low entropy. On the other hand, a set where all elements have the same probability will have maximal entropy. Classical measures of entropy (such as Kolmogorov–Sinai entropy) require vast amount of data to estimate entropy

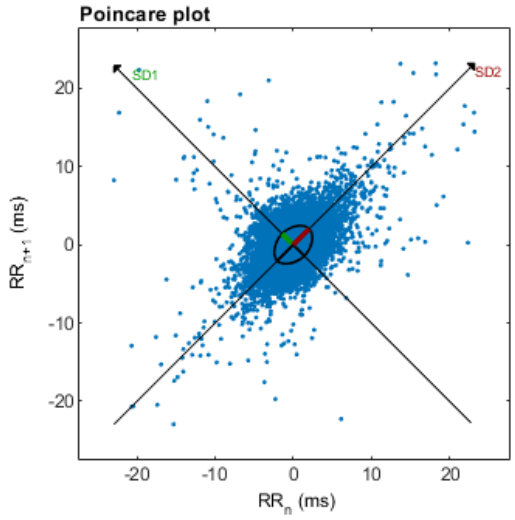


Figure 1.6: An example of Poincaré plot, showing the scatterplot of successive RR intervals, $SD1$, and $SD2$ axes. Credit: Johannes Mockenhaupt, wikimedia.org.

correctly, measures such as ApEn were developed to be resilient to limited amount of data. ApEn in particular was developed specifically to quantify the novelty and unpredictability of HR data, and has since then been applied to other domains (e.g. finance, psychology, etc.).

Detrended Fluctuation Analysis Detrended Fluctuation Analysis (DFA) is a modification of root mean square analysis of a random walk, and allows the detection of self-similarity present in non-stationary time series. DFA captures long term correlations within time series, at different time scales [169]. In the fractal analysis of the NN series, DFA yields two parameters, α_1 and α_2. A direct link between DFA and spectral analysis has been established, demonstrating that α_1 and α_2 are related to the conventional LF/HF and VLF/LF ratios. Consequently, these parameters should not be considered as novel, independent predictors [242]..

Correlation Dimension Correlation Dimension (CD) is a measure for the amount of correlations present in the NN time-series [144]. CD is a method used in chaos theory, useful to analyse non-linear system dynamics, in particular is used to estimate the dimensionality of a set of random points [18, 79]. CD was introduced by Grassberger and Procaccia in 1983 [79] as a measure of fractal dimension.

1.1.3 Known associations between HRV and Health outcomes

In this section, I examine the medical evidence regarding which health outcomes can be predicted by Heart Rate Variability. The analysis encompasses HRV measurements taken while the subject is in a stationary cardiopulmonary state, as well as HRV data recorded during exercise.

HRV at rest

HRV is traditionally acquired with the subject in stationary cardiopulmonary state (i.e. without recent physical activity), in a sitting position. HRV is calculated from 24 hours measurements of heart activity, 5 minutes (short), or less than 5 minutes (ultra-short) [203].

HRV is reported as predictor for several health conditions:

CVD Most studies look at HRV as a predictor of cardiac failure after the first episode [226]. In this study, I am interested in the probability of experiencing the first episode and how HRV can be utilised as a predictor. Hillebrand et Al. [87] report that 1% increased SDNN decreases risk of first CVD episode by 1% and the correlation depends on the duration of the measurement (low frequencies matter the most). Lanza et Al. [113] report that "in patients with unstable angina with preserved myocardial function, HRV (SDNN, SDANN, VLF, LF, HF, LF:HF) is an independent predictors of in-hospital and medium-term mortality". Zhang et Al. [249] found that HRV at night is predictive of CVD, indicating that circadian activity (not visible in short measurements like in traditional HRV analysis) might be informative of cardiovascular health. Low HRV (RMSSD, SDSD, pNN50, and to a lesser extent SDNN) and high Heart Rate are assciated to *coronary heart disease*, *myocardial infarction (MI)*, cardiac revascularisation procedures, and cancer; and a U shape was observed for cancer mortality, with higher risk in the lowest and highest tertiles of pNN50 [49, 194]. *Unstable angina* correlates with Low LF and LF:HF [194]. Values of SDNN24: below 70ms are associated with an increased risk of *Congestive Heart Failure (CHF)*, by a factor of 2.5. Other predictors include SD, SDNN, ULF, LF and DFA, and lower daytime LF predicts sudden death [194]. Rest, exercise, and ambulatory HRV is predictive of all-cause and cardiovascular mortality [213].

Left Ventricular Dysfunction (LV) At very early stage of LV development HR is attenuated because of reduced parasympathetic activation. Persistent inflammation with increased levels of cytokines (small proteins that are crucial in controlling immune system cells activity in the blood) plays a role in LV progression [163]

Stroke Lees et Al.[118] report that HRV features may have utility as a biomarker for stroke and for poststroke complications and/or functionality; SDNN, VLF, LF, and LF/HF derived from 24-hour ambulatory ECG differed significant between the 3 patient groups (with normal, mild dysfunction, and moderate renal dysfunction) and healthy controls. HF is the strongest predictor of stroke, then SDNN24, then LF.

Diabetes Lower SDNN, RMSSD and HF, higher LF are also visible in young subjects with type 1 diabetes, compared to healthy subject of the same age [94]. Diabetics patients had lower vagal tone (HF power) than non-diabetics after adjustment for age, race, and gender [226]. Both short term and 24h HRV are prognostic factors for type 1 and 2 diabetes, and are predictive of cardiovascular morbidity and mortality [163].

Hypertension Vagal tone is lower in subjects with hypertension: HF power adjusted for age, race, gender, smoking,

diabetes, and education was significantly lower in the hypertensive group; Low LF power was associated with the development of hypertension in men but not in women; RMSSD adjusted for age, race, study center, diabetes, smoking, education, and BMI was lower at baseline among those persons with hypertension [226].

Hypercolesterolemia Total cholesterol and low-density lipoprotein (LDL) were inversely associated with 24-h HRV (both in CVD and healthy groups); short-term HRV RMSSD was inversely related to LDL cholesterol [226].

Other Thayer reports that "HRV has been associated with immune dysfunction and inflammation, [...] CVD, diabetes, osteoporosis, arthritis, Alzheimer's disease, periodontal disease, and certain types of cancers as well as declines in muscle strength and increased frailty and disability" [226]

Other factors that affect HRV, that I will need to take into consideration in my experiments, include:

Smoking Nicotinic receptors affect the activity of the vagus nerve, and can affect High Frequencies of the HRV spectrum [163]. Vagal tone increased after one week of smoking cessation [226].

Lifestyle-related risk factors :

- a lack of physical activity and the abuse of tobacco, alcohol, and drugs have also been associated with autonomic imbalance and decreased parasympathetic activity [226]. Exercise increases HRV in healthy individuals. In patients with myocardial infarction, exercise is associated with changes in LF reactivity, higher, SDNN, RMSSD, pNN50 and lower LF:HF (both in hypertensive and normotensive subjects). In patients with CHF improves HRV to a lesser extent. In patients with unstable angina exercise ameliorates HRV but it's not a prognostic factor. In patients with diabetes there are no conclusive results [194].
- both acute and chronic smoking was associated with decreased vagal tone. Likewise, smoking was associated with a significantly increased LF:HF ratio [226].
- The single most replicable effect of aerobic training on cardiac function is a decreased resting HR (and increase vagal tone) [226].

Non-modifiable risk factors :

- RMSSD decreases on average 3.6 milliseconds every 10 years of age [226].
- reduced HRV have been found in individuals with a family history of CVD risk factors like hypertension and diabetes [226].

Work stress SDNN was significantly lower among those categorised in the high job strain group; shift workers had significantly decreased SDNNi [226].

HR during exercise and CVD

Heart Rate Variability has been studied mainly in a stationary status, with the subject at rest. Many HRV features have been proposed, on the time domain (SDNN, RMSSD, pNN50, etc); the frequency domain (HF, LF, etc); entropy measures.

In contrast, the dynamics of heart rate during exercise and how it changes with health status and health outcomes has not been studied with equal attention, as shown by the fact that there are no specific HRV features designed to capture meaningful aspects of the heart rate dynamics during exercise, and most features are not even usable if the HR is not stationary [230, 145, 171, 70]. This is surprising, because HR response during exercise is associated with cardiac death and HR to mortality [70], yet limited studies look into the prognostic significance of HRV during exercise. Maximum HR reached during exercise has been studied in correlation to CVD, autonomic imbalance may precede CVD symptoms and events [70]. Inadequate HR response during exercise is a predictor of sudden cardiac death, coronary artery disease, myocardial infarction, unstable angina, revascularisation [70]. Heart Rate Recovery (HRR) is affected by smoking status, age, fitness level, resting HR. Lower HRR is associated with coronary heart disease and all cardiovascular death [70].

1.1.4 Continuous monitoring enabled by wearable devices

Wearable physiological monitors have become increasingly popular, often worn during people's daily life, collecting data 24 hours a day, 7 days a week. In the last decade, these devices have attracted the attention of the scientific community as they allow us to automatically extract information about user physiology (e.g., heart rate, sleep quality and physical activity) enabling inference on their health.

Thanks to the recent technological advancements of wrist-worn wearable devices equipped with heart rate sensor, it is now possible to semi-continuously and passively measure cardiac activity during the people's daily life. For this reason and because of the low cost of these devices, they might have a great impact on the medical field due to the fact that it is possible to assess HRV features 24 hours per day, 7 days per week [81].

Wearable physiological monitoring, exploiting devices with unobtrusive (often wrist-worn) format, offers substantial promise for improving the care of patients in clinical settings, and for enabling individuals to better manage their own health. With recent advances in consumer markets, including devices such as fitness trackers and "smart watches", the use of wearable monitors is becoming increasingly commonplace. However, very few of these devices penetrate into use at scale within either clinical settings or for permitting patients to track their own health outside clinical environments [38, 223], with only small numbers of studies that have been described in the literature [205, 40, 200].

HRV and the involved physiological pathways is a field still actively studied, despite the fact that HRV has been studied for decades. In modern days, with the advent of Digital Signal Processing techniques, HRV and its relationship with health conditions has been extensively studied. Many studies have outlined how HRV can be used

Table 1.1: Summary of differences in HRV data using clinical grade and wearable devices

	HRV analysis with clinical grade devices	**HRV analysis with wearable devices**
Measurement device type	Clinical grade, reliable, low error	Commercial grade, unreliable, low signal to noise ratio
Measurement setting	Controlled (hospital visits)	Unreliable (daily life)
Cost	High	Low
Data is representative daily life	No (biased by clinical setting)	Yes (collected during daily activities)
Continuous passive monitoring possible	No (data collected in clinical environment)	Yes
Confounding factors (posture, caffeine, etc)	Can be controlled	Can not be reliably controlled

as a prognostic factor for many health conditions [87, 113, 112, 249, 49, 194, 213, 163, 118, 94, 226, 163]. With the advent of Machine Learning, many models have been proposed that can classify or even predict future health status and outcomes from HRV [142, 5, 36, 126, 32, 155, 208, 128, 2].

However, all the models found in literature were built on HRV data extracted from Electrocardiography (ECG) data, collected from reliable clinical grade devices, in controlled environment (i.e. the datasets are usually based on data collected during hospital visits). This ensures that the data reflects the underlying physiological process as closely as technically possible. However, this limits the applicability of such models to data collected during hospital visits: consumer grade ECG patches are available for the fitness market, but they it is not feasible to assume that, with the exception of athletes, normal population would wear them regularly, for their daily activities. It would be desirable to be able to use the abundant data collected by wearable devices to continuously monitor the health status of subjects, passively. This would enable the creation of models that could monitor the health status in the background, following the subtle changes of physiology for seasonal effects, normal physiological phenomena, but also for pathological phenomena. In theory this would allow the detection of the onset of health issues, potentially even before the symptoms are noticeable by the subject. Such a continuous, passive monitoring system, could operate in the background without requiring any user interaction, only requiring the use of a wearable device and access to data for analysis.

Table 1.1 shows a summary of the challenges and opportunities offered by the use of wearable devices for HRV analysis compared to clinical grade devices.

1.1.5 Error propagation in HRV analysis

A major obstacle to the use of wearable devices to extract HRV data in a remote patient monitoring, or in a hospital-at-home setting, is the lack of clinical validation: the lack of characterisation of their ability to estimate clinically-relevant physiological parameters, such as, in the case of cardiac applications, understanding the propagation of

error from the PPG waveform acquired within the device, through to estimation of the HR, and ultimately onto the HRV features that are used to track the state of cardiac health of a patient.

HRV data from wearable devices is collected in a non-controlled environment. The subject might misplace the sensor, or move excessively, causing missing data in the NN time-series. To extract HRV features, missing data needs to be interpolated. This process introduces artefacts in the time-series.

In this section I discuss the error induced by the PPG sensor (instead of the ECG sensor), and the effect of missing data in the final HRV features extracted.

HRV from Photoplethysmography data

HRV analysis is traditionally carried out via ECG collected in a controlled environment; analysis thereby proceeds on the assumption of noise-free data. Many HRV features aim to capture fast variability (high frequencies) in RR time-series, which is easily and adversely affected by the presence of outliers. For this reason, the time-series of RR intervals is typically reviewed manually, before their use in analysis.

The use of PPG instead of ECG for HRV analysis is not extensively explored, and especially for those cases in which the PPG is acquired via consumer-grade wearable devices. While both ECG and PPG may appear to capture aspects of the same physiological phenomenon, ECG records the electrical activity responsible for heart contractions, whereas PPG records the peripheral circulation's response to these contractions. Consequently, PPG is susceptible to interference from various factors, such as blood pressure and artery stiffness, which can introduce variability in the signal used for HRV analysis. Various studies have explored the accuracy of HRV features extracted from clinical-grad PPG devices [17, 97, 127, 132, 195, 202, 225], and the quality of HRV from a PPG acquired from smartphones has been described [170]. These studies have, in general, concluded that HRV analysis from PPG acquired under carefully-controlled conditions is reliable, because most HRV features estimated via PPG show a very high correlation with the same features estimated via ECG. A notable exception to the latter is pNN50 [97, 132], which was shown to be less reliable when estimated via the PPG than from the ECG.

In section 2 I study the propagation of errors induced by motion artefacts in data collected from a PPG wearable device onto HRV features. I discuss how most noise is caused by artefacts induced by motion. The difficulty to obtain motion free data during the day might be the reason why many wearable device manufacturers opted to report HRV analysis only for data collected during sleep [23].

Effects of missing data

During the day, approximately 1% of beats are to be expected to be ectopic [103] when they are recorded using gold standard instruments (i.e., Electrocardiography). An ectopic beat is a disturbance of the cardiac rhythm that induces premature ventricular or atrial contraction. The physiological artefact producing inconsistent beat heavily affects the HRV spectrum in the high frequency range, and could result in erroneous results [168]. The

removal of ectopic beats is crucial to ensure the accuracy and reliability of the derived measures [22]. Ectopic beats, which are premature or abnormal heartbeats originating from locations other than the Sinoatrial Node, can introduce irregularities in the heartbeat intervals. These irregularities may lead to misinterpretations of HRV parameters, potentially confounding the analysis by introducing artificial variability that is not reflective of the autonomic nervous system's influence on the heart. Besides physiological artefacts, the presence of motion and mechanical artefacts induced by external stimuli can result in a significant number of inconsistent beats when data are recorded using wrist-worn wearable devices [107, 39]. Since missing data are common in the NN-interval time-series derived from wrist-worn wearable device, they could complicate the analysis of HRV features making it sometimes impossible. To make reliable HRV analysis, previous studies suggested several pre-processing methods for NN-intervals time-series (e.g., deletion, interpolations and filtering). However, these pre-processing methods have their own distinct effect on HRV analysis yielding different results [168].

The simplest way of handling the inconsistent NN-intervals provided in literature is to delete them [197]. In this approach, the abnormal NN-intervals are removed and the remaining NN-intervals are merged together. A limitation of the deletion approach is that it reduces the overall length of the HRV signal. This may significantly influence HRV spectrum [197]. Other interpolation methods maintain the original number of samples, but, by manipulating the duration of NN-intervals, they also change the overall duration by some amount. There are several interpolation approaches useful for handling inconsistent NN-intervals, i.e., zero degree, linear and cubic spline [138]. Zero degree replaces the inconsistent NN-intervals with the mean of the closest normal values. Linear interpolation fits a straight line over the inconsistent NN-intervals to obtain normal values. The most popular interpolation approach is the spline of order three (i.e., cubic spline). It fits a third degree polynomial smooth curve through a number of data points to obtain new values. This approach is recommended when there is only small number of inconsistent NN-intervals [138].

Finally, it was found that the interpolation introduces power in the low frequency band (LF) and reduces high-frequency components (HF) power [39]. This aspect affects frequency domain HRV features, while little effect was found in time domain HRV features [107].

1.1.6 Limitations of conventional HRV features to represent heart dynamics

Time vs duration

The entirety of HRV analysis is built on top of inter-beat-interval time-series, obtained by calculating the difference between the timestamps of consecutive detected heartbeats. This moves the analysis from the domain of time to the domain of duration. This first step is always present in all processing pipelines and is the prerequisite for all HRV features, in time, frequency, and non-linear domains.

However, as I will discuss in detail in Chapter 3 this has the effect of discarding important information when filtering out outliers from the NN-intervals sequence, such as ectopic beats, or motion artefacts. Removing heartbeats

from the NN-interval sequence, or attempting to interpolate, results in shifting the whole sequence, with effects on the whole spectrum: high frequencies will be affected by the interpolation method; and low frequencies will be affected by the drift induced by shifting the data.

Imagine a signal composed by a perfect sinusoidal signal. Its spectrum will be composed by a single peak. Now imagine shifting the signal in a single point, e.g. removing or adding some samples at random places. The spectrum of the modified signal will have more energy in the high frequencies (necessary to explain sudden changes of instantaneous heart rate), and what once was a single peak will become less defined, with a sudden change in phase of the main signal. Different interpolation strategies will all affect the spectrum, in different ways.

As will be discussed in Chapter 3 by interpolating missing heart beats in the timestamps time series, instead of in the duration time series, it is possible to mitigate this phenomenon. It is surprising that such a simple mitigation technique is not common practice in HRV analysis.

Equivalence of conventional HRV features

Throughout the evolution of HRV analysis, numerous features have been suggested; however, a pervasive challenge lies in their substantial correlation and overlap, coupled with a tendency to encapsulate limited information and address only specific subsets of physiological phenomena. For instance, many existing features predominantly focus on frequencies shorter than 5 minutes, thereby restricting the analysis to the Autonomic Nervous System (ANS) dynamics while overlooking the intricate contributions of the Sinoatrial Node (SAN). In the pursuit of advancing our understanding of cardiac regulatory mechanisms, there arises a critical need to develop more comprehensive and nuanced HRV features that transcend these limitations and offer a more holistic representation of the underlying physiological processes. In section 1.1.2 I have presented the HRV features conventionally used in HRV analysis. Most conventional HRV features are strongly correlated between each other [187, 22]:

- SDNN measures the overall power of the measured HRV spectrum. when measured over 5-minute windows is influenced by both PNS and SNS [203]. When measured over 24 hours is mostly affected by ultra low frequencies that, as discussed in section 1.1.1, are affected by SAN. As will be discussed in Chapter 6, SDNN24 is the "gold standard" for cardiac health [203].

- SDNNI, assessed across a 24-hour duration, represents the mean standard deviation of consecutive 5-minute data windows. However, its reliance on 5-minute data intervals leads to the exclusion of contributions from components with periods longer than 5 minutes, such as ultra-low-frequency variations. Consequently, this limitation results in an oversight of Sinoatrial Node activity, as the nuanced changes occurring on a longer timescale are disregarded. Furthermore, the averaging nature of SDNNI diminishes the impact of rapid fluctuations, ultimately culminating in a correlation with very-low-frequency components.

- pNN50 performs a non-linear filtering of changes in HRV, applying a hard filter at 50 msec to the considered

data. This means that it only considers contributions from fast components of HRV (PNS). It correlates with HF and RMSSD.

- RMSSD looks at the difference between consecutive heartbeats, calculating the total power of their differentials. RMSSD is identical to SD1 (Poincaré Plot), correlates with HF, and is affected by PNS.
- ULF band is only measurable with recordings of 24h of duration, and focuses on frequencies lower than 0.003 Hz. As we will see in Chapter 6 it is influenced by SAN activity, it correlates with SDNN24, and can be estimated from semi-continuous measures of HR (without 24h ECG), i.e. from a fitness tracker.
- VLF band focuses on frequencies between 0.003 and 0.04 Hz. As discussed it correlates with SDNNI.
- LF band look at frequencies between 0.04 and 0.15 Hz, and is normally estimated on measurements between 2 minutes and 5 minutes. It is affected by both the activity of PNS and SNS. No strong correlations with other HRV features are known.
- HF band focuses on frequencies between 0.15 and 0.4 Hz. Respiration is a major contributor. SNS does not affect HF, but PSN does. HF has strong correlation with all the HRV features that focus on fast changes, such as RMSSD and pNN50.
- SD1 and SD2 (from the Poincaré Plot) correlate with LF and HF power, and with RMSSD [203].
- Frequency domain features show strong correlation with Entropy measures [187].

Lack of focus on SAN activity

HRV analysis is based on heart activity measured either over 24 hours, or more frequently over 5 minutes. The choice of 5 minutes can be explained by the need to collect at least a few complete cycles of the components of the spectrum in the LF range: 0.04 Hz, that corresponds to a period of 24 seconds. In 5 minutes we can collect 10 complete cycles of a signal at 0.04 Hz. This will allow us to estimate the spectrum in the LF range with sufficient accuracy. 5 minutes can therefore be explained as a practical way to obtain the minimum data necessary to extract information about a relevant physiological phenomenon without requiring unnecessary lengthy procedures. However, this choice implicitly puts all the focus on ANS activity, and ignores SAN.

HRV in non-stationary cardiac state

Another limitation of traditional HRV features is that they necessitate the stationarity of the cardiovascular system. Time-domain features rely on the assumption that there is no trend in the NN-intervals series. For example, SDNN (being a standard deviation) would be strongly affected by the presence of a trend in the data. For this reason HRV analysis is almost uniquely concerned with measures of the subject in a stationary cardiopulmonary condition.

Nevertheless, understanding how the heart responds to physical activity provides valuable insights into overall cardiopulmonary fitness and health. Although there is notable research on heart rate variability during physical exercise [145, 95, 186], the predominant focus has been on HRV at rest, reflecting a more extensive body of literature. Few HRV features have been specifically designed to characterise heart dynamics in non-stationary cardiovascular conditions, highlighting a gap in the exploration of HRV beyond resting states. This gap will be further addressed in section 1.1.3.

Among them is Heart Rate Recovery (HRR), which focuses only on the reduction in beats per minute of the heart rate right after the subject stopped performing physical activity and came back in a resting state, reducing the complexity of the physiological reaction of the cardiovascular system to physical activity to a single number. HRR has been measured to be independent from HRV at rest in the same subjects [95], suggesting that it might be expression of different physiological phenomena, therefore potentially containing additional useful information. However, the exploration of this potential could have been more extensive, and relying on a single metric (the reduction in heart rate from peak to 1 minute after the end of physical exercise) is unlikely to capture the full spectrum of nuances in how efficiently the cardio-respiratory system responds to increased demand. Therefore, I posit that a more comprehensive model of the time series of heart rate and physical workload could enhance the development of cardiovascular health prediction models. In Chapter 7, I discuss the construction of a detailed model of heart dynamics during exercise, and subsequently, I also use this model to predict health outcomes.

1.2 Research Shortcomings addressed in this thesis

In this Chapter, a critical review of the existing literature has revealed several noteworthy gaps which this research aims to address. Firstly, while substantial work has been done on HRV analysis using data from clinical-grade devices, there is a paucity of research utilising data from consumer-grade, wearable devices. This thesis seeks to explore the validity and applicability of HRV data derived from such devices in continuous, passive health monitoring.

Another significant gap identified is the lack of detailed characterisation of the error propagation in HRV features when using data from wearable devices. The existing literature mostly focuses on data quality from controlled, clinical settings, leaving a research void in understanding how motion artefacts, missing data, and other common issues in wearable device data affect HRV analysis.

Furthermore, conventional HRV analysis primarily focuses on short-term variability, often neglecting the Sinoatrial Node activity and the impact of non-stationary cardiovascular states on HRV. This research aims to extend the understanding of HRV by exploring these less-studied aspects, particularly how SAN activity and non-stationary cardiac states influence HRV.

In addition to these identified gaps, this thesis extends the state of the art in several key areas. Firstly, it delves

into the analysis of how interpolation techniques for missing data affect HRV features. Current literature often overlooks the impact of different interpolation methods on the accuracy of HRV metrics. This study systematically evaluates various interpolation techniques, providing insights into their implications for HRV analysis in both clinical and non-clinical settings.

Traditionally, time series analysis in chronobiology utilises cosinor-based rhythmometry, often relying on single component cosinor models for their simplicity and ease of interpretation, despite their limited fit due to the complex nature of circadian rhythm effects on heart rate. More sophisticated approaches, like multiple component cosinor models or Machine Learning models, provide a more accurate fit but lack the interpretability of simpler models. There is a clear need in chronobiological analysis for a model that not only accurately fits the data but also maintains ease of interpretation, effectively bridging the gap between complexity and practicality.

Sinoatrial Node activity is often overlooked in HRV analysis due to its requirement for data spanning multiple hours, whereas traditional HRV analysis is based on inter-beat intervals and is typically feasible only for short durations, usually a few minutes. SDNN24, which is the standard deviation of NN intervals over a 24-hour period, is regarded as a predictor of cardiac health. Its calculation typically requires 24 hours of ECG data, often gathered using a Holter monitor. The complexity of this measurement significantly limit the practicality of SDNN24 as an early predictor of cardiac deterioration. If it were feasible to passively and continuously collect the necessary data, for instance through commercial-grade fitness trackers, SDNN24 could be more effectively utilised for early cardiac health monitoring.

Finally, although Heart Rate Variability has been extensively researched in resting states, there is a scarcity of studies on HRV during physical activity and its potential for predicting health outcomes, specifically investigating the unique insights HRV can offer beyond what is available from Electronic Health Records (EHR). This could significantly expand the application of HRV in preventive healthcare and patient monitoring. This research gap highlights the need to explore HRV's full potential in different physiological states.

In summary, this thesis will address these gaps through a series of experiments and analyses, with the overarching aim of enhancing the predictive power and applicability of HRV analysis in a range of health contexts.

1.3 Outline of the thesis

The reminder of the thesis is structured as follows:

- In the second, third and fourth Chapters of the thesis, I investigate the sources of noise in HRV analysis, with a particular emphasis on data passively collected from wearable devices. This exploration is motivated by the recognition that error propagation in this context has not been thoroughly examined:
  - Chapter 2 is based on a paper published on *Healthcare Technology Letters* [149]. In this Chapter I analyse how noise in inter-beat-intervals collected from a PPG sensor on fitness trackers propagates to

HRV features;

- Chapter 3 is based on a paper published on *Sensors* [153]. In this Chapter I study the impact of commonly used strategies to compensate for missing data in the inter-beat-intervals time series and propose a different approach that minimised reconstruction error;
- In Chapter 4, based on a paper published on *Sensors* [193], I look at ultra-short HRV features and how their estimation is affected by motion artefacts and missing data, both phenomena heavily present in data collected form wearable devices.

- In Chapters 5 and 6, I scrutinise the utilisation of routinely collected heart rate (HR) data from fitness trackers to estimate crucial aspects of cardiac health. This investigation is prompted by the recognition that the exploration of ultra-low frequencies, Sinoatrial Node dynamics, and circadian effects on HRV data derived from wearable devices has not been thoroughly addressed in the existing literature:
  - In Chapter 5, based on a paper published on *Pysiological Measurements* [150], I present a model to analyse the impact of circadian rhythm on resting heart rate. I develop a methodology that allows for accurate fitting of the data while preserving the ease of interpretation of the cosinor model commonly used for circadian analysis, reducing the complexity of a model with many components to only 3 free parameters;
  - In Chapter 6, based on a paper published on *Sensors* [152], I show how the SDNN24 HRV feature (gold standard for cardiac health) can be estimated from HR data only, commonly collected by fitness trackers, instead of requiring a Holter device. This enables the continuous monitoring of cardiac health, using simple fitness trackers, allowing widespread passive screening and early detection of cardiac activity deterioration.

- In Chapter 7, I detail the outcomes of an experiment using UK Biobank data to identify health outcomes predictable by HRV during exercise. This involved training a BERT-based large language model on Electronic Health Records (EHR) to encapsulate each participant's medical history into a fixed-size vector. The study aimed to predict future diagnoses of various health outcomes, utilising HRV data to assess its independent informative value. This experiment might indicate for which health conditions it would be valuable to consider data collected by fitness trackers in conjunction with EHR, to assess risk.

- In the last Chapter I draw my conclusions about my work, and discuss potential future work.

Chapter 2

Profiling the Propagation of Error from PPG to HRV Features in a Wearable Physiological-Monitoring Device

Wearable physiological monitors are becoming increasingly commonplace in the consumer domain, but in literature there exists no substantive studies of their performance when measuring the physiology of ambulatory patients. In this Chapter, I investigate the reliability of HRV analysis from inter-beat-intervals collected from commercial grade wrist worn wearable devices. I will study the error propagation from raw data to some of the most used HRV features in literature in the time domain and in the frequency domain. Our experiments quantify the propagation of error from (i) PPG acquired by pulse oximetry, to (ii) estimation of heart rate and the inter-beat-interval time-series, and (iii) subsequent calculation of HRV features. Our experiments confirm that motion artefacts account for the majority of this error, and show that the unreliable portions of heart rate data can be removed, using the accelerometer sensor from the wearable device. Our experiments further show that acquired signals contain noise with substantial energy in the high-frequency band, and that this contributes to subsequent variability in standard HRV features often used in clinical practice. I finally show that the conventional use of long-duration windows of data is not needed to perform accurate estimation of time-domain HRV features.

2.1 Introduction

In this study, I investigate this propagation of error, with the aim of improving understanding of the accuracy of the estimated HRV features; I use as my experimental device a commonly-used wearable consumer device, the

Microsoft *Band 2*[1]. This study was conducted in late 2018. The Microsoft Microsoft *Band 2* has since been discontinued, but at the time it was providing developers open access to the data that was able to collect. The sensors included: optical heart rate monitor; three-axis accelerometer; gyrometer; GPS; microphone; ambient light sensor; galvanic skin response sensors; UV sensor; skin temperature sensor; capacitive sensor; and barometer. This device was selected because it allowed access to raw accelerometry data as well as to inter-beat-intervals (most commercial wearable devices only allow access to analysed data, i.e. steps and HR, not low level data, i.e. inter-beat-intervals and triaxial accelerometry data). The most commonly-occurring mode of error in such wearable devices has previously been identified as being due to movement artefact [164], which is unsurprising given the sensitivity of the pulse oximetry process (which yields the PPG waveform from which HR and HRV parameters are estimated) to the small changes in light intensity that arise due to movement of the patient. However, while some studies have investigated how to mitigate the effects of motion artefacts correcting the raw PPG signal [83, 116], I could find a small amount of studies investigating the possibility to assess PPG reliability from accelerometer [117], and no study investigating the propagation of error from PPG to HRV features.

This investigation provides the following contributions: (i) investigation of a possible approach to determine when the time-series of pulsatile intervals (estimated from the PPG) should be discarded; (ii) estimate the expected error on subsequently-derived HRV features as movement of the patient increases; and (iii) suggest a means of filtering to minimise the error of the HRV features.

It has been reported that use of the PPG for HRV analysis is only reliable if the cardiovascular system is in a stationary condition [225]; i.e., that the PPG cannot reliably be used when the subject is recovering from physical activity. In my experiment, I investigate this finding (and ultimately find no difference in accuracy after a short period of exercise than when compared with PPG from a resting condition). I here describe my investigation of the accuracy of HRV features estimated from a wrist-worn PPG sensor in a *non-controlled* environment, as would be represented by a typical setting for the use of wearable devices. This is a crucial preliminary task in HRV analysis from wearable devices, and, to my knowledge, no study has yet been carried out on this subject.

To my knowledge this is the first analysis of the error propagation from inter-beat intervals to HRV features. This Chapter is based on work published in *Healthcare Technology Letters*[149] in 2018. The research and results presented here were conducted in collaboration with Bartoloni, L and Colombo, M, at that time colleagues at BioBeats, who contributed with collecting the data and running the analysis.

2.2 Experimental Design

I collected 30 minutes of data from the wearable device (the Microsoft *Band 2*) of 5 subjects, all male, healthy, Fitzpatrick scale[2] Type III (medium white skin tone), mean age 32 (standard deviation 6). These data include (i)

[1] `https://en.wikipedia.org/wiki/Microsoft_Band_2`
[2] `https://en.wikipedia.org/wiki/Fitzpatrick_scale`

a time-series of the accelerometry for each of three orthogonal axes of motion and (ii) a time-series of pulse-to-pulse intervals, estimated by the band from its PPG sensor. I concurrently collected R-R intervals using an ECG sensor (acting as the reference device, with limitations discussed in section 2.5) from a *Polar H7* chest-mounted strap. I will hereafter use the term *P-P intervals* to refer to the pulse-to-pulse intervals (acquired from the wrist-worn wearable PPG device), and *R-R intervals* to refer to beat-to-beat intervals (acquired from the chest-mounted ECG system).

Our experimental protocol comprises four sections:

rest For $t = [0\ 300]$ s, the subject was requested to sit as still as possible, refraining from talking or moving the arm on which the wearable PPG system was mounted. This first section of the experiment is aimed at establishing a reliable baseline for both motion artefacts and resting heart rate.

stress For $t = [300\ 600]$ s, the subject was requested to climb and then descend two storeys of stairs without stopping. This second section is aimed at measuring the error induced by motion artefacts, and to raise substantially the heart rate of the subject.

recovery For $t = [600\ 1200]$ s, the subject was again requested to sit as still as possible. The aim of this third section is to establish the propagation of error from a PPG sensor, in absence of movement, when the subject is not in a stationary cardiac state.

free For $t = [1200\ 1800]$ s, the subject is requested to remain sitting, but is permitted to perform normal seated activities; e.g., talking, using a laptop computer, etc. This last section of the experiment is aimed at collecting data that is representative of PPG data that can be expected to be collected in users every-day life, when they are not explicitly asked to stay still, and are free to perform daily activities. This is the data that can be expected to be used to HRV analysis at rest with the subject awake, and is therefore interesting to quantify the expected error present in HRV features collected in this condition.

2.3 Methods

This section describes the steps taken in analysing the data acquired from each subject. Aggregated results for all subjects will be presented in section 2.4.

Accelerometry

From the three time-series of accelerometry x_t, y_t, z_t, corresponding to motions in each of three orthogonal directions, I consider absolute magnitude of the first-order differences, which is invariant to rotations of the device:

$$w_t = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2 + (z_t - z_{t-1})^2} \tag{2.1}$$

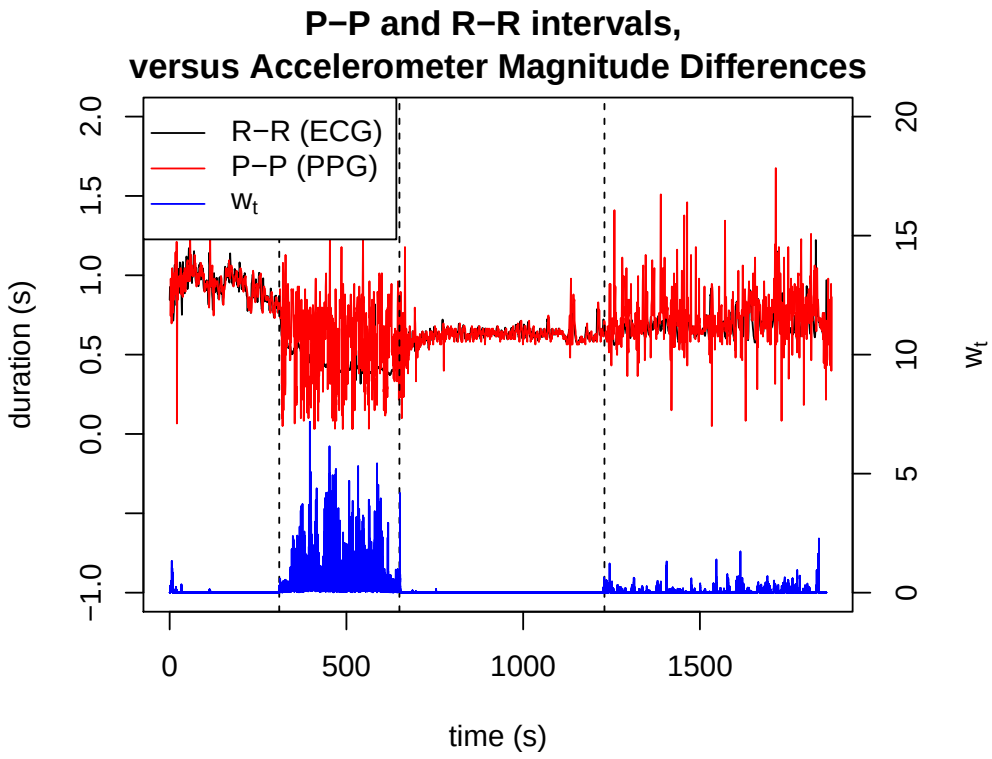


Figure 2.1: Time-series of P-P and R-R intervals acquired from PPG and ECG (upper plot), and corresponding accelerometry time-series w_t (lower plot) for an exemplar patient, where divisions between the four sections of the experimental protocol are shown as vertical dashed lines.

which are subsequently averaged over a window of duration τ, containing n samples,

$$W_t(n) = \frac{1}{n} \sum_{i=t-n}^{t} w_i \tag{2.2}$$

The time-series w_t is shown for an exemplar subject in figure 2.1. As expected, it may be seen from the figure that the accelerometry time-series increases with the activity undertaken during the second phase of the experiment (in which the subject was climbing and descending stairs), and that there was very little activity during the first and third sections (during which the subject remained still). The "free-action" fourth section shows moderated accelerometer activity, with occasional transients, arising from rapid and infrequent movements of the wrist on which the device was worn.

The figure also shows the differences between P-P and R-R intervals estimated from the PPG and ECG sensors. During the first and third sections, the time-series of P-P and R-R intervals are similar. As motion increases, it may be seen that the fourth section corresponds to frequent substantive differences between P-P and R-R intervals, with frequent spikes in P-P intervals not present in R-R intervals, and that these differences become larger during the increased activity of the second section.

2.3.1 Outlier Removal

Before subsequent processing, I performed outlier removal to discard P-P intervals with physiologically-implausible values; e.g. increased P-P intervals (and therefore HR) that occur faster than could be feasibly produced by physiology. This was straightforwardly performed by computing a 10s moving average $\mu_{10}(t)$, and then discarding P-P intervals $PP(t)$ for which $\frac{|PP(t)-\mu_{10}(t)|}{\mu_{10}(t)} \geq 0.5$.

2.3.2 HRV Analysis

I estimate some of the HRV features defined earlier in section 1.1.2: AverageNN (Average heartbeat duration), RMSSD (Root Mean Square of Successive Differences); the percentage of inter-beat intervals that differ more than 50 milliseconds from the previous interval (pNN50); SDNN (Standard Deviation of NN intervals); and Symphatovagal Balance Index, the ratio of Low Frequencies and High frequencies (SVI). These are estimated, together with W_t, over a short-time window of duration $\tau = 40$ s, with subsequent windows overlapped by 20 s.

Figure 2.2 shows HRV features estimated for an exemplar subject. During the second section of the experiment, it may be seen that the increased movement of the subject results in substantial errors in the PPG signal that propagate to the HRV features.

For this exemplar participant, I can see that the standard deviation of the accelerometry time-series W_t is correlated to the magnitude of the errors in the time-domain HRV features, with correlation coefficients: $\rho = 0.91$, 0.84, 0.67, and 0.64 for AverageNN, SDNN, pNN50, and RMSSD, respectively. However, this correlation does not exist for the frequency-domain HRV feature, SVI, for which the corresponding correlation coefficient is $\rho = -0.03$.

2.3.3 Discarding Data with Low Signal Quality

Having confirmed above the intuition that error in HRV feature estimation is correlated with movement, I subsequently choose to estimate signal quality via the accelerometry time-series W_t, such that HRV estimates may be discarded if W_t exceeds some threshold κ.

I therefore calculated the distribution of error for each HRV feature, as a function of the value of W_t (the average of w_t over the window of duration τ), and noticed that error for most of the HRV features (Average NN, SDNN, RMSSD, and pNN50) quickly increases for $W_t > 0.02$: the error for Average NN goes from 0.0013 ($W_t < 0.02$) to 0.0142 ($W_t > 0.02$); the error for SDNN goes from 0.0061 ($W_t < 0.02$) to 0.0139 ($W_t > 0.02$); the error for RMSSD goes from 0.0079 ($W_t < 0.02$) to 0.0153 ($W_t > 0.02$); the error for RMSSD goes from 0.057 ($W_t < 0.02$) to 0.125 ($W_t > 0.02$). SVI is the only HRV feature whose error doesn't change with $W_t > 0.02$. Following the observation that, for most of the HRV features, the error is an order of magnitude lower than the value of the corresponding HRV feature for $W_t < 0.02$, and that the error becomes of the same order of magnitude as the value of the HRV feature for $W_t > 0.02$, I choose a threshold of $\kappa = 0.02$ for the remainder of this work. I emphasise that this is a

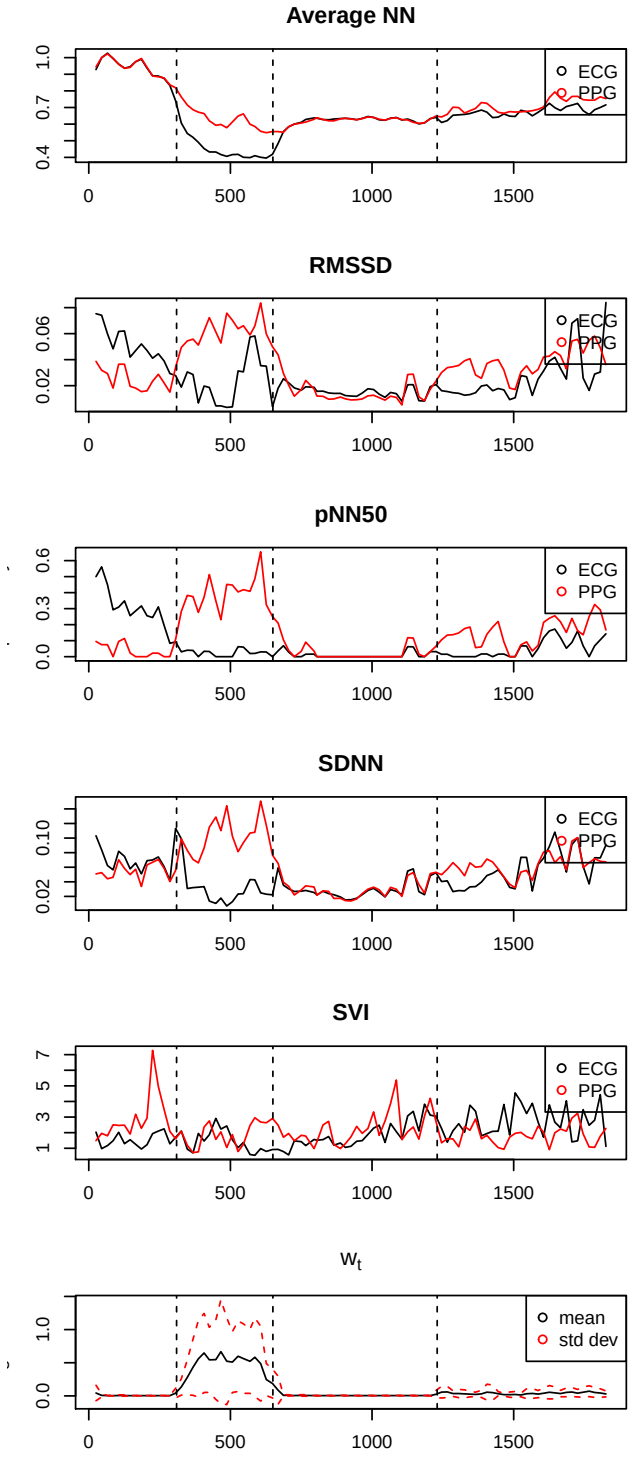


Figure 2.2: HRV features estimated from P-P intervals acquired from PPG (red) and from R-R intervals acquired from ECG (black), in reading order, for one of the subjects in the study. The lower plot shows the time-series of accelerometry W_t, with mean and one s.d. values within a window of $\tau = 40$ s shown in black and red, respectively.

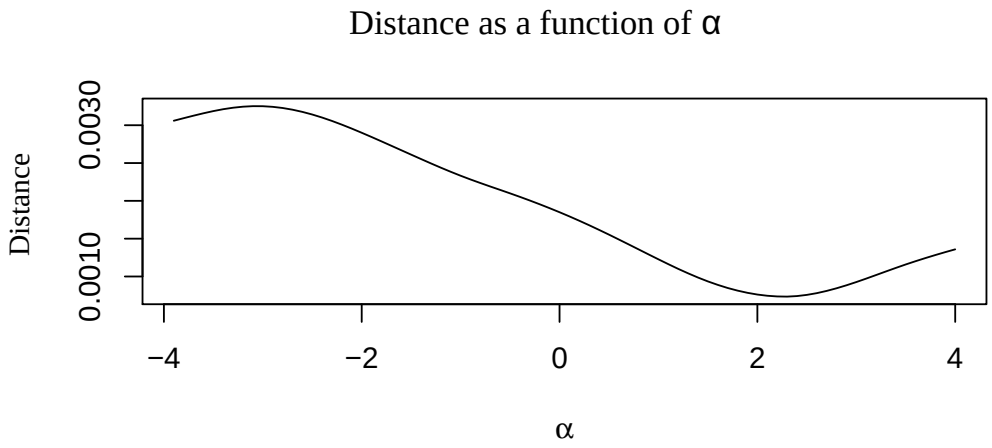


Figure 2.3: Squared-distance d between pre-processed P-P and R-R intervals, as a function of time-offset α between them (seconds).

candidate κ, given the prototype nature of the study, conducted on a small patient group, without a test set, and which would naturally be more principled in formulation for a subsequent larger study.

2.3.4 Signal Realignment

The final step in the pre-processing is to find (and correct for) the time-offset between the PPG and ECG signals. Even in presence of perfect synchronisation between the PPG and ECG devices, the P-P and R-R timeseries are not aligned for physiological reasons: the ECG signal detects the electric impulse of the heart, that is instantaneous with the heart contraction, while the PPG sensor is placed on the wrist, and it detects the flow of blood in the subject circulatory system, as shown in figure 1.5. The blood takes time to flow from the heart to the wrist [177], and this delay between ECG and PPG can vary as a function of many factors, including blood pressure, artery stiffness, subject's height, and should therefore be calculated independently for each subject.

The alignment is performed by resampling both P-P and R-R timeseries at 10Hz, and finding the value of α that minimises the squared-distance d between the P-P intervals and the R-R intervals lagged by a time-offset α.

Figure 2.3 shows this relationship for an exemplar patient, which is reproduced across other subjects (not shown here for brevity). In this case I choose a value of $\alpha = 2.2$ s, which corresponds to the value that minimises d.

2.4 Results

This section presents results of (i) investigating the effect of changing window-size τ when calculating HRV features; (ii) investigating the signal-to-noise ratio for the time-series of P-P interval estimated from the PPG, with respect to the reference R-R intervals derived from the ECG; and (iii) the distribution of errors for HRV features throughout

Table 2.1: RMSE of HRV features, for window size τ

τ (s)	**AverageNN**	**SDNN**	**RMSSD**	**pNN50**	**SVI**
40	0.0480	0.0334	0.0564	0.2285	0.6471
60	0.0489	0.0316	0.0566	0.2294	0.5683
80	0.0499	0.0294	0.0575	0.2279	0.6080
100	0.0504	0.0287	0.0583	0.2330	0.5982
120	0.0506	0.0280	0.0575	0.2259	0.6613

the four stages of the experimental protocol.

2.4.1 Effect of Window Size

Table 2.1 shows the root mean-square error (RMSE) of each HRV feature for various window sizes τ. It may be seen that the RMSE does not change substantially with varying τ errors increase with window size for AverageNN ($\approx 10\%$), RMSSD ($\approx 10\%$), and pNN50 ($\approx 2\%$), while the RMSE decreases for SDNN ($\approx 25\%$) and SVI ($\approx 30\%$). However, it may be seen from the table that the decrease of SVI is not proportional to changes in τ.

It is common practice to require long-duration windows of data to estimate HRV features reliably. The experiment shows that short-duration windows of data still allow to perform accurate estimation of time-domain HRV features.

2.4.2 Signal-to-Noise Ratio

Noting that the ECG-derived time-series of R-R intervals RR is used as the reference, we can define the time-series of residuals $r_t = RR - PP$, where the time-series of re-aligned P-P intervals from the PPG is PP.

Figure 2.4 shows the spectra of RR compared with the residual time-series r_t when calculated with and without pre-processing of the time-series of P-P intervals derived from the PPG.

Additionally, we can plot the signal-to-noise ratio (SNR) between RR (the "signal") and $r_t = RR - PP$ (the "noise") for the cases with and without pre-processing of PP, as shown in figure 2.5.

Figures 2.4 and 2.5 show that the power in the signal exceeds that of the noise for the ultra-low frequency band (i.e., those frequencies below LF). This is an expected result, because power in this band corresponds to slowly-changing physiological phenomena that modulate the time-series of P-P and R-R intervals - this corresponds to heart rate and its slow changes after the effect of physical activity (increased HR during physiological stress, and reduced HR during recovery). Heart Rate is resilient to errors because is proportional to the average Inter-Beat-Interval, and computing the average is equivalent to applying a low pass filter on the original time-series, which has the effect to remove outliers. This is an expected result: motion artefacts will induce spurious peaks in the PPG signal, resulting in quick changes in the Heart Rate. These changes are expected to mainly induce instantaneous changes in the time series and therefore affect high frequencies. As discussed, Heart Rate is the result of a low pass filter on the Inter-Beat Interval time series, therefore most of the induced noise will be filtered out and HR

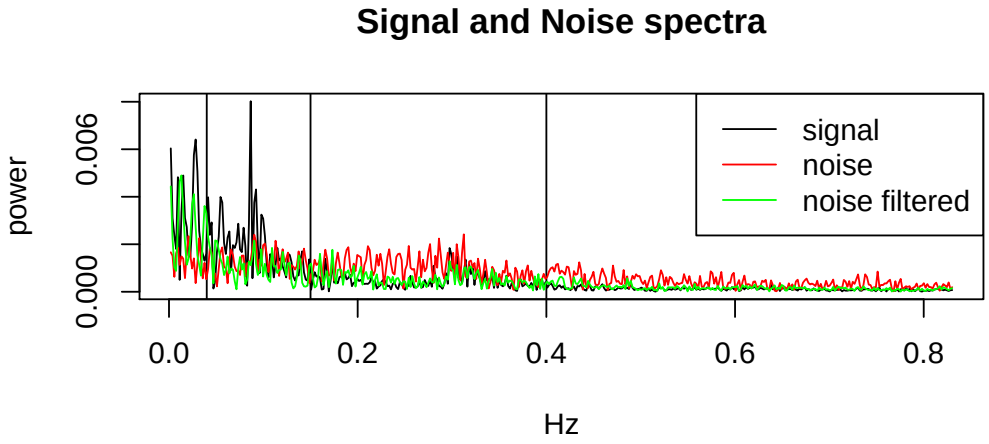


Figure 2.4: Spectra for the RR signal vs. residuals r_t where the latter have been calculated from the original ("noise") and pre-processed ("noise filtered") signals PP, shown in red and green, respectively. The vertical lines show the boundaries of the frequency bands used in HRV analysis: LF between 0.04 and 0.15 Hz, and HF between 0.15 and 0.4 Hz.

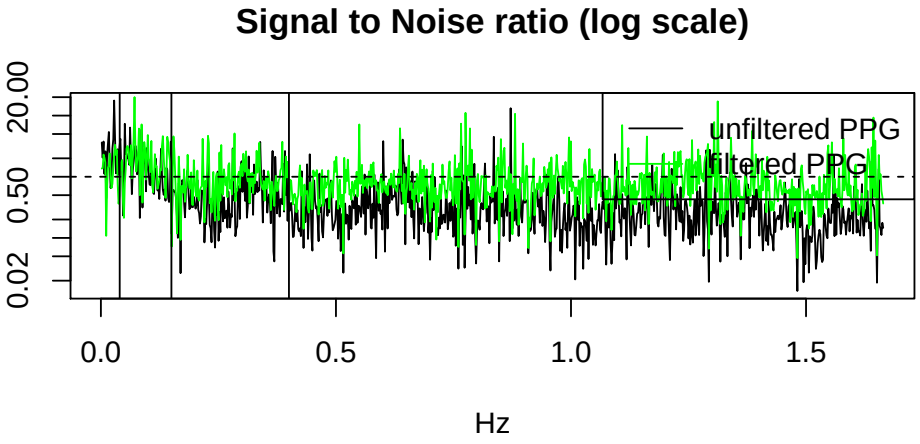


Figure 2.5: SNR between RR (signal) and residuals r_t (noise), with and without pre-processing of PP shown in green and black, respectively. A horizontal line shows SNR = 1.0.

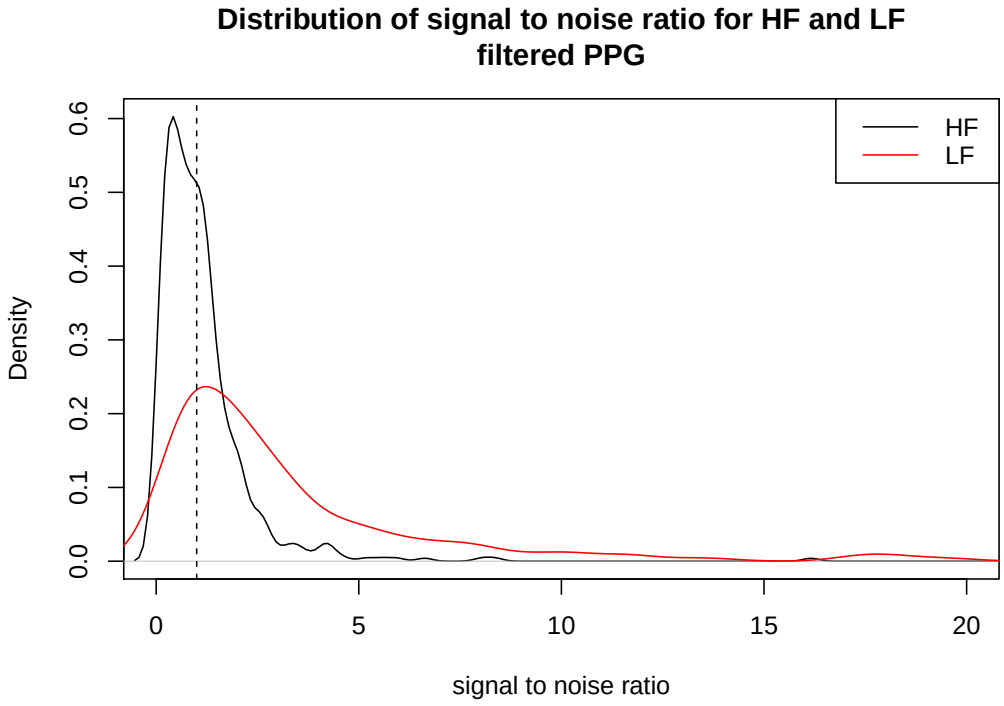


Figure 2.6: Distribution of signal-to-noise ratio in HF and LF bands for all subjects. SNR = 1.0 is shown with the vertical dashed line; values to the left of this line correspond to a proportion of subjects' data that falls below SNR = 1.0.

estimated from P-P was expected to be close to the original HR from R-R. This is verified by the analysis on the collected data.

Both the LF and HF bands have a very low SNR. The LF band frequently exceeds SNR = 1.0 for the case in which the P-P intervals from the PPG have been pre-processed. HF band has a signal-to-noise ratio close to, or below, SNR = 1.0, which corresponds to the fact that the noise is at least as powerful as the signal in this band. The absolute power in the HF band is low with respect to the lower-frequency bands (as shown in figure 2.4). This latter effect may explain why SVI is as unreliable as observed in the results: SVI is the ratio between LF and HF, with HF in the denominator, it is therefore sensitive to HF noise. The pre-processed ("filtered") PPG has a higher signal-to-noise ratio than the unprocessed ("unfiltered") PPG in the HF band. This experiment points at the fact that SVI is not a reliable HRV feature and should be used with caution, ensuring pre-processing steps aiming at identifying outliers and motion artefacts are performed.

Figure 2.6 shows the distribution of signal-to-noise ratios for HF and LF bands across all subjects, where I emphasise that these distributions are normalised histograms. The figure shows that HF values frequently fall below the value of SNR = 1 (i.e., that part of the distribution shown in black that lies to the left of the vertical dashed line). This indicates that HF values are not reliable for the majority of subjects. LF values appear to be more reliable than HF, since a larger proportion of the distribution (shown in red) falls above SNR = 1.

2.4.3 Distributions of HRV Errors for all subjects

In the previous sections I applied the accelerometer based noise filtering and subsequent HRV analysis to the data collected from a single user. In this section I replicate the same procedure to the data collected from all users. For every HRV feature I aggregate the results and I calculate the mean, the standard deviation, the 25th, 50th, and 75th percentiles. I present those results as statistical descriptors of the expected error propagation from PPG to HRV features.

Table 2.2 shows the distribution of accelerometry W_t over all subjects in the four stages of the experimental protocol (*rest*, *stress*, *recovery*, and *free*). As expected, the *rest* and *recovery* stages have similar accelerometry, with more than 75% of the W_t values below the threshold κ value of $W_t = 0.02$). During the *free* stage, values of W_t fall below 0.02 less than 25% of the time. During the *stress* stage, W_t consistently exceeds the threshold κ.

Table 2.2: Distribution of accelerometry W_t, showing the mean and s.d. across all subjects, and the percentiles on the distribution of W_t at 0.1, 0.25, 0.75, and 0.9.

	mean	std dev	0.25	0.5	0.75	0.9
rest	0.016	0.032	0.002	0.002	0.010	0.072
stress	0.406	0.202	0.212	0.471	0.577	0.600
recovery	0.012	0.035	0.002	0.003	0.007	0.011
free	0.052	0.032	0.027	0.057	0.075	0.089

Table 2.3 shows the distribution of errors in the time-domain HRV features, for each phase. Reflecting the distribution of the accelerometry W_t, the *rest* and *recovery* stages have low error most of the time (75%), the *free* stage has low error between 25% and 50% of the time, and the *stress* stage is consistently associated with large error in HRV features.

As noted previously, SVI appears to be generally less reliable than the time-domain HRV features, across all subjects. The large errors present in the estimation of SVI, reported in table 2.3, are due to the low signal-to-noise ratio of the HF band.

2.5 Conclusions

I have shown that PPG data acquired from a consumer-grade wrist-worn wearable device are highly susceptible to motion artefacts. I have analysed the noise profile, and shown that the frequency-domain HRV feature SVI is not reliable, because the signal-to-noise ratio in the HF band (which is in the denominator of the quotient that defines SVI) often falls below a value of SNR = 1.0.

Our results demonstrate that HRV features extracted when the user is still are reliable, even if the user is not in a cardiovascular stationary state (such as in the third stage of the experimental protocol, "recovery"), which therefore does not agree with the findings reported in literature [225].

I have shown that it is possible to use the accelerometry data of the wearable device to calculate W_t, and

Table 2.3: Distribution of errors of [a] AverageNN, [b] SDNN, [c] RMSSD, [d] pNN50, and [e] SVI, showing the mean and s.d. across all subjects, and the percentiles on the distribution of the error at 0.1, 0.25, 0.75, and 0.9.

AverageNN	mean	std dev	0.25	0.5	0.75	0.9
rest	0.004	0.010	0.001	0.001	0.002	0.008
stress	0.127	0.052	0.088	0.124	0.155	0.194
recovery	0.007	0.020	0.001	0.001	0.003	0.011
free	0.043	0.032	0.009	0.042	0.072	0.082
SDNN	mean	std dev	0.25	0.5	0.75	0.9
rest	0.012	0.011	0.004	0.009	0.016	0.022
stress	0.132	0.036	0.111	0.133	0.166	0.179
recovery	0.006	0.014	0.001	0.003	0.006	0.015
free	0.030	0.022	0.009	0.028	0.049	0.061
RMSSD	mean	std dev	0.25	0.5	0.75	0.9
rest	0.025	0.014	0.016	0.024	0.031	0.038
stress	0.078	0.022	0.064	0.079	0.089	0.107
recovery	0.008	0.011	0.003	0.004	0.008	0.019
free	0.033	0.020	0.017	0.034	0.049	0.056
pNN50	mean	std dev	0.25	0.5	0.75	0.9
rest	0.238	0.107	0.167	0.235	0.293	0.377
stress	0.511	0.086	0.438	0.525	0.565	0.634
recovery	0.021	0.049	0.000	0.000	0.030	0.060
free	0.174	0.116	0.076	0.165	0.260	0.332
SVI	mean	std dev	0.25	0.5	0.75	0.9
rest	1.536	1.298	0.745	1.022	2.099	2.377
stress	0.611	0.535	0.147	0.451	0.772	1.515
recovery	0.659	0.452	0.332	0.537	0.932	1.390
free	1.308	0.994	0.512	1.112	2.057	2.948

estimate the quality of the corresponding PPG signal. Therefore, unreliable data can be automatically discarded using a threshold κ on W_t. Such an operation could be used to selectively turn off the PPG sensor, thereby saving battery life on the wearable device, and also avoiding HRV analysis using unreliable data. I derive a value of κ that retains data with error an order of magnitude lower than the signal for most HRV features, discarding data with higher error. The value of κ should be refined with a larger dataset.

I have shown that the conventional use of long-duration windows of data is not needed to perform accurate estimation of time-domain HRV features. I emphasise that the results described in this section represent a preliminary study, and future work should include validation of these results with more subjects to verify my findings. The constraint of a limited participant pool, comprising only five subjects, of similar characteristics (age, gender, fitness level), was attributed to the choice of allocating the limited resources available for subsequent investigations of this thesis. Future investigations should strive to expand the participant base, ensuring a more comprehensive and robust assessment of the outcomes.

An additional constraint pertains to the reliance on a commercial-grade chest strap electrocardiogram (ECG) as the gold standard, with the initial assumption of its error-free nature, which, upon closer examination, is susceptible to motion artefacts. In hindsight, employing a Bland-Altman analysis would have been more judicious, acknowledging and addressing potential discrepancies in the reference standard rather than assuming its infallibility.

Chapter 3

Analysis of the Impact of Interpolation Methods of Missing RR-Intervals Caused by Motion Artefacts on HRV Features Estimations

3.1 Introduction

In section 1.1.4 I have discussed how the increased adoption of wearable devices, capable of passively capturing vital signs, creates an opportunity for continuous monitoring of users health status [81]. These wrist-worn wearable devices, however, produce several inconsistent RR-intervals caused not only by physiological phenomena (i.e. ectopic beats, atrial fibrillation, etc.), but mainly by motion and mechanical artefacts induced by external stimuli. The number of abnormal RR-intervals increases from 1%—when heart beats are recorded with gold standard technology (i.e., electrocardiography)—[103] to more than 10%—when they are recorded with wrist-worn wearable devices. As discussed in Chapter 2, the biggest issue about the data recorded by wearable devices is the missing values due to outliers removed for portions of data affected by motion artefacts. This missing data could negatively affect the assessment of HR response and estimation of HRV, that could in turn provide misleading insights concerning the health status of the individual.

However, standard methods for calculating HRV features from the time-series of RR-intervals require accurate beat detection. Hence, handling the missing values became a fundamental aspect to correctly evaluate users' physiological response. These missing values affect the HRV analysis producing misleading results [107]. In previous stud-

ies, the inconsistent RR-interval data were handled by reconstructing the missing values using nearest-neighbour, linear, cubic spline and piece-wise cubic Hermite interpolation methods [107]. However, these methods can also introduce changes in the reconstructed time-series that could corrupt the signal spectrum [39], thus reducing the ability to estimate both time or frequency domains HRV features.

In the study presented in this Chapter, I focus on healthy subjects with normal heart activity and investigate the effects of missing variation of the timing between beats (RR-intervals) caused by motion artefacts on HRV features estimation by randomly introducing missing values within five min time windows of RR-intervals obtained from the *nsr2db* [1] PhysioNet dataset [78] by using Gilbert burst method [75].

I then evaluate several strategies for estimating HRV in the presence of missing values by interpolating periods of missing values, covering the range of techniques often deployed in the literature, via linear, quadratic, cubic, and cubic spline functions. I thereby compare the HRV features obtained by handling missing data in RR-interval time series against HRV features obtained from the same data without missing values.

Finally, we assess the difference between the use of interpolation methods on time (i.e., the timestamp when the heartbeats happen) and on duration (i.e., the duration of the heartbeats), in order to identify the best methodology to handle the missing RR-intervals. The main novel finding of this study is that the interpolation of missing data on time produces more reliable HRV estimations when compared to interpolation on duration. Hence, I can conclude that interpolation on duration modifies the power spectrum of the RR signal, negatively affecting the estimation of the HRV features as the amount of missing values increases. I can conclude that interpolation in time is the optimal method among those considered for handling data with large amounts of missing values, such as data from wearable sensors.

This Chapter is adapted from a paper published in *Sensors* [153] in 2019. The research and results discussed herein were collaboratively conducted with Rossi, A. from the University of Pisa, and Cairo, M. from BioBeats. Their contributions primarily focused on data analysis.

3.1.1 Contributions

To the best of my knowledge, this work is one of the first studies investigating the effect of high percentage of missing values (i.e., 30%, 50% and 70%) on HRV analysis. In previous studies, the inconsistency of RR-intervals was due to a small number of ectopic beats, while wrist-worn wearable devices introduce motion and mechanical artefacts that produce a huge quantity of abnormal heart beats.

Moreover, to the best of my knowledge, this is the first study to analyse the effect on HRV features of interpolation on time versus interpolation on duration. I show the difference among interpolation methods (i.e., no-interpolation, nearest neighbour, linear, quadratic and cubic spline) on both time and duration time-series in order to detect which interpolation method yields lower error in HRV features estimations.

[1] `https://physionet.org/content/nsr2db/1.0.0/`

I conclude by showing that interpolation on time is the best choice for pre-processing RR time-series with missing values, contradicting the approach traditionally followed, based on durations time-series.

3.2 Materials and Methods

3.2.1 Dataset

In the study presented in this section, I used *nsr2db* (Normal Sinus Rhythm RR Interval Database) PhysioNet dataset [78]. This dataset contains beat annotations of 54 normal sinus rhythm subjects (30 men: 28–76 years; 24 women: 58–73 years) extracted from 23 h long electrocardiogram (ECG) recordings, digitised at 128 samples per second, and beat annotations obtained by automated analysis with manual review and correction. The selection of the *nsr2db* ECG dataset was driven by our intention to initiate the study with a dependable time series, onto which we introduced noise replicating the characteristics induced by motion artefacts. While a PPG dataset was a viable alternative, the unavailability of datasets with sufficient data and a paired trustworthy ground truth led us to opt for the approach of introducing synthetic noise to a reliable time series.

In order to compute HRV features, the 23 h time series of ECG recording of each user were split into 5 min windows. Moreover, to investigate the effect of missing values on HRV analysis, artificial missing RR-intervals (i.e., 30%, 50% and 70% of missing values) were inserted into the 5 min windows.

The missing values were created in accordance with a burst Gilbert model [75] that simulates burst-error with a two-state Markov chain (i.e., good as 0 and bad as 1) [93]. I simplify the burst model to only have one parameter p so that:

- the model transitions from state 0 (error free) to state 1 (error) with probability p
- the model transitions from state 1 (error) to state 0 (error free) with probability $1 - p$
- when in state 0 (error free) the model remains in state state 0 probability 1-p
- when in state 1 (error) the model remains in state state 1 probability p
- when in state 0 the data is always error free
- when in state 1 the data is always missing

In this simplified version of the burst Gilbert model, the model's behaviour is completely determined by a singular parameter, denoted as p. This parameter not only dictates the model's dynamics but also equates to the expected proportion of missing data introduced into the time series.

The missing values were introduced in the time series when the state of the two-state Markov chain is equal to 1. Examples of 30%, 50% and 70% of missing values created by Gilbert model are provided in Figure 3.2.

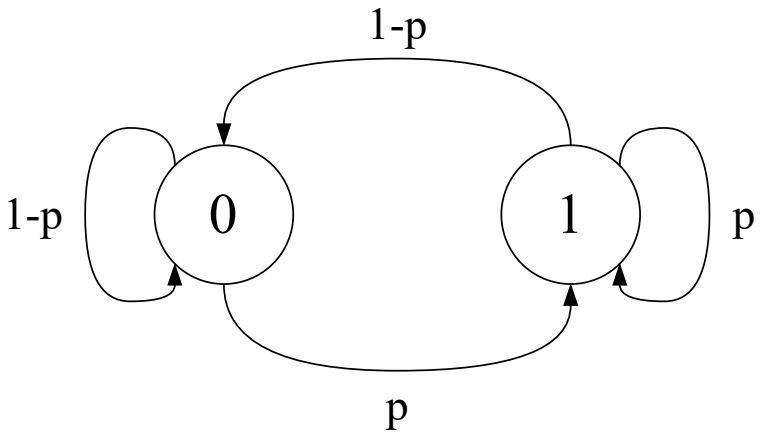


Figure 3.1: Gilbert model simulates burst-error with a two-state Markov chain (i.e., 0 and 1).

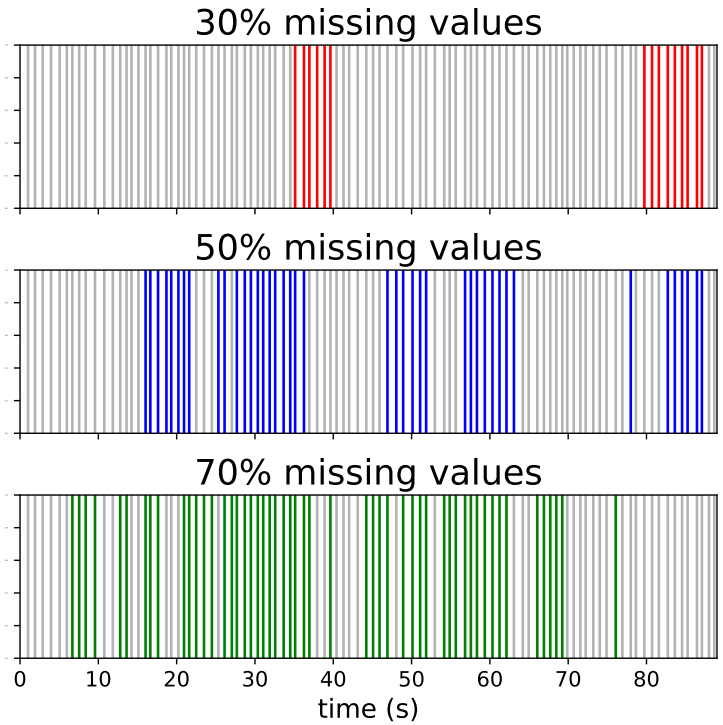


Figure 3.2: Examples of 30%, 50% and 70% of missing values created by Gilbert model. The coloured lines refer to missing beats, the grey lines refer to heart-beats kept in the time-series.

The Gilbert model was selected for simulating missing data due to its simplicity and its ability to generate errors distributed in bursts, mirroring real-life scenarios. This choice aligns with the anticipated occurrence of artefacts, such as those induced by motion artefacts, during alternating periods characterised by user relative stillness and moments of user movement.

3.2.2 Missing Values Interpolation

Considering the scenario where data is missing at timestamp i, such that $A < i < B$, and with the values at time A and B known (X_A and X_B respectively), six distinct interpolation methods were applied to estimate the missing values:

- No interpolation: this approach does not create interpolated values of missing RR-intervals. Differently to the Deletion method that remove missing values merging non-consecutive beats that induce in missing interpretation of HRV features, the no-interpolation method maintains the missing values into the RR-intervals time series.

- Nearest neighbour: the nearest neighbour or proximate interpolation is the easiest interpolation method [211]. This interpolation assigns the value of the closest known (existing) neighbour to the missing- value as shown in Equation (3.1).

$$X_i = \begin{cases} x_B & \text{if } i < \frac{a+b}{2} \\ x_A & \text{if } i \geq \frac{a+b}{2} \end{cases} \tag{3.1}$$

  where a and b are the indexes of x_A and x_B. Interpolated data by this method are discontinuous and it often yields the worst results [121]

- Linear: this method fits a straight line passing through points x_A and x_B [76]. Interpolated data by the linear model are bound between x_A and x_B as showed in Equation (3.2).

$$X_i = \frac{x_A - x_B}{a - b}(i - b) + x_B. \tag{3.2}$$

  Gaunck et al. [76] demonstrated that this method is efficient, and most of the time it is better than non-linear interpolations for predicting missing values in environmental phenomena with constant rates. In addition, they also found that in average this interpolation model underestimated the real values but it strongly depends on the distribution of the data.

- Quadratic: differently from the linear interpolation model, the quadratic function needs three points of interest

to interpolate missing values in a time series as showed in Equation (3.3).

$$X_i = x_B \frac{(i-b)(x_C - x_A)}{2(b-a)} + \frac{(i-b)^2(x_A - 2x_B + x_C)}{2(b-a)^2}. \tag{3.3}$$

Compared to the linear model, quadratic interpolation is found to be in general more accurate [121].

- Cubic spline: fitting data points using polynomials of degree higher than one leads to problems of oscillation outside the fitted points, known as Runge's phenomenon [48]. This problem can be avoided by using a spline, a function defined piece-wise by polynomials, using data points as control points instead of forcing the fitted function to pass through the data points. Cubic splines are composed of piece-wise third-order polynomials. By using third degree polynomials is possible to ensure that the resulting curve is smooth [48], avoiding the problem of the straight polynomial interpolation that tends to induce distortions on the edges of the polynomials, given by the fact that, in general, the first and second derivative of the function defined by piece-wise polynomials will not be continuous at the edges of polynomials. With cubic splines, it is possible to force the first and second derivatives of consecutive polynomials to be equal, ensuring smoothness of the resulting curve.

I applied each of the interpolation methods listed above to heartbeats expressed as a sequence of durations and as a sequence timestamps, then analysed the error in HRV features estimations, in order to identify the best approach.

The on-duration approach is that which is mostly commonly used in the literature to handle missing values. The data used as input to the interpolation methods was the sequence of durations of the heartbeats (the RR-intervals), obtained by subtracting the timestamp of each heartbeat from the timestamp of the subsequent heartbeat in the sequence of heartbeats.

Differently, I propose the the on-time approach whereby interpolation methods are applied to the sequence of timestamps of the heartbeats, postponing the differentiation pre-processing step that transforms timestamps into durations to after the interpolation step.

As shown in Figure 3.3 the difference between the on-time and on-duration approaches is the order of the processing steps: in the on-duration approach the timestamps are converted to durations as the first processing step; in the on-time approach this step is performed after interpolation is performed.

To better illustrate the differences between interpolation on time and duration, I simulated 100 heartbeats and then I randomly generated 10% of missing RR-intervals in this artificial time-series by using Gilbert burst approach. The length of the RR-intervals time-series changes when I interpolate the missing values on duration, while it remains the same when I interpolate on time (Table 3.1). This result suggests that the interpolation on duration moves beats away from their original position in time, introducing changes to the spectrum, while interpolation on time preserves the position on time of retained heartbeats. In particular, Table 3.1 shows that the low RR-intervals error (i.e., average difference between heartbeats duration) is obtained with linear interpolation

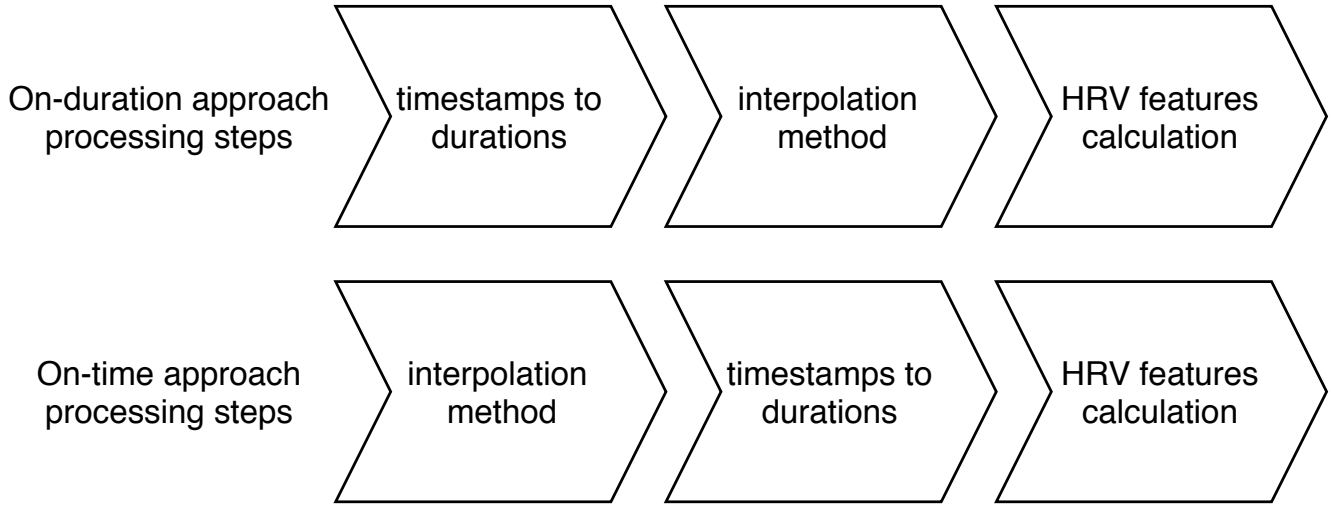


Figure 3.3: Processing steps in the on-time and in the on-duration approaches.

Table 3.1: Difference between duration and time interpolation by using different approach (i.e., no-missing values, nearest neighbour, linear, quadratic, and cubic spline). The best values for RMSE and RE are in bold.

	Window Time (s)		**RMSE (s)**		**RE (%)**	
Interpolation	**Time**	**Duration**	**Time**	**Duration**	**Time**	**Duration**
No-missing values	90.11		—		—	
Nearest	—	91.95	—	0.096	—	5.11
Linear	90.11	91.83	**0.075**	0.090	**3.70**	4.86
Quadratic	90.11	92.13	0.084	0.107	4.35	5.83
Cubic spline	90.11	92.24	0.085	0.109	3.46	6.63

on time. The nearest interpolation on time was not performed because interpolating with this approach is useless, as the interpolated values introduced into the time-series would have the same time as the closest beat, creating physiologically impossible data.

Figure 3.4 provides more detailed analyses of the difference between linear interpolation on both duration and time. The cumulative error when the missing values are interpolated on duration increases incrementally (i.e., the higher is the number of missing values, the higher is the cumulative error) depending on the interpolation type used (e.g., linear, quadratic and cubic spline). Differently, time interpolation did not introduce change in time series length due to the fact that this approach estimates intermediate values between the time when two observed beats happen in accordance with interpolation type.

3.2.3 Feature Engineering

To obtain HRV features, I analysed real (i.e., without missing values values), non-interpolated, and interpolated (i.e., with different percentage of artificial missing values values) 5 min ECG time series.

The variables that I include in my analysis, in both time and frequency domain, are defined as:

- Time domain (defined in section 1.1.2): HR mean; RMSSD; SDNN; pNN50;

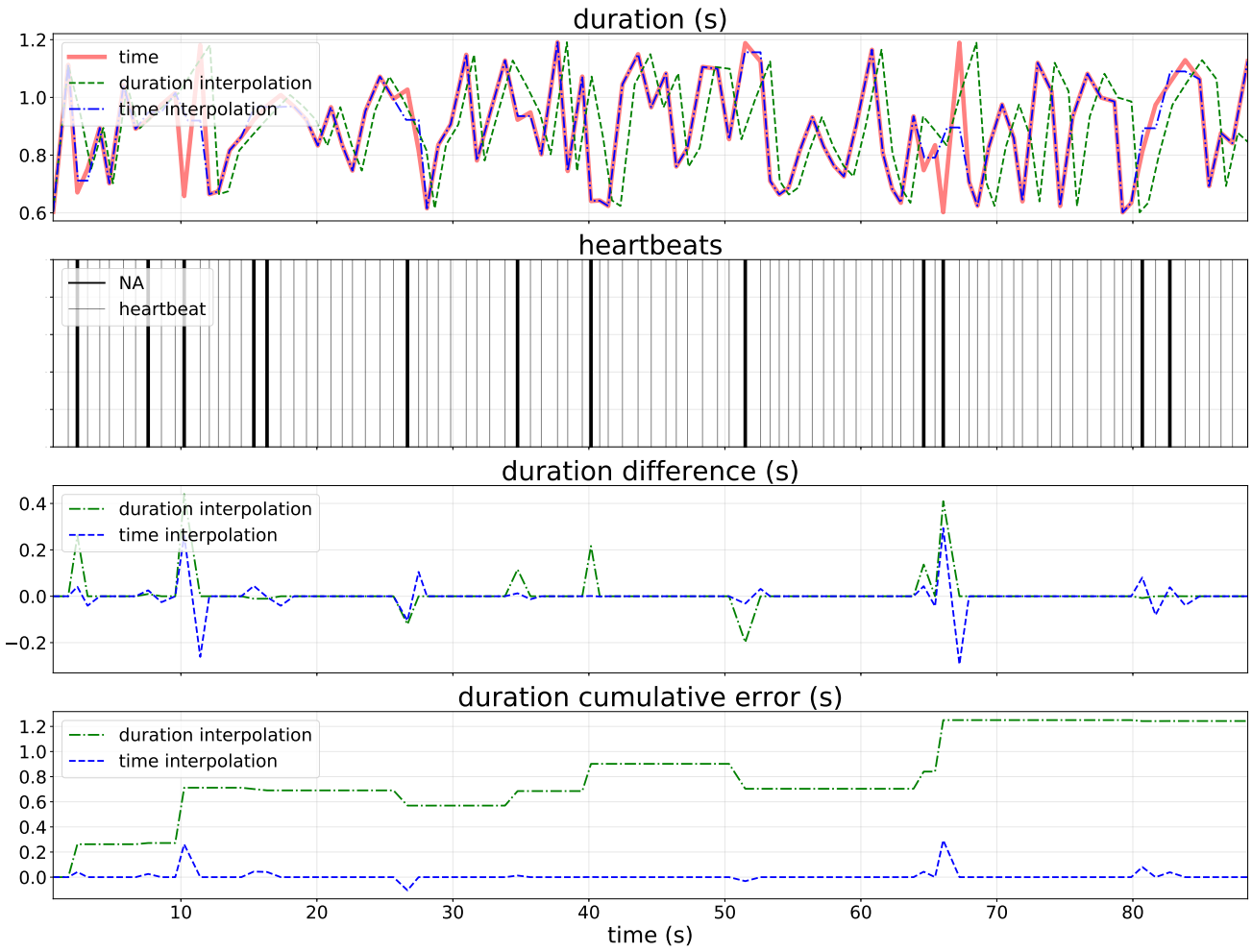


Figure 3.4: Difference between linear interpolation on time and duration. Red solid line refers to real variation of the timing between beats (RR-intervals) time series, green dashed line refers to the on-duration approach, and green dot dash line refers to the on-time approach.

Table 3.2: Descriptive statistic of HRV features. Mean and 95% coefficient intervals (CI) are provided for all the feature.

HRV Features	**Mean**	**95% CI**
IBI (s)	0.78	[0.54, 1.11]
pNN50 (n)	8%	[4%, 16%]
RMSSD (s)	0.039	[0.017, 0.36]
SD1 (s)	0.027	[0.012, 0.26]
SD2 (s)	0.077	[0.040, 0.25]
SDNN (s)	0.059	[0.017, 0.25]
VLF (s^2)	0.87	[0.22, 4.15]
LF (s^2)	0.477	[0.12, 5.57]
HF (s^2)	0.28	[0.050, 3.024]
total power (s^2)	1.91	[0.53, 21.44]
LF/HF (s^2)	2.9	[1.2, 10.2]

- Frequency domain (defined in section 1.1.2): PSD in the VLF, LF, and HF bands; LF/HF ratio (ratio between LF and HF expressed as ms^2); and total power (Power in all the frequency ranges, i.e., $\leq$0.4);
- Non-linear features (defined in section 1.1.2): Poincaré plot (SD1 and SD2)

3.2.4 Success Metrics

I assessed the difference between HRV variables computed on real time series and the ones with missing values by RMSE. Additionally, the relative errors (REs, see Equation (3.4)) were used to assess the effects of the missing data on the HRV features compared with the parameters calculated from the RR-intervals time-series without missing data.

$$RE = \frac{|x_{real} - x_k|}{x_{real}} * 100, \tag{3.4}$$

where x_{real} refers to the HRV features computed from RR-intervals time-series without missing values, while x_k refers to the values obtained from interpolated time-series.

3.3 Results and Discussions

3.3.1 Results Summary

I analysed 15,359 RR-intervals time-series of 5 min in this study. Table 3.2 shows the descriptive statistic of the HRV features extracted from all users in the dataset. In particular, the users have an average heart rate of about 75 $\pm$ 14 beats per minute.

Table 3.3 presents the interpolation methods demonstrating the lowest error across varying amounts of missing values for each HRV feature. The findings reveal that the on-time approach, involving interpolation on the timestamp of heartbeats, consistently yields more reliable estimations of HRV features compared to the on-duration approach, which involves interpolation on the interval duration between two consecutive heartbeats. Notably, in

Table 3.3: Best performing interpolation approach (i.e., with lowest RE) for each HRV feature in each percentage of missing values evaluated. The error in estimating HRV features is reported using RE and RMSE.

Missing Values (%)	HRV	Interpolation Domain	Interpolation Method	RE (%)	RMSE
30	RMSSD (s)	No-interpolation		14.65	0.38
	SDNN (s)	Time	quadratic	9.42	0.34
	PNN50 (n)	No-interpolation		24.37	1.51
	SD1 (s)	No-interpolation		14.68	0.27
	SD2 (s)	Time	quadratic	8.57	0.47
	VLF (s^2)	Time	quadratic	14.50	0.82
	LF (s^2)	Time	quadratic	26.87	2.01
	HF (s^2)	Time	quadratic	32.18	4.48
	LF/HF (s^2)	Time	cubic	41.39	1.73
	total power (s^2)	Time	quadratic	17.16	6.26
50	RMSSD (ms)	No-interpolation		23.13	0.76
	SDNN (s)	Time	quadratic	15.47	0.41
	PNN50 (n)	No-interpolation		39.01	2.35
	SD1 (s)	No-interpolation		23.18	0.54
	SD2 (s)	Time	quadratic	13.49	0.49
	VLF (s^2)	Time	quadratic	23.72	0.40
	LF (s^2)	Time	quadratic	42.42	1.12
	HF (s^2)	Time	quadratic	52.56	2.48
	LF/HF (s^2)	Time	cubic	58.07	2.26
	total power (s^2)	Time	quadratic	27.59	3.96
70	RMSSD (s)	No-interpolation		34.37	0.91
	SDNN (s)	Time	quadratic	22.76	0.47
	PNN50 (n)	Time	linear	63.90	3.88
	SD1 (s)	No-interpolation		34.46	0.59
	SD2 (s)	Time	quadratic	19.19	0.51
	VLF (s^2)	Time	quadratic	29.73	0.52
	LF (s^2)	Time	quadratic	56.41	1.45
	HF (s^2)	Time	quadratic	72.98	3.34
	LF/HF (s^2)	Time	cubic	72.07	2.80
	total power (s^2)	Time	quadratic	72.07	5.27

none of the considered cases did the on-duration approach emerge as the method with the lowest error. In this table I provide the results of the best interpolation approach for each HRV feature and for all the percentages of missing values. The RE and RMSE values provided in this table refer to the error induced by missing values when I compare HRV features obtained from the real RR-intervals time-series versus the ones obtained from interpolated time-series. The best interpolation methods provided in Table 3.3 refer to the ones with lower RE. For all of the HRV features, the highest was the percentage of missing RR-intervals, and also the parameters estimation errors. This was due to the fact that the power spectrum of the RR-intervals signal changes with the number of missing values. The choice of the interpolation method also added different types of noise to the signal. As shown in Table 3.1, the interpolation on time, or not interpolation at all, produces more reliable HRV features compared to interpolating on duration. Table 3.3 only reports the best interpolation method, in appendix E I reported the results of a larger selection of interpolation methods, for all levels of missing values, for all HRV features.

The lowest errors on HRV features estimation with missing RR-intervals are obtained using the no-interpolation or the interpolation on time approaches, while the interpolation on duration approach consistently yields the worst results (Table 3.3). Even if low time-series difference were detected in simulated linear interpolation on time (Figure 3.4 and Table 3.1), Table 3.3 suggests that the best interpolation method depends of the HRV features that we want to assess. Moreover, this table also shows that, as suspected, the higher is the percentage of missing values, the higher is also the HRV feature estimation error (i.e., RE and RMSE).

3.3.2 HRV Features

Time Domain

RMSSD and PNN50 do not require any interpolation to obtain reliable estimations for all the percentages of missing values, while SDNN needs quadratic interpolation on time (see Table 3.3). A possible explanation of this result is that RMSSD and PNN50 capture fast changes in heart activity, i.e., high spectrum frequencies, and SDNN captures slow changes, i.e., very low spectrum frequencies, as discussed in Chapter 6. Moreover, interpolation methods, especially interpolation on duration, act as low pass filters, affecting the signal measured by the HRV features (Figure 3.5). No interpolation changed the spectrum, but did not introduce fictuous durations, thus minimising the impact on successive differences of durations, that were the first computation step of both RMSSD and PNN50.

Frequency Domain

Figure 3.5 shows the Lomb–Scargle spectral analysis for different percentages of missing values and for each interpolation method on both time and duration. This figure shows that different interpolation methods introduce different deformations in the resulting power spectra. It is interesting to notice that performing no interpolation results in a flatter spectrum, more similar to a white noise.

In the frequency domain, the interpolation method that produces the lowest error is the quadratic on time (see Table 3.3). This figure shows that, as the amount of missing values increases, not interpolating (i.e. the no-interpolation approach) caused the spectrum to change (dampening low frequencies and enhancing high frequencies), flatten it, making it similar to the spectrum of white noise. Figure 3.5 also shows that cubic spline interpolation on time tends to dampen low frequencies while enhancing high frequencies; that cubic spline interpolation on duration tends to dampen all frequencies; and that quadratic interpolation on duration tends to enhance all frequencies. Finally, Figure 3.5 also shows that linear and quadratic interpolations on time and that nearest neighbour and linear interpolation on duration have minimal impact on both low and high frequencies, with quadratic interpolation on time having the least effect on all frequencies.

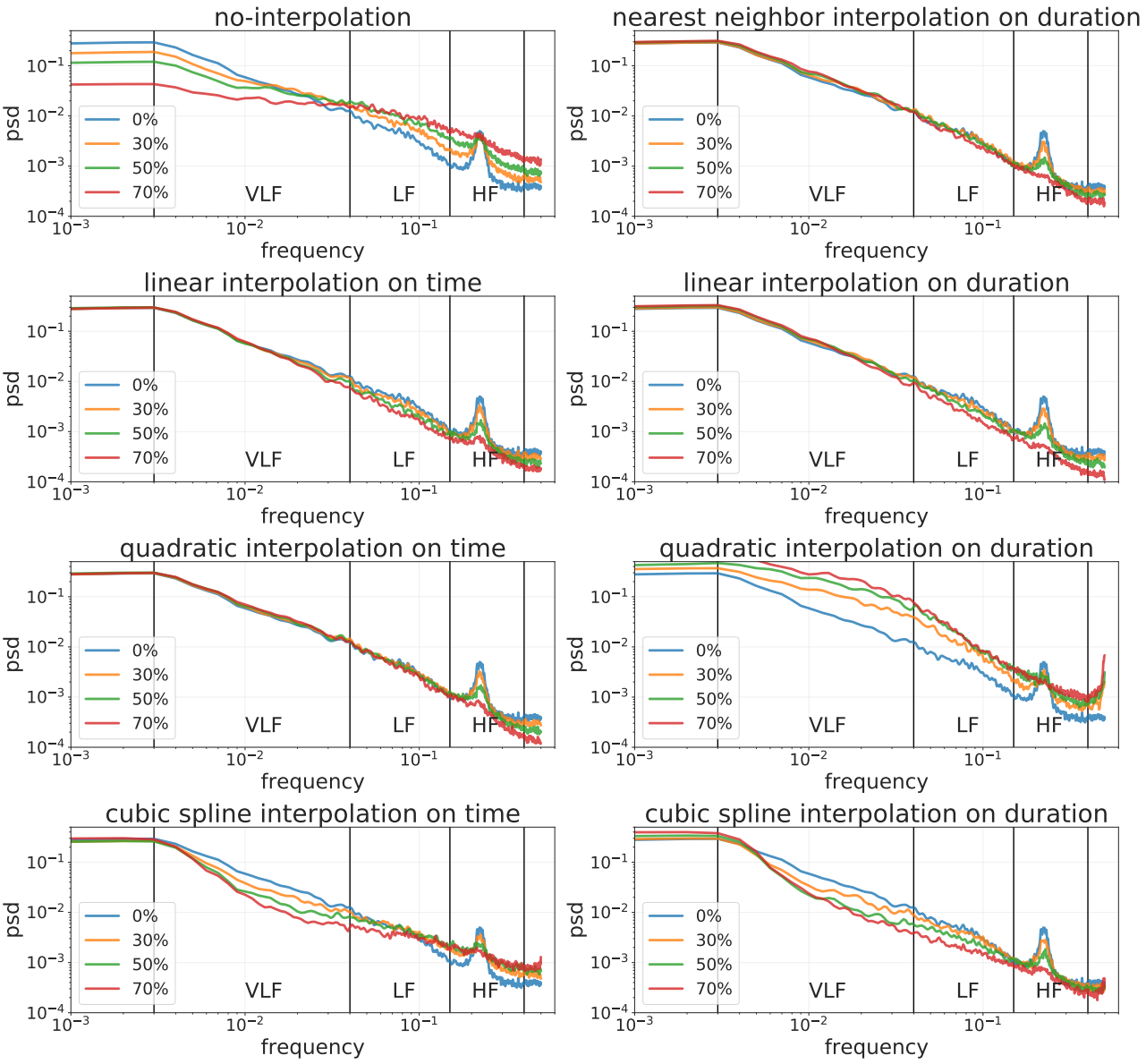


Figure 3.5: Frequency analysis of a user's RR-intervals time-series recorded in 5 min with different percentages of missing values (i.e., 0%, 30%, 50% and 70%) handled with different interpolation methods (i.e., nearest neighbour, linear, quadratic and cubic spline) on both time and duration.

Non-Linear Domain

SD1 does not require any interpolation to handle missing values, while SD2 needs quadratic interpolation on time to obtain reliable result (see Table 3.3). To give an explanation of these results, in Figure 3.6 I provide an example of the relationship between two consecutive RR intervals RR_i and RR_{i+1} (i.e., Poincaré plot) where SD1 and SD2 are extracted. This figure shows Poincaré plots obtained after interpolating missing RR-intervals by using different interpolation methods on both time and duration. When the missing values were interpolated on time, the variability of SD1 reduced as the percentage of missing values increased, while the SD2 remained constant. Differently, the interpolation on duration introduced errors on both SD1 and SD2 increasing their variability as the missing values increased. Finally, in this figure it can be seen that no-interpolation and quadratic interpolation on time introduced less error compared to the other method on SD1 and SD2, respectively.

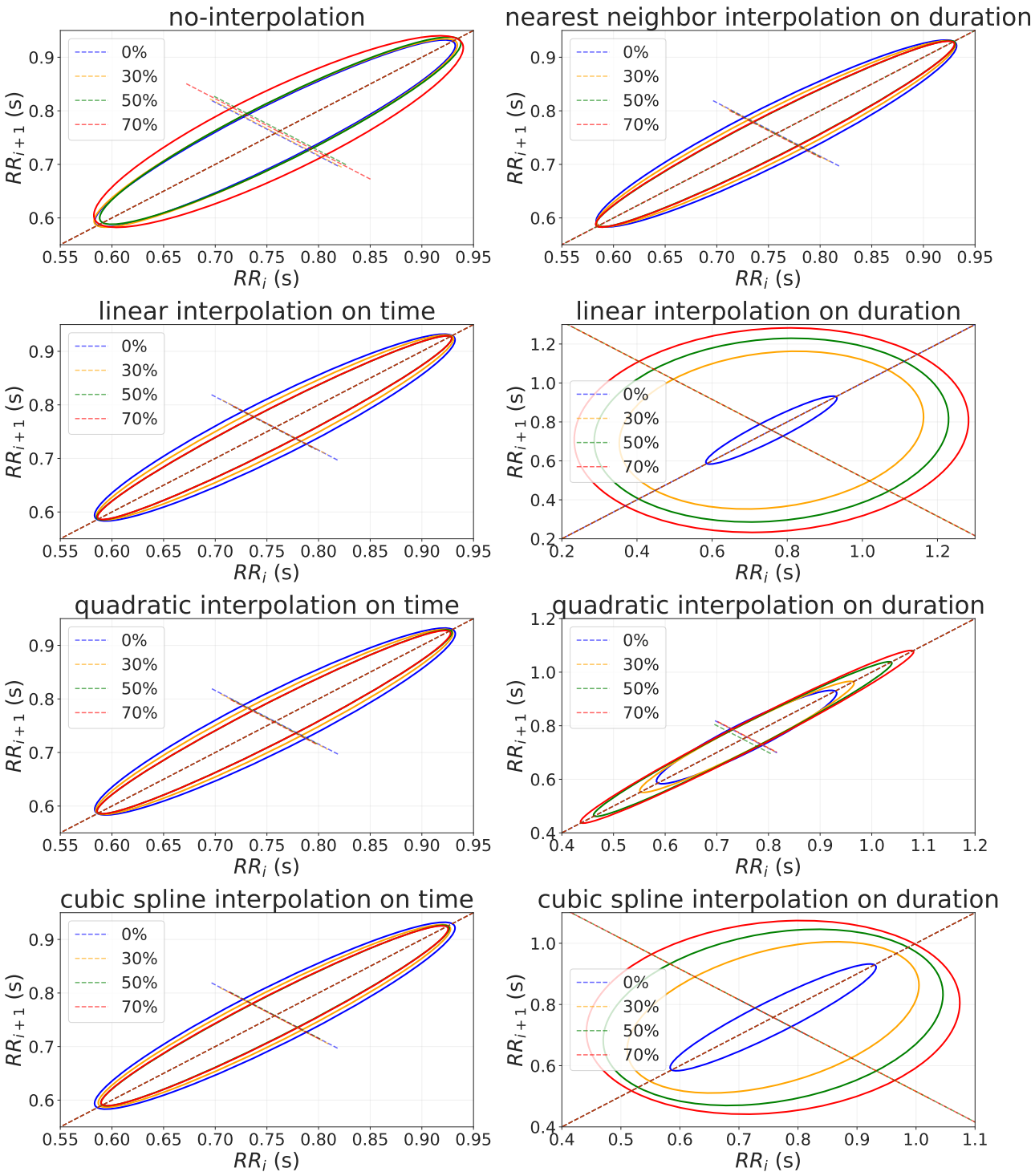


Figure 3.6: Poincaré plot of a user's RR-intervals time-series recorded in 5 min with different percentage of missing values (i.e., 0%, 30%, 50% and 70%) handled with different interpolation methods (i.e., nearest neighbour, linear, quadratic and cubic spline) on both time and duration.

3.4 Conclusions

In this section I quantify the expected error propagation of missing values in RR-intervals time-series to HRV features, as a function of pre-processing interpolation approach, and amount of missing data. The primary findings of this study contradict prevailing practices in the HRV analysis field. Specifically, it is revealed that the interpolation of missing values in RR-intervals time-series on time (i.e., the timestamps of heartbeats) consistently yields more reliable estimations of HRV features compared to interpolation based on duration.

By using this pre-processing approach, the quantification of the expected error on HRV features caused by a huge amount of missing values (e.g., motion artefacts on a wrist-worn wearable device) can support better estimations of users' healh and well-being, by assessing their HRV features. This enables continuous passive monitoring of users' cardiovascular activity in a non-obtrusive way, collecting data during their daily activities that could enable further research on preventative health.

A limitation of this study is that I focused solely on healthy subjects with normal heart activity, restricting the analysis to significant amounts of missing values induced by motion artifacts and overlooking physiological phenomena such as ectopic beats

Future studies will be useful for researcher and companies, which give insight into heart rate variability recorded by wrist worn IoT wearable devices, in order to better understand the potentiality of the data extracted from these devices to make inference about people heath status. Future work is needed to assess the influence of missing values simulated in accordance with motion and mechanical artefacts induced by external stimuli during data acquisition by using wrist worn wearable devices.

Chapter 4

Error estimation of ultra-short heart rate variability parameters: effect of missing data caused by motion artefacts

Analysis of Heart Rate Variability from ultra–short measurement windows is desirable in order to increase the applicability of HRV features to wrist-worn wearable devices equipped with heart rate sensors. This section is focused in particular on two of the most widely used HRV parameters: the standard deviation of inter-beat intervals (SDNN) and the root Mean Squared error of successive inter-beat intervals differences (RMSSD), see section 1.1.2 for their definitions. As I have discussed in Chapter 2, one of the main challenges with extracting these HRV parameters from wrist-worn devices is that their data are affected by the motion artefacts. For this reason, estimating the error caused by this huge quantity of missing values is fundamental to obtain reliable HRV parameters from these devices. To this aim, in this section I simulate missing values induced by motion artefacts (from 0 to 70%) in ultra-short time window (i.e., from 4 minutes to 30 seconds) by random walk Gilbert burst model [75] in 22 young healthy subjects. I find that 30 seconds and 1 minute ultra-short time windows are required to estimate RMSSD and SDNN, respectively. Moreover, due to the fact that ultra-short time windows do not permit to asses very low frequencies and the SDNN is high affected by these frequencies, the bias for estimating SDNN continue to increase as the time window length decreases. Differently, small error is detected in RMSSD up to 30 seconds due to the fact that it is highly affected by high frequencies which are possible to be evaluated even if the time window length decreases.

This Chapter is based on a paper that appeared in *Sensors* [193] in 2020. The work presented here was a joint effort with Rossi, A and Pedreschi, D, who were instrumental in the data analysis phase of the research.

4.1 Introduction

As discussed in section 1.1.4 wearable devices hold great potential for the their ability to passively track patients in their daily life, enabling unobtrusive systems that could detect early signs of deterioration of vital signs.

Several studies estimate the reliability of HRV analysis from IBI time series shorter than 5 minutes for the clinical purposes demonstrating that 10 and 30 seconds recordings are the minimum time required to obtain accurate measures of RMSSD and SDNN, respectively [154, 10, 25, 167]. However, to the best of my knowledge, no studies have been conducted in order to estimate effect of missing values induced by wrist-worn devices equipped with HR sensors on ultra-short HRV parameters.

Hence, the aim of this section is to quantify the error induced by ultra-short IBI time series (i.e., from 4 minutes to 30 seconds) and missing values (from 0 to 70%) in order to detect if it is possible to accurately estimate SDNN and RMSSD by using wrist-worn wearable devices. To this aim, I simulate missing values induced by motion artefacts by Gilbert burst model which create burst-error with a two-state Markov chain in 22 young healthy subject.

4.2 Materials and Methods

4.2.1 Data

In the study presented in this section, I used the Multilevel Monitoring of Activity and Sleep in Healthy people (MMASH) dataset [191, 192, 77], that provides 24 hours of continuous beat-to-beat heart data (*IBI*), triaxial accelerometer data, sleep quality, physical activity and psychological characteristics (i.e., anxiety status, stress events and emotions) for 22 healthy young males (age = 27.29±4.21 yrs; height = 179.91±8.22 cm; weight = 75.05±12.79 kg). Participants' anthropomorphic characteristics (i.e. age, height and weight) were recorded at the start of the data recording. Moreover, participants filled in questionnaires that provide information about participants psychological status, i.e. chronotype, anxiety status and sleep quality. During the 24 hours of the data recording, participants wore two devices that continuously recorded heart response (Polar H7 ECG heart rate monitor - Polar Electro Inc., Bethpage, NY, USA) and the actigraphy data (ActiGraph wGT3X-BT - ActiGraph LLC, Pensacola, FL, USA). Moreover, the perceived mood were recorded at different times of the day and the daily stress were provided before sleeping in order to summarise the individual's stressful events of the day. Finally, twice a day (i.e., before going to bed and when they woke up) the subjects collected saliva samples at home in appropriate vials in order to assess Melatonin and Corisol saliva concentration. More details about experimental setup of MMASH dataset are provided in the data descriptor paper written by Rossi et al. and published on MDPI Data journal in 2020 [191].

Participants wore the Polar H7 chest strap for 24 hours (between 9AM and 9AM of the next day), and were instructed to wear the chest straps during both day (during physical activities too) and night. To the aim of

this study, the IBI time-series of each participants were split in 5 minutes time windows in order to compute gold standard HRV parameters. Moreover, to assess the effect of length of the time window to HRV estimation, ultra-short time windows (from 4 minutes to 30 seconds) were selected in each 5 minutes time window and the HRV parameters were computed. Finally, I introduce missing values to assess their effect on HRV features. More details about the dataset pre-processing and the HRV estimation were provided in Section 4.2.2 and Section 4.2.3, respectively. Finally, the data analysis details are provided in Section 4.2.4.

In accordance with the Helsinki Declaration as revised in 2013, the study was approved by the Ethical Committee of the University of Pisa (#0077455/2018).

4.2.2 Data pre-processing

First, by using the Python *hrv-analysis* library [1] 2.18±1.29% of RR-intervals in the dataset were detected as ectopic beats (i.e., disturbance of the cardiac rhythm frequently related to the electrical conduction system of the heart) or missing values induced by motion artefacts. These missing values were reconstructed via quadratic interpolation applied on the time domain, i.e. the heartbeats timestamp, instead of the duration domain, i.e. the duration of the heartbeats, as suggested discussed in Chapter 3.

In order to compute gold standard SDNN and RMSSD HRV features, the participants' 24 hours RR-intervals time series were split into 5 minutes windows (6,130 five-minutes time windows were extracted). Moreover, to detect the influence of ultra-short HRV features in each 5 minutes window were randomly selected ultra-short window (4 minutes, 3 minutes, 2 minutes, 1 minutes and 30 seconds).

Finally, in order to simulate motion artefacts observable during *IBI* recording by wrist-worn devices equipped with PPG sensors, artificial missing values (i.e., 30%, 50% and 70%) were created for each 5 minutes and ultra-short time windows in accordance with a random walk Gilbert burst model which simulates burst-error with a two-state Markov chain (i.e., good as 0 and bed as 1) [93].

4.2.3 HRV parameters

HRV is usually calculated from 5 minutes *IBI* time series [161]. For the aim of this study I focused on two most popular time-domain HRV measurements: SDNN and RMSSD. These two HRV metrics are computed in each 5 minutes window and in each ultra-short window with and without missing values.

Moreover, to better investigate the effect of missing values in the ultra-short HRV parameters, the power spectrum of the *IBI* time series was computed. The power spectrum $S_{xx}(f)$ of a time series $x(t)$ describes the distribution of power into frequency components (ω) composing that signal. The minimal frequencies that are possible to observe for each ultra-short time series were defined as ratio between 1 and the its length expressed in seconds (*1/sec*), while the maximum is set at 0.4 in accordance with the literature. In order to derive the Power

[1] `https://pypi.org/project/hrv-analysis`

Spectrum Density of the time series $x(t)$, the Lomb–Scargle Periodogram was used instead of Fourier transformation because the inter-beats intervals are not uniformly distributed [130, 199]. VLF, LF, HF, and total power (Power in all the frequency ranges, i.e., $\leq$0.4) were obtained by calculating the integral of the power in the relevant frequency range in the spectrum, as showed in Equation 4.1

$$PSD_{freq} = \int_{min_{freq}}^{max_{freq}} S_{xx}(\omega)d\omega \tag{4.1}$$

4.2.4 Data analysis

The relationship between "gold standard" HRV parameters computed on 5 minutes window and ultra-short HRV parameters with and without missing values are assessed by Pearson's correlation coefficient. Values higher than 0.8 reflects a strong correlation. The error was assessed using the Root Mean Squared Error. The magnitude of the difference was assessed by Cohen's d effect size (ES). It is computed as the ratio between the mean difference between HRV gold standard parameters and ultra-short ones, and the pooled standard deviation as showed in Equation 4.2. ES $<$ 0.2, (0.2-0.5], (0.5-0.8], and $>$ 0.8 are considered trivial, small, and medium and large effect size. For example, if the effect size is equal to 0.04 the two groups' means differ by 0.04 standard deviations and this difference should be considered trivial.

$$ES = \frac{\overline{HRV_{5minutes}} - \overline{HRV_{ultra-short}}}{(\sigma_{HRV_{5minutes}} + \sigma_{HRV_{ultra-short}})/2} \tag{4.2}$$

Spearman's rank correlation coefficient was computed in order to assess the relationship between HRV features.

All the data pre-processing, feature extraction and data analyses are conducted by using Python 3 programming language.

4.3 Results

Descriptive statistics of all the participants' psycho-physiological features provided in MMASH dataset are provided in the data descriptor paper [191].

Table 4.1 and table 4.2 report the descriptive statistics and statistical results of SDNN and RMSSD for all of the windows and missing values percentage, respectively. Small effect size and moderate correlation (i.e., effect size $<$ 0.38, r $>$ 0.69; see Table 4.1) are detected for ultra-short SDNN values with time window longer than 2 minutes showing a bias lower than 10.70 ms (RMSE $<$ 26.84 ms) without a statistically significant systematic error. Differently, moderate effect size (i.e., effect size $>$ 0.45) is detected for time windows shorten then 1 minute. Differently, RMSSD shows a trivial effect size and moderate/strong correlation (i.e., effect size $<$ 0.20, r $>$ 0.75; see Table 4.2) for almost all of the time windows showing a bias lower than 4.83 ms (RMSE $<$ 21.47 ms) with a small systematic error (not statistically significant). In particular, only 70% of the missing values in 1 minutes and

Table 4.1: Descriptive statistics and statistical analysis of SDNN for all of the time window and missing values. The statistical analysis is focused to compare 5-minute HRV parameters to ultra-short ones. Double and single asterisks refer to the trivial and small effect size of ultra-short HRV features compared to 5-minute ones, respectively.

time window	Missing values	SDNN	r	rmse	bias	systematic error	ES
5min	—	72.70±30.23	—	—	—	—	—
4min	**0%**	69.32±35.83	0.79	23.17	-6.92±22.11	0.08	0.20*
	30%	69.02±36.16	0.78	23.87	-7.20±22.75	0.08	0.21*
	50%	68.62±36.26	0.77	24.71	-7.48±23.55	0.09	0.21*
	70%	67.64±36.66	0.74	26.02	-8.27±24.67	0.10	0.23*
3min	**0%**	65.96±31.99	0.76	22.88	-7.64±21.57	0.04	0.26*
	30%	65.72±32.18	0.76	23.19	-7.84±21.82	0.05	0.26*
	50%	65.31±32.33	0.75	23.71	-8.19±22.25	0.05	0.27*
	70%	64.54±32.98	0.73	24.65	-8.76±23.04	0.08	0.29*
2min	**0%**	63.57±31.39	0.71	25.15	-9.51±23.28	0.03	0.34*
	30%	63.22±31.56	0.70	25.66	-9.76±23.73	0.04	0.35*
	50%	62.82±31.85	0.70	25.98	-10.60±23.95	0.05	0.37*
	70%	61.94±32.35	0.69	26.84	-10.70±24.62	0.07	0.38*
1min	**0%**	58.95±30.35	0.63	29.03	-13.50±25.70	0.01	0.45*
	30%	58.54±30.53	0.63	29.46	-13.84±26.00	0.02	0.45*
	50%	58.04±30.78	0.62	29.98	-14.18±26.42	0.02	0.47*
	70%	56.61±31.31	0.60	30.98	-15.28±26.94	0.04	0.52
30s	**0%**	51.41±28.62	0.53	34.18	-19.97±27.74	-0.01	0.67
	30%	50.77±28.80	0.52	34.73	-20.49±28.04	-0.01	0.70
	50%	49.95±29.05	0.51	35.42	-21.16±28.41	0.00	0.71
	70%	48.11±29.63	0.48	36.85	-22.63±29.08	0.02	0.77

30 seconds windows show small effect size and moderate correlation (i.e., effect size > 0.20, $r < 0.62$) with a bias higher than 5.37 ms (RMSE > 24.01 ms), while other time series length with different percentage of missing values shows trivial ES.

Figure 4.1 compares the PSD of the 5-minute time series and the ultra-short ones with missing values. The shorter is the time series, the lower are the Very Low Frequencies (VLF) that are possible to evaluate. In particular, frequencies lower than 0.004 (1/240s), 0.006 (1/180s), 0.008 (1/120s), 0.017 (1/60s) and 0.03 (1/30s) Hz are not possible to evaluate for 4 minutes, 3 minutes, 2 minutes, 1 minute and 30 seconds, respectively. However, even if the time series decrease, it is possible to assess both LF and HF. Due to the fact that RMSSD reflects the HF (a strong correlation was detected between RMSSD and HF in 5-minute time series - Table 4.3), it is possible to accurately estimate RMSSD even if the length of the time series decrease due to the fact that all HF could be assessed. Differently, missing very low frequencies negatively affect SDNN estimation as the length of the time series decreases. In particular, the moderate correlation detected between SDNN and Total power (Table 4.3) demonstrate that short time series do not permits to assess all the frequencies in the power spectrum negatively affecting the SDNN estimation.

In line with the results found in Table 1.3 and Table 1.2 for SDNN and RMSSD, respectively, Table 4.4 shows trivial and small effect size for LF and HF for all of the ultra-short time series with almost all of the missing values (except for 30s time series with 70% of missing value). Moreover, trivial and small effect size for time series longer

Table 4.2: Descriptive statistics (mean±standard deviation) and statistical analysis of RMSSD for all of the time window and missing values. The statistical analysis is focused to compare 5-minute HRV parameters to ultra-short ones. Double and single asterisks refer to the trivial and small effect size of ultra-short HRV features compared to 5-minute ones, respectively.

time window	Missing values	RMSSD	r	rmse	bias	systematic error	ES
5min	—	42.49±20.75	—	—	—	—	—
4min	**0%**	42.34±22.12	0.93	7.87	-0.56±7.85	0.16	0.03**
	30%	42.02±22.96	0.90	9.97	-0.87±9.93	0.21	0.05**
	50%	41.35±24.06	0.85	12.78	-1.50±12.69	0.26	0.07**
	70%	39.04±26.09	0.72	18.34	-3.69±17.96	0.32	0.16**
3min	**0%**	42.15±22.36	0.92	8.82	-0.57±8.81	0.19	0.04**
	30%	41.73±23.21	0.88	10.97	-0.96±10.93	0.24	0.05**
	50%	41.08±24.45	0.83	13.86	-1.56±13.76	0.29	0.08**
	70%	39.00±26.89	0.71	19.21	-3.48±18.89	0.36	0.16**
2min	**0%**	41.83±22.70	0.90	9.85	-0.82±9.82	0.19	0.04**
	30%	41.29±23.60	0.86	12.05	-1.31±11.98	0.25	0.07**
	50%	40.40±24.75	0.80	14.80	-2.17±14.64	0.28	0.11**
	70%	38.38±27.60	0.68	20.63	-4.00±20.24	0.38	0.17**
1min	**0%**	41.26±23.47	0.86	12.02	-1.22±11.96	0.24	0.06**
	30%	40.54±24.57	0.81	14.42	-1.90±14.29	0.29	0.09**
	50%	39.52±26.23	0.75	17.72	-2.84±17.50	0.35	0.13**
	70%	36.73±29.70	0.62	24.01	-5.37±23.40	0.44	0.22*
30s	**0%**	38.73±24.13	0.79	15.13	-2.34±14.95	0.30	0.11**
	30%	37.66±25.34	0.73	17.83	-3.37±17.51	0.33	0.15**
	50%	36.07±27.20	0.65	21.41	-4.83±20.86	0.39	0.19**
	70%	32.63±30.47	0.53	27.14	-7.80±26.00	0.47	0.31*

Table 4.3: Spearman's rank correlation coefficient between HRV features in 5-minute time series.

	SDNN	RMSSD
VLF	0.62	0.19
LF	0.51	0.43
HF	0.45	**0.80**
Total Power	**0.73**	0.45

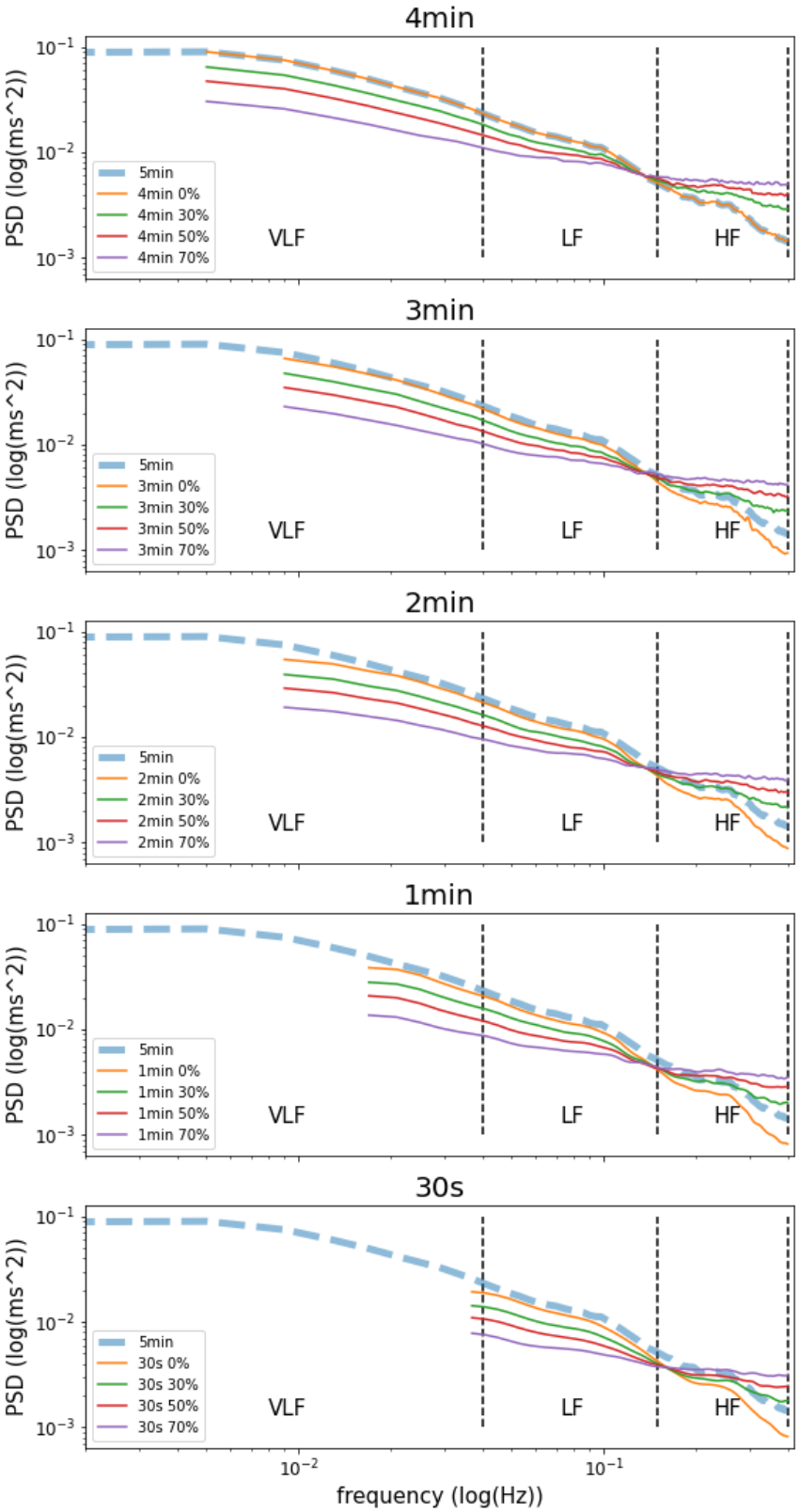


Figure 4.1: Power Spectrum Density plot for time-series with different time series length and missing values percentage. Blue dot line in each plot reflects the power spectrum of the 5-minute time window.

Table 4.4: Descriptive statistics (mean±standard error of the mean) of VLF, LF, HF and Total power of the power spectrum density for all of the time window and missing values. The statistical analysis is focused to compare 5-minute HRV parameters to ultra-short ones. Double and single asterisks refer to the trivial and small effect size of ultra-short HRV features compared to 5-minute ones, respectively.

time window	Missing values	VLF	LF	HF	total power
5min	—	1,891.44±46.10	1,208.11±30.84	689.49±24.84	3,907.50±86.95
4min	***0%***	1,533.88±37.96**	1,208.11±30.84**	689.48±24.84**	3,549.94±80.76**
	30%	1,128.62±27.88*	1,027.24±17.10**	954.92±22.97**	3,208.93±59.54**
	50%	853.10±22.20*	917.61±19.01**	1,123.03±30.88**	2,976.95±66.27**
	70%	567.70±15.31	819.50±27.86**	1,313.26±32.97*	2,778.61±73.01**
3min	***0%***	1,127.11±25.10*	1,110.41±19.26**	547.88±25.18**	2,897.76±58.20**
	30%	827.45±18.66*	937.27±13.08**	785.34±13.43**	2,640.77±40.29*
	50%	625.49±13.81	822.96±12.18*	960.21±17.47**	2,484.46±40.24*
	70%	429.55±9.17	719.21±13.36*	1,137.89±20.49*	2,350.27±41.60*
2min	***0%***	1,035.59±24.53	1,082.92±18.27**	518.80±16.84**	2,743.70±48.45*
	30%	753.04±17.71	904.07±14.64**	743.62±14.54**	2,486.35±41.52*
	50%	572.83±13.52	786.10±12.41*	895.60±15.57**	2,326.40±38.32*
	70%	394.79±9.66	682.95±14.42*	1,067.63±28.46*	2,204.56±50.53*
1min	***0%***	616.19±16.25	1,060.77±17.50**	491.83±10.75**	2,272.02±38.02*
	30%	453.45±11.82	880.46±14.18*	693.12±12.15**	2,110.12±34.48*
	50%	337.15±8.56	741.32±12.09*	831.16±14.37**	1,976.68±32.74*
	70%	226.16±5.78	623.91±11.57*	964.02±17.28*	1,867.88±33.62
30s	***0%***	—	1,043.22±19.46**	497.03±10.04**	1,634.33±27.33
	30%	—	820.17±15.69*	628.70±11.05**	1,521.87±26.04
	50%	—	671.04±13.42*	720.85±13.05**	1,451.23±25.83
	70%	—	539.98±11.27	839.28±16.60*	1,425.51±27.72

than 1 minute and of 4 minutes were detected for total power and VLF, respectively. In particular, only total power of 4 minutes time series with all of missing values percentage and 3 minutes time series with 0% of missing values show a trivial ES. Additionally, Table 4.4 and Figure 4.1 show that VLF and LF are underestimated as the percentage of missing values increase, while HF is overestimated.

4.4 Discussion

Confirming the findings of previous studies [154, 10, 25, 167], we found that RMSSD requires at least 30 seconds to obtain reliable values (trivial effect size; Table 4.2), and that SDNN requires at least 1 minutes of data to provide reliable results. The novel result provided in this study is that in RMSSD is highly affected by missing values compared to SDNN (i.e., 0.12±0.04 and 0.04±0.03, respectively) showing however a small increment in effect size from all of the time series lengths as their percentage increase. Our findings confirm the correlations between HRV features, discussed in section 1.1.6

To truly estimate SDNN we need precise information from the whole power spectrum, i.e. from VLF to HF. I have found a moderate/strong relationship (r=0.73) between total power and SDNN obtained from 5-minute time series. In particular, SDNN is more correlated with VLF than with LF and HF (Table 4.3) and the power of VLF (49.92%) is larger than the power of LF (31.88%) and HF (18.20%) by several orders of magnitude (Table

4.4 resulting in a smaller contribution of HF and LF to SDNN then VLF [22]. Hence, the small effect size and RMSE higher than 22.88 ms between SDNN obtained from 5-minute time series and ultra-short ones are induced by the fact that it is not possible to evaluate all the VLF required to estimate SDNN (Figure 4.1). For example, in 2 minutes IBI time series is possible to evaluate the power of frequencies higher then 0.008 losing information of frequencies between 0.0033 and 0.008. For this reason, SDNN shows a bias of about -9.51±23.28 (RMSE=25.15) with a small effect size (effect size = 0.34). For this reasons, SAN is less reflected in ultra-short SDNN estimation compared to 5-minute time windows one. Moreover, very small change in effect size (0.04±0.03) between 0 to 70% of missing values in all the time window demonstrated that SDNN is accurate even if motion artefacts affect the IBI time series. Hence, SDNN is more affected by the length of the time window than the missing values that could be detected throughout the IBI time series.

Differently, RMSSD is more sensitive to high-frequency (HF) fluctuations [204] compared to LF, VLF and total power. Strong correlation is detected between RMSSD and HF obtained in 5-minute IBI time series (r=0.80, Table 4.3). Hence, even if the length of the IBI time series decrease it still possible to accurately assess the HF. Beak et al. [10] have found that RMSSD and in particular HF calculated using ultra-short IBI time series could be a good surrogate to those calculated using standard 5-min intervals. For this reason and as shown in Table 4.2, RMSSD is found to be accurate for all the IBI time-series without missing values (ES<0.11, RMSE<15.13 ms, r<0.79 and bias<-2.34 ms). For all the time window, effect size shows a small increment with the missing values up to 50% (0.05±0.02), while increase of about 0.15±0.03 for RMSSD estimated with 70% of missing values. However, trivial effect size was observed for almost all the time windows with 70% of missing values. Table 4.4 and Figure 4.1 show trivial effect size in HF up to 50% of missing values for all the time windows, while a small effect size was detected for time windows with 70% of missing values shorten than 1 minute.

These results highlight the fact that approaches to interpolate missing values are not required for RMSSD but they are needed for SDNN estimation as also discussed in Chapter 5. RMSSD captures fast changes in heart activity (HF), while SDNN is a indices of all the power spectrum and in particular of VLF (Table 4.3 and Table 4.4). For this reason, interpolation methods that act as low pass filters affecting the power spectrum density of IBI time series have a negative impact on RMSSD, while a small effect on SDNN [153]. Differently, the no interpolation approach did not introduce white error, thus minimising the impact on IBI successive differences, that is the first computation step of RMSSD. Moreover, due to the fact that SDNN can be estimate from the whole power spectrum the estimation of missing frequencies is required to improve accuracy of SDNN.

4.4.1 Effect of sample size on standard deviation

In table 4.1 it can be seen that as the number of Inter Beat Intervals is reduced (either by shorter measurement window, or by having a larger number of missing values), also the corresponding SDNN becomes lower, even if the underlying measured signal has not changed.

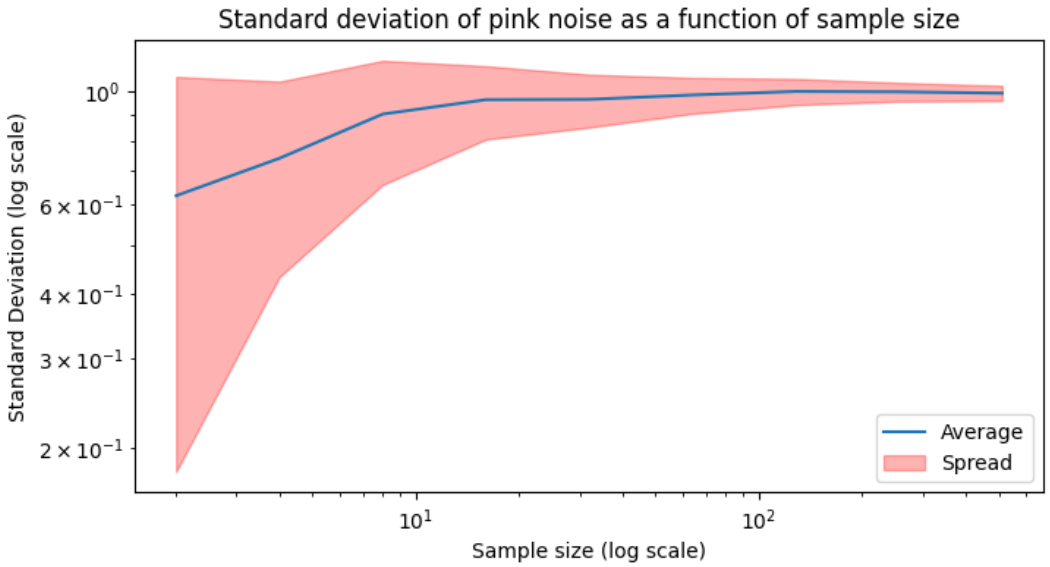


Figure 4.2: Standard deviation (average and spread) of a pink noise signal as a function of the sample size.

In the subsequent discussion, I provide a mathematical explanation for this phenomenon, elucidating the bias introduced in SDNN calculations when considering a subset of the original data. In-depth exploration of this phenomenon is reserved for Chapter 6. Essentially, this effect can be attributed to the standard deviation's direct proportionality to the power of the entire spectrum of the analysed sequence. Longer sequences encompass a wider range of frequencies, especially the lower frequencies, resulting in a larger spectrum. Considering the human heart rate variability spectrum, it displays a profile characterised by increased power in lower frequencies following a $1/f$ shape, often referred to as *pink noise* [190], longer sequences naturally yield higher total spectrum power and, consequently, higher standard deviation.

To further illustrate the concept I have generated 2^{20} random numbers with pink noise shape, and extracted random elements from this sequence, with a sample size between 2^1 and 2^{10}, repeating the extraction 100 times for each sample size. I have then calculated the standard deviation of those samples, and studied how the standard deviation changes as a function of the number of samples extracted from the original signal.

Figure 4.2 shows the average standard deviation and its spread, as the number of samples increases. It can can be seen that the average standard deviation increases as the sample size increases (and its variability decreases, as the estimation becomes more and more precise with more data available).

Furthermore, the calculation of SDNN typically employs the biased formula (dividing by n), which introduces a sample size-dependent bias. In contrast, the unbiased formula, dividing by $n - 1$, adjusts for this variability in the number of samples.

This provides an explanation on why we see a reduction in SDNN as the measurement window shortens and as the number of missing values increases.

4.5 Conclusions

As shown in previous studies, it is possible to accurately estimate RMSSD and SDNN by using 30 seconds and 1 minute time windows, respectively. In particular, in the study presented in this section I highlight the fact that ultra-short IBI time series do not permit to assess all the VLF increasing the bias on SDNN. Moreover, to the best of my knowledge, this is the first study that evaluates the effect of missing values on ultra-short HRV data. I found that the missing values affect SDNN estimations, while have a moderate effect on RMSSD. In particular, the shorter the measurement window, and the higher is the missing values, the higher is the bias in SDNN.

Chapter 5

A computationally efficient algorithm to obtain an accurate and interpretable model of the effect of circadian rhythm on resting heart rate

In this and the following Chapter, my attention shifts towards examining heart dynamics across time durations typically overlooked in conventional HRV analysis, specifically spanning from hours to days. This temporal scope aligns with the Ultra Low Frequency bands within the HRV spectrum. As previously introduced in Sections 1.1.1 and 1.1.2, this segment of the spectrum holds particular significance as it provides valuable insights into the functioning of the Sinoatrial Node.

Due to the inherent challenges associated with measuring Ultra Low Frequency components from ECG, which would necessitate users to endure prolonged discomfort wearing ECG devices, or from wearable PPG devices, which are prone to significant data gaps caused by normal daily activities and motion, the feasibility of capturing HRV data for more than a few minutes is rare. Consequently, the bulk of HRV analysis has predominantly concentrated on the frequency range that can be measured within a brief time frame of less than 5 minutes, primarily influenced by the Autonomic Nervous System. This focus has left the exploration of Sinoatrial Node activity largely uncharted.

In this Chapter, my focus shifts to investigating the influence of circadian rhythm on resting heart rate. I introduce a model that expands upon the widely adopted cosinor model [43], which traditionally relies on a single periodic sinusoidal component. The extended model incorporates multiple components, providing an improved fit for the heart rate data, all while maintaining the user-friendly nature of the original cosinor model. Remarkably,

this expanded model requires only three parameters to describe its characteristics

In Chapter 6, I provide compelling evidence demonstrating that SDNN24, the standard deviation of heart beat durations over 24 hours, a specific HRV feature widely regarded as the *gold standard* for assessing cardiovascular health, can be accurately estimated using HR data collected from commercial-grade fitness trackers. This achievement surpasses the need for the extensive 24-hour ECG monitoring typically associated with SDNN24 assessment. I delve into the underlying rationale, highlighting that frequent HR measurements are capable of yielding reliable estimates of the Ultra Low Frequency band within the HRV spectrum. Notably, this band is predominantly governed by the Sinoatrial Node, the heart's pacemaker, which elucidates why this particular HRV feature serves as a robust predictor of cardiac health.

Circadian rhythm is a strong factor of variability in heart rate, yet few models attempt to accurately capture its effect on heart rate. In this Chapter I present a mathematical derivation of the single-component cosinor model [43] with multiple components that fits user data to a predetermined arbitrary function (the expected shape of the circadian effect on resting heart rate), thus permitting me to predict the users circadian rhythm descriptive parameters (i.e., *MESOR*, *Acrophase* and *Amplitude*) with a high accuracy.

I show that my model improves the accuracy of HR prediction compared to the single component cosinor model (10% lower RMSE), while retaining the readability of the fitted model of the single component cosinor. I also show that the model parameters can be used to detect sleep disruption in qualitative observations. The model is computationally cheap, depending linearly on the size of the data. The computation of the model does not need the full dataset, but only two surrogates, where the data is accumulated. This implies that the model can be implemented in a streaming approach, with important consequences for security and privacy of the data, that never leaves the user devices.

The model I developed goes beyond fitting circadian activity on resting heart rate, and it can be used to fit arbitrary periodic real valued time series, vectorial data, or complex data, with a known expected shape.

This Chapter is derived from a study featured in *Physiological Measurements* [150], published in 2019. The research was a collaborative effort with Bartoloni, L from BioBeats, who contributed to the mathematical development of the model, and Rossi, A. from the University of Pisa, who contributed the data analysis.

5.1 Introduction

Heart Rate is an important vital signs, that can be easily measured using wearable devices equipped with a PPG sensor. However, HR can not be easily interpreted without the proper context. A measured HR value of 70 bpm could be considered high for an athlete while asleep, or low for someone that does not usually train, after a long walk. Several factors are known to affect HR: circadian rhythm will make the HR higher during the day and lower during the night; during physical activity (e.g. a long walk) the muscles will require more blood, causing the heart

to beat faster [217]; fever, consequence of an infection or a flu, will cause the heart to beat faster; the need for mental performances, e.g. a student during an exam, also causes the heart to beat faster [24]; age, sex, and BMI also change the expected heart activity of the user [203]. To be able to interpret HR measures and understand their physiological meaning, is therefore necessary to build a model of the subject, and to estimate the Resting HR (RHR) for that particular subject at that particular time of the day. Once the expected RHR is calculated, the difference between the measures HR and the RHR can provide valuable insight about the user physiological health and well-being.

RHR is widely investigated in literature because it is an independent predictor of cardiovascular and all-cause mortality in men and women [69, 114, 198, 124, 110]. However, these studies are focused on a single point measurements without evaluating the variation of RHR through the day that could provide more information about health status of the population. In 1982 Faria and Drummond [61] detected a physiological response fluctuation during the day of RHR and resting temperature, demonstrating a relationship with physical activity tasks. All the studies previously cited evaluated the cardiovascular disease risk and all-cause mortality from RHR and not from HR because of RHR reflects the basal activation of the Autonomic nervous system that is found to be a predictor of several pathologies. The variation of HR among days could be only due to the different physical activity performed in a specific day [67] and not from a real change in health status, that can be detected analysing the daily fluctuation of RHR.

Physiological measures such as HR and RHR are continuous variables that are subject to the influence of external and internal stimuli, changing over a daily cycle of about 24h with patterns that have been defined as circadian rhythms [109, 61]. The time series obtained from wrist wearable device are often noisy (i.e., variations in the biological system that are not part of the deterministic portion of the signal and that are derived from external errors such as instrument inaccuracy), short (to save battery) and sparse (i.e., unequal time intervals between observations, missing datapoints) [64]. Despite these time series problems and the fact that the real signal is not a simple sinusoidal, HR circadian rhythms analysis is usually performed using the cosinor method [43] with a single component because the resulting parameters can be easily interpreted. In particular, the cosinor curve with single component provides 3 parameters that describe the circadian rhythm:

MESOR (Mid-line Estimating Statistic Of Rhythm) is an estimation of central tendency of the distribution of values across the cycles of the circadian rhythm computed using a cosine function;

Amplitude is the difference between the peak and the mean value of a wave;

Acrophase is the time of the day at which the peak of the circadian rhythm occurs.

It was found that the shape of the HR curve as a function of the hour of the day significantly differs from a simple cosine: during the night the HR becomes slower for approximately 6 hours, while during the day it rises for approximately 18 hours, with 3 peaks corresponding to meals [137, 157]. However, only a few studies in literature

propose a cosinor method with multiple-component to analyze periodic components in non-sinusoidal longitudinal time series [64, 43]. The result of the multiple-component cosinor analysis includes several parameters which are not easy to interpreted in comparison to the three parameters provided by single-component cosinor. Multiple-component model was found to be more suitable to approximate the signal's waveform when it deviates from sinusoidal. However, the single-component cosinor model is still widely used to investigate physiological circadian rhythm because of its simplicity and readability [43].

Some models have been proposed to fit the data more accurately. Leise proposed the use of Wavelets [120, 119]. Fossion proposed the use of Multiscale Adaptive Analysis [68]. Those approaches improve the goodness of fit, but fitting parameters are difficult to interpret, compared to the single component cosinor. An additional limitation of those methods is that they are only concerned with the frequencies present in the data, without requiring the signal to follow a specified shape.

With the wide adoption of fitness devices, concerns about the security and privacy of the data collected by these devices have been raised [63, 4]. Among principles of *Privacy by Design* are *Data Minimisation*, which states that only the data strictly necessary to perform the intended activities should be collected, and *Full Functionality*, which states that privacy should not impair the functionalities of the product [26]. However, the standard approach of fitness companies is to upload all the user data to their servers for processing, and performing all the data analysis on the cloud, requiring access to users raw heart rate data. In practice, *Data Minimisation* is usually sacrificed in favour of *Full Functionality*.

5.1.1 Contributions of this study

In this section I present a mathematical derivation of the single-component cosinor model with multiple components that fits user data on an predefined arbitrary shape (i.e., the model which describes the circadian rhythm of the population), thus enabling the prediction of the user's circadian rhythm effect on resting heart rate.

Unlike the cosinor model with multiple components provided by Cornelissen [43], my model provides the same three easy-to-interpret parameters which can be derived from the cosinor model with single component (i.e., *MESOR*, *Amplitude*, *Acrophase*).

Our model allows to assign weights to every data point used for fitting the expected shape to the user data. This enables advanced analyses with arbitrary error models, like assigning less importance to data points collected while the user was moving (expecting motion artefacts, as explained in Chapter 2). Using weights also it makes possible to implement real-time streaming analysis of the data, performing a new analysis as new data arrives, assigning less importance to older data. In this study I show how to implement such a real-time strategy assigning exponentially decaying weights to old data, obtaining a circadian model that continuously changes with new data. I show how the three fitted parameters (i.e., *MESOR*, *Amplitude*, *Acrophase*) can be used to interpret behavioural and physiological changes.

The model presented in this section is computationally light and, for this reason, it is suitable to be executed on a wearable device. The complexity of the accumulation phase is linear on the number of data points, while the complexity of the resolution phase depends on the number of frequencies used and not on the number of data points.

To the best of my knowledge, this the first study that provides an algorithm that fits the circadian rhythm by using a multiple component cosinor, forcing the data to fit a particular shape, i.e. the population average effect of RHR circadian rhythm, only leaving three free parameters (i.e., the same used by single one). Moreover, my algorithm allows to perform all the data sensitive calculations on the wearable device. As a consequence users raw heart rate data could be kept on the wearable device, while allowing companies to obtain the output of the analysis, implementing both the *Full Functionality* and the *Data Minimisation* Privacy by Design principles.

Predicting a user's RHR is an important step to interpret the measured HR data: the residual between measured HR and estimated RHR can then be correlated with physical activity, and inference of valuable insight on this user's well-being and health becomes possible.

5.1.2 Background

Short and sparse time series were historically analysed by the single-component cosinor model. However, this model is not suitable for unevenly sampled, noisy time series (i.e., the one showing several outlier values), with missing data such as the data recorded by wearable heart rate trackers.

Cornelissen at al. [43] extend the single-component cosinor to a multiple-component model in order to analyse the time series in chronobiology. Instead of solving a system of 3 equations in 3 unknowns to define the parameters of the circadian rhythm curve (i.e., *MESOR*, *Amplitude*, *Acrophase*) there are 2c + 1 parameters to estimate (where c refers to the number of harmonics selected to describe the circadian). This approach better approximates the signal's waveform of the circadian rhythm compared to single-component cosinor model thanks to the higher number of cosinor components used to fit the model on the circadian rhythm curve. For example, a 2-component cosinor model that describes periods of 24 and 12 hours has been found to be well-suited to approximate the nightly drop of blood pressure [82]. However, interpreting and comparing the parameters derived from this model prove challenging, particularly in contrast to results obtained through the single-component cosinor method [64]. For instance, a slight alteration in the phase of a harmonic can lead to significant changes in the resultant function's shape, and parameters such as *Amplitude* may be distributed across several components.

Fitting unevenly sampled data to sinusoidal functions can also be performed by Least-squares spectral analysis techniques, such as the Lomb-Scargle Periodgram [234], or the Vaníček method [224]. The setting of our problem is different, as in I am interested in fitting functions with a particular shape, that is not a simple sinusoidal, but an arbitrary combination of cosines, with fixed amplitudes and locked phases.

5.2 Methods

5.2.1 The mathematical model

Prior from the general population

I model the prior of RHR over time $A(h)$ with the sum of c cosines, as defined in the multiple-component cosinor in Cornelissen at al. [43]. This equation can be rewritten using Euler's complex exponential formula as Equation 5.1. The number c of components to use is a free parameter.

$$\begin{aligned} A(h) &= \beta + \sum_{k=1}^{c} \alpha_k \frac{e^{\mathrm{i}2\pi k(h+\phi_k)} + e^{-\mathrm{i}2\pi k(h+\phi_k)}}{2} \\ &= \beta + \sum_{k=1}^{c} \frac{\alpha_{|k|}}{2} e^{\mathrm{i}2\pi k(h+\phi_k)} + \sum_{k=1}^{c} \frac{\alpha_{|-k|}}{2} e^{\mathrm{i}2\pi(-k)\left(h+\phi_{|-k|}\right)} \\ &= \beta + \sum_{k=1}^{c} \frac{\alpha_{|k|}}{2} e^{\mathrm{i}2\pi k(h+\phi_k)} + \sum_{k=-c}^{-1} \frac{\alpha_{|k|}}{2} e^{\mathrm{i}2\pi k\left(h+\phi_{|k|}\right)} \end{aligned} \tag{5.1}$$

The parameters β, α, and ϕ correspond to the *MESOR*, *Amplitude*, and *Acrophase*, and can be found as explained in Cornelissen at al. [43].

If I define $\psi_0 = 0$, $\psi_{k\neq 0} = \frac{\alpha_{|k|}}{2} e^{i2\pi k\phi_{|k|}}$ I can rewrite Equation 5.1 as Equation 5.2

$$A(h) = \beta + \sum_{k=-c}^{c} \psi_k e^{\mathrm{i}2\pi kh} \tag{5.2}$$

Note that I have defined ψ so that ψ_{-k} is equal to the complex conjugate of ψ_k. Therefore, the number of real parameters has not changed.

Equations 5.1 and 5.2 are equivalent to cosinor analysis [43] and the Vaníček method [224]. From the next section I introduce novel contributions.

Fitting user data with locked parameters

Single component cosinor can only accurately fit data modulated by a simple sinusoidal factor. Circadian effect on RHR is unlikely to be fully captured by such a simple function. Multiple component cosinor can fit data more accurately [43], but the resulting set of fitting parameters, (a *MESOR*; an *Acrophase* and an *Amplitude* for each component) can not be easily interpreted, making it difficult to interpret the parameters of each component. Moreover, fitting a model for each user with all the free parameters will fit shapes that might not be what circadian is supposed to be, and it will be difficult to get an idea of what the phase and amplitude are.

I use the $A(h)$ model provided in the previous section as the prototype (and prior) for users RHR throughout the day (circadian rhythm). However, every user has a different RHR at night, during the day, and might go to sleep and wake up at different times than most of the users (that generated the model). For this reason I define a

refined version of the model where some parameters are let free to adapt to user's data. In particular, Equation 5.3 shows the formula that locks parameters of the c components only allowing 3 degrees of freedom (the same used in single component cosinor analysis: *MESOR*, *Amplitude*, and *Acrophase*). This is equivalent to single component cosinor analysis, using a different function than a cosine.

$$U(\theta, h) = \theta_0 + \theta_1(A(h - \theta_2) - \beta) = \theta_0 + \theta_1 \sum_{-c \leq k \leq c} \psi_k e^{\mathrm{i}2\pi k(h-\theta_2)} \tag{5.3}$$

The free parameters in Equation 5.3 are:

θ_0 the average RHR of the user;

θ_1 the amplitude of the oscillation of the user RHR with respect to the prior. Values of θ_1 close to 1 indicate that the range of the resting heart rate is similar to the prior, less than 1 indicate that the user resting heart rate has a smaller range compared to the prior, etc.;

θ_2 the phase of the user circadian rhythm with respect to the phase of a cosine with period equal to one day. θ_2 ranges between -0.5 and 0.5. Values of θ_2 close to 0 indicate that the user is synchronised with the prior.

The θ parameters allow to express the user RHR, $U(h)$, as a vertical shift (θ_0), horizontal shift (θ_2), and a change in amplitude (θ_1) of the population prior $A(h)$.

Loss function

To find θ that best approximates $U(h)$ for a particular user, I fit $U(h)$ to the RHR data collected for that user, minimising the distance L between the data and $U(h)$ as shown in Equation 5.4, where y_j is the j^{th} measure of RHR of the user, x_j is the time of the measure, and w_j is the weight assigned to the j^{th} measure.

$$L(\theta) = \sum_j w_j \left\| y_j - U(\theta, x_j) \right\|^2 \tag{5.4}$$

Finding the θ that minimises L is not trivial, because I use periodic functions that are not simple cosines, but a particular combination of cosines, defined by the α and ϕ parameters, that are not allowed to change. The literature of cosinor based rhythmometry does not offer any solutions, neither analytic nor numeric. Therefore, I developed an analytic solution to finding the θ that minimises L.

First, I introduce the following notation to further simplify formulae:

$$\zeta = e^{-\mathrm{i}2\pi\theta_2} \tag{5.5}$$

$$F_d = \sum_j w_j y_j e^{\mathrm{i}2\pi d x_j} \tag{5.6}$$

$$S_d = \sum_j w_j e^{\mathrm{i}2\pi d x_j} \tag{5.7}$$

I can now develop Equation 5.4 using the definition of U in Equation 5.3 into the following

$$L(\theta) = \sum_j w_j y_j^2 + \theta_0^2 \sum_j w_j + \theta_1^2 \sum_{k,m} \psi_m \psi_k \zeta^{k+m} S_{k+m} - 2\theta_0 F_0 + \theta_1 \sum_m \psi_m \zeta^m \left(2\theta_0 S_m - 2F_m\right) \tag{5.8}$$

where the indices k and m range between $-c$ and c.

To minimise the loss I set to zero the partial derivatives with respect to θ_0, θ_1, θ_2. I obtain the system in Equation 5.9

$$\begin{cases} \theta_0 S_0 + \theta_1 \sum_k \psi_k \zeta^k S_k - F_0 & = 0 \\ \sum_m \psi_m \zeta^m \left[\theta_0 S_m + \theta_1 \sum_k \psi_k \zeta^k S_{m+k} - F_m\right] & = 0 \\ \sum_m m\psi_m \zeta^m \left[\theta_0 S_m + \theta_1 \sum_k \psi_k \zeta^k S_{m+k} - F_m\right] & = 0 \end{cases} \tag{5.9}$$

Substituting θ_0 and then θ_1 in the equations of the system 5.9 I find a polynomial in ζ:

$$0 = \sum_{m,l,k} m\psi_m \psi_l \psi_k \zeta^{m+l+k} \Psi_{m,l,k} \tag{5.10}$$

where

$$\Psi_{m,l,k} = F_k S_l S_m + F_k S_0 S_{m+l} - F_0 S_k S_{m+l} - F_m S_l S_k - F_m S_0 S_{k+l} + F_0 S_m S_{k+l} \tag{5.11}$$

Since the indices m, l, and k range between $-c$ and c, and $\forall m.\Psi_{m,m,m} = 0$, the polynomial in Equation 5.10 has degree $6c - 2$.

The roots of the polynomial in ζ can be found with spectral methods, i.e. using the eigenvalues of the companion matrix [57] to find an approximation, and then the solutions can be refined using the Newton-Raphson method [74].

From the roots of the polynomial I can determinate the values of θ_2, and subsequently of θ_1 and θ_0, where the total derivative of $L(\theta)$ is zero. I compute $L(\theta)$ using Equation 5.8 on these points to determine the absolute minimum.

Our model is based on the weighted least squares approach. Data do not need to respect the assumption of homoscedasticity, and weights can be designed to focus the analysis on certain areas of the data, e.g. give less weight to data collected in movement.

Numerical considerations

The system in Equation 5.9 depends on data indirectly through the values of F and S. Moreover, also the square loss function can be expressed in term of these values (and the total sum of y_i^2). This means the whole algorithm

does not require access to the full dataset, but only a limited set of values (namely the values of S, F and the sum of squares of y values if we want to output an absolute value for the loss). All these values are moreover an additive function of data sets, that is for two disjoint set of values $A = \{x_i, y_i\}$ and $B = \{x_j, y_j\}$ $S_{k,A\cup B} = S_{k,A} + S_{k,B}$ and $F_{k,A\cup B} = F_{k,A} + F_{k,B}$. This implies that it can be computed separately for each subset of samples, and then added together to obtain the total value: for example it can be accumulated directly in the user's device, with big advantages in computational complexity, data transfer and privacy. In particular the asymptotic analysis of the algorithm complexity is the following (where N is the number of data points, c the number of components):

- The accumulation of data into the surrogate values S and F is $\mathcal{O}(N \cdot c)$
- To solve the polynomial I need to find eigenvalues of a $6c - 2$ sided matrix which requires $\mathcal{O}(c^3)$ operations
- To compute the loss for a given set of fit parameters it requires $\mathcal{O}(c^2)$ operations, which has to be repeated for each candidate solution, for a total of $\mathcal{O}(c^3)$ operations

The total complexity is thus $\mathcal{O}(c^3 + Nc)$, in particular it only depends linearly on the size of data.

5.2.2 Experimental setting

To test the accuracy of my model of circadian effect on RHR against the traditionally used single component consior, I used the data collected by Biobeats, a corporate wellbeing company[1], through their product Biobase, a stress management app that uses a wrist worn wearable device (Biobeam) to passively collect activity data (number of steps performed every 20 seconds), sleep phases by a triaxial accelerometer with a sample rate of 100Hz, and heart beats by photoplethysmography sensors (i.e., Interbeat Interval durations are measured for 2 minutes every 30 minutes, with a sample rate of 50hz; and the average heart rate of 1 minute of heart activity is measured every 10 minutes, with a sample rate of 25hz).

To test my model I used a dataset that includes 47 individuals (F=52%, M=47%), with an age range of 20 - 63 years (M = 29.29, SD = 11.38), who agreed to participate to the experiment and wore a wrist worn wearable device that collected heart rate every 10 minutes, activity data (steps) every 20 seconds, and sleep phases every night, for four weeks. The number of HR measurements per user was on average 3,471 (SD 2,964).

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study was approved by the ethics committee of the University of Surrey, UK, and written informed consent was obtained from all participants (identifier: 1190-PSY-16), as part of a larger study [30].

[1] `https://www.biobeats.com/`

Building a prior by fitting β and ψ on the general population

To build a prior for the resting HR I proceeded as follows:

- for each user I kept only time series with at least 100 heart rate samples. After this pre-processing step dataset had a total of 163,179 HR measures;
- I split the dataset using a leave-one-out strategy, training the prior on a dataset that did not include the user for which I wanted to calculate the θ parameters;
- for each HR measure, I calculated the amount of steps taken by the user in the previous 5 minutes;
- I removed from the dataset the HR points where the number of steps in the previous 5 minutes was larger than 10. This step was carried out to ensure I estimate RHR, instead of HR that is influenced by the physiological request of physical activity. This introduces bursts of missingness, for example, when the user walks for commuting. However, our model is inherently designed to be resilient to unevenly sampled data; hence, there is no requirement for imputing missing data. After this pre-processing step the number of HR datapoints decreased from 163,179 to 132,498;
- on each user I first performed cosinor analysis on a single component to find the *Acrophase* (every user will generally have different sleep habits);
- I then aligned the data, forcing all users to have the same phase as a cosine with period equal to 1 day. This step is performed to align all users to the same sleep habits, and derive the shape of the effect of circadian rhythm on resting heart rate on a population level;
- finally, I used multiple-components cosinor on the phase aligned data from all users, using $c = 4$ cosines, with period of 1 day, half a day, a third of a day, and a quarter of a day.

Figure 5.1 shows one of the priors fitted using the leave-out-out approach. In that case the values for β and ψ, from equation 5.2, are:

- $\beta = 68.75$;
- $\psi_1 = 1.63$;
- $\psi_2 = 0.69e^{-\mathrm{i}0.97\pi}$;
- $\psi_3 = 0.13e^{-\mathrm{i}0.06\pi}$;
- $\psi_4 = 0.19e^{\mathrm{i}0.49\pi}$.

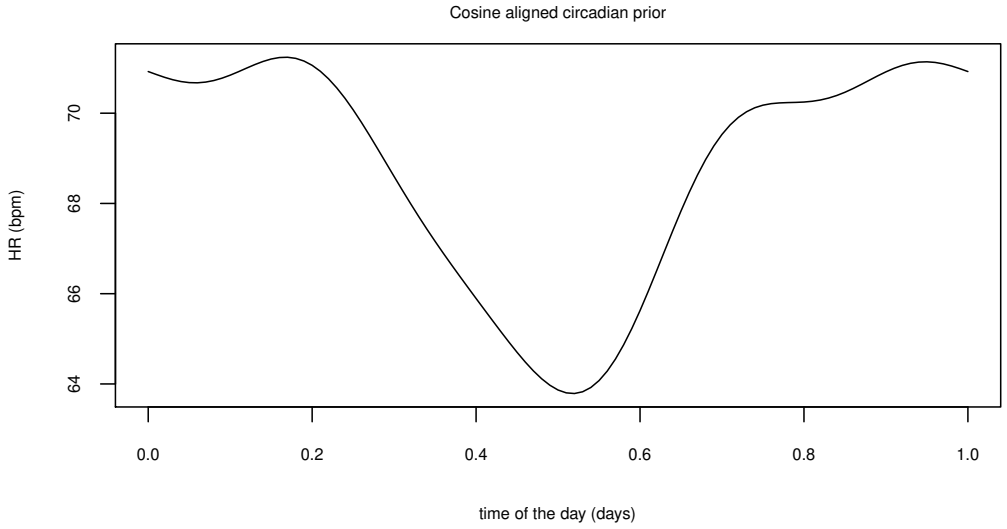


Figure 5.1: Shape of one of the priors fitted using a leave one out approach. The x-axis represents the time of the day, expressed in days: 0.5 indicates midday, 0.0 and 1.0 indicate midnight. For simplicity, I chose to align the data to a cosine, although any other phase would have been mathematically equivalent. It is still possible to shift the prior to accommodate a different expected sleep pattern by merely adjusting an argument in ψ_1. For instance, incorporating π would imply that the prior assumes the lowest point in resting heart rate would occur at midnight.

β represents the population average. ψ_1 describes the first of the 4 components cosinor fitting. The modulus expresses how important this component is in the fitting, and the argument expresses the phase of this component. ψ_1 has argument equal to zero, because I explicitly aligned all users to a cosine, therefore the first component is necessarily aligned with the cosine, having no shift in phase. It can be seen that the first component captures 62% of the overall information, as the modulus of ψ_1 is equal to 0.62 multiplied by the sum of the moduli of all ψ_1, ψ_2, ψ_3, and ψ_4, which means that using single component cosinor we would capture only approximately 40% of the information.

Building the models of a user resting heart rate by fitting θ on the data of that user

To obtain the analysis of the user of interest, I proceeded as follows:

- for every HR measure, I calculated the number of steps in the 5 minutes preceding the timestamp of the HR measure;
- I removed all the HR measures with a corresponding number of steps larger than 10. Therefore I removed all the HR measures that happened during, or immediately after, some physical activity;
- with a rolling window approach, for every HR measure
  - I applied exponential decay to the previous HR measures. The decay function is applied to the weights w_i used in the loss function L. $w_i = e^{-\Delta t}$, where Δt is the time, in days, between the datapoint i and

the latest datapoint, that triggered the update of the model. Therefore, the weight approximately halves every day;

– I then calculated the surrogates S and F, solved the polynomial in ζ, finding the θ parameters of Equation 5.3 that minimise the loss function;

Figure 5.2 shows θ_0, θ_1, and θ_2, calculated at each HR measure.

The choice of using an exponential decay function over time ensures that recent data is weighted more than old data. This approach allows the model to evolve over time, reacting to changes in sleep habits and physiological status. An alternative approach could have been to consider measures in finite time moving windows. However, this latter approach could yield unpredictable results in presence of wide sections of missing data, a frequent situation in data from field experiments, caused by users forgetting to put the band back on after recharging it. An exponential decay approach ensures that even in presence of wide sections with missing data, a usable model is always available.

Figure 5.2 shows seven days of HR and sleep data of one of the users, and the estimated circadian parameters:

- the first graph shows the HR measures of the user as black dots, and the RHR as estimated by the circadian model as red dots. The x-axis is time expressed as UNIX-timestamps, the y-axes is HR in beats per minute;
- the second graph shows the estimated *MESOR*, θ_0, evolving over time, changing for every HR measure. The x-axis is time expressed as UNIX-timestamps, the y-axes is HR in beats per minute;
- the third graph shows the estimated *Amplitude*, θ_1, evolving over time, changing for every HR measure. The x-axis is time expressed as UNIX-timestamps, the y-axes is the ratio between the amplitude of the user model and the amplitude of the population model. A value around 1 (the dotted horizontal line) indicates that the circadian amplitude is in line with the population average, values above 1 indicate that the amplitude is larger than the population average, values below 1 that the amplitude is below the population average;
- the fourth graph shows the estimated *Acrophase*, θ_2, evolving over time, changing for every HR measure. The x-axis is time expressed as UNIX-timestamps, the y-axes is phase in days with respect to a cosine with period 1 day, with range between -0.5 and 0.5 days. The horizontal dotted line shows a phase of zero, indicating that the user sleeps at midday;
- the fifth graph shows the sleep data measures by the wearable, in the graph I show at what time the user went to bed and woke up, as a line. The horizontal dotted line indicates midnight. I did not use this data in my model, and is shown here as a reference for the sleep behaviour of the user.

Looking at the estimated circadian parameters, it can be seen that on the fourth night, slightly after timestamp=1546300000, the *MESOR* increases and the *Amplitude* decreases, indicating non-optimal physical conditions of the user, and that the *Acrophase* suddenly changes, increasing from the stable average. Observing the sleep data

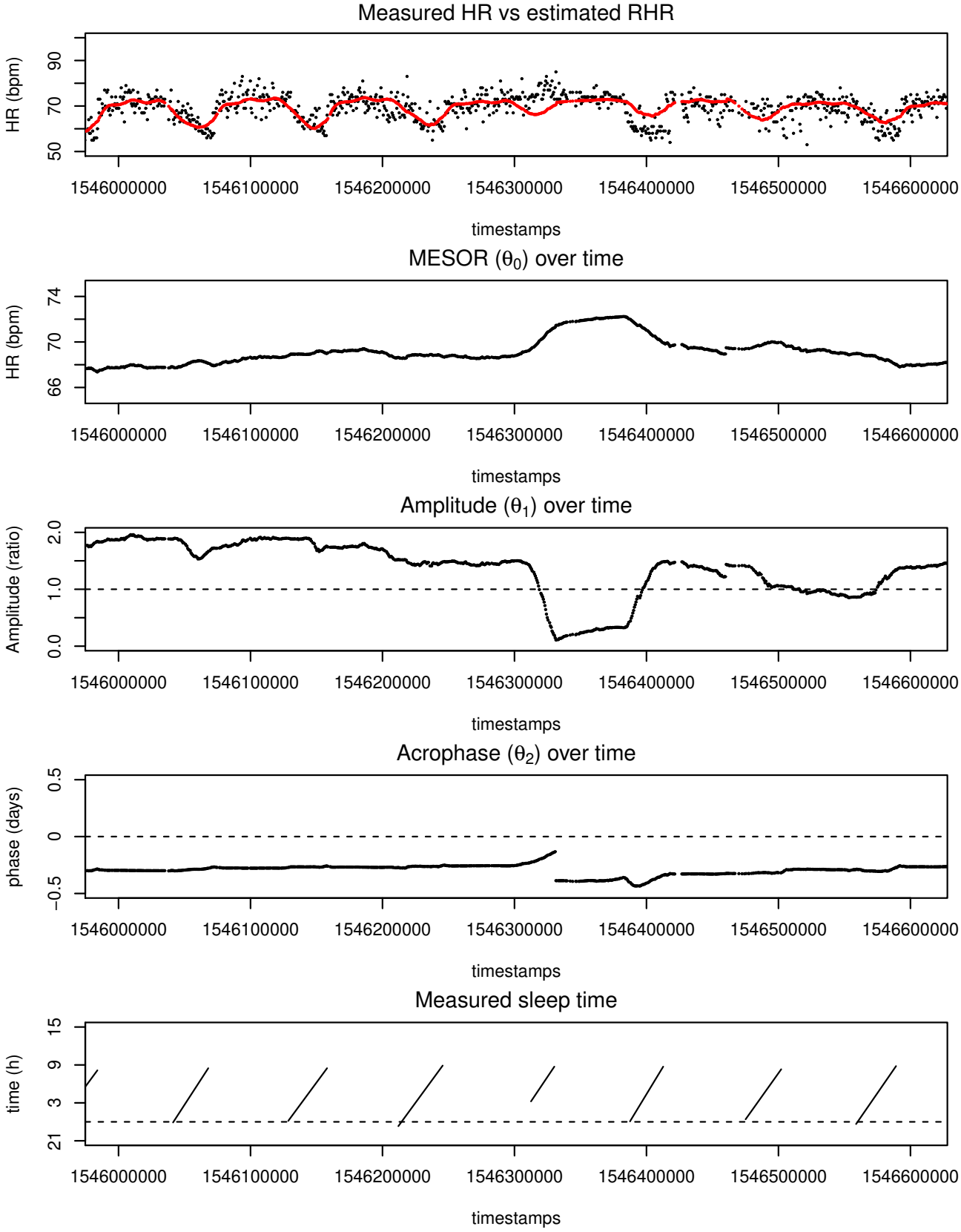


Figure 5.2: A week of HR and sleep data of a user, and fitted circadian parameters. The first row shows the measured HR (black dots) versus the estimated RHR (red line); the second, third, and fourth rows show the evolution of θ_0, θ_1, and θ_2 over time; the fifth row shows the measured sleep data. The unit of the x axis is seconds, the data is represented in Epoch Unix Timestamps (seconds since January 1st, 1970 at UTC).

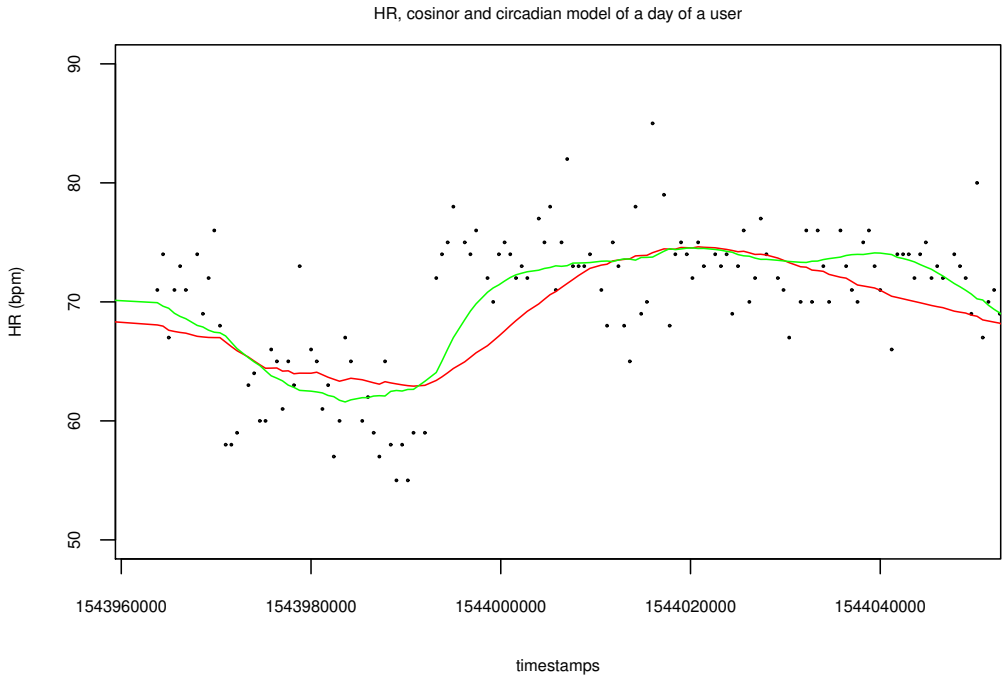


Figure 5.3: One day of HR (beats/minute) data of a user, with fitted single component cosinor (red line) and my model (green line), recalculating the models for every HR datapoint with exponential decay on data

from the band it can be seen that on that same night the user suddenly changed sleep habits (from a healthy stable time to bed of about 11pm, to 3am). The lack of proper sleep can be seen in the HR data and is reflected in the changes in θ_0 and θ_1. The sudden change in the user sleep habit is reflected in the sudden change in θ_2.

Figure 5.3 shows the data of one day of a user, with the fitted single component cosinor model and my circadian model. It can be seen that my model better captures the sudden change in HR when the user awakens, and the fact that the time awake is longer than the time asleep.

Predict user's resting heart rate

I applied the method described in the previous sections to obtain an evolving model of the users. I used those models to predict the next HR, given all the user's previous HR datapoints. I compared the prediction accuracy of my model against the single component cosinor model.

The cosinor RMSE is 5.73 ± 2.39, my circadian model RMSE is 5.15 ± 1.07. My circadian model RMSE is in average 10% lower than the cosinor RMSE. A paired t-test returns, 0.59 mean of the differences between cosinor RMSE and my circadian RMSE, with a 95% confidence between 0.02 and 1.15 (p-value$<$ 0.05).

I also tested my model against the Null Hypothesis that simply using the previous HR measure to predict the current HR measure yields better results than modelling circadian effect on HR. The Null Hypothesis model has RMSE equal to 5.94 ± 1.38. A paired t-test returns a mean of differences between my circadian model RMSE and the Null Hypothesis model RMSE of 0.79, with a 95% confidence between 0.57 and 1.01 (p-value$<$ 0.001).

The paired t-test between cosinor RMSE and Null Hypothesis RMSE returns a non-significant difference of means (p-value > 0.5).

The loss function $L(\theta)$ can be used to calculate the standard error, useful to have an estimation of the distribution of the predicted data. This information could be used to automatically analyse HR activity, finding anomalous data points that could be caused by factors not modelled by circadian effect on RHR, such as physiological conditions.

5.3 Conclusion

In this section I presented a novel mathematical model that can be used to approximate a user's resting heart rate. The model achieves the expressiveness of the multiple component cosinor, i.e. the model fits the data. At the same time the parameters of the model are easy to interpret parameters: *MESOR*, *Amplitude* and *Acrophase*, the same produced by the single component cosinor model. The model does not require uniformly sampled data, and allows to weight each datapoint independently. Hence, the results showed in this section suggest that it is possible to accurately predict circadian RHR rhythm by using wrist worn PPG device permitting to make inference about health status of the people at home without the need of expansive devices. Moreover, my model allows the fitting to require a signal with a predetermined shape (a combination of sinusoid functions with fixed amplitudes and phases). I use this characteristic to force the shape of the expected effect of circadian rhythm on RHR, an operation not easily doable with the method used in literature.

Continuously assessing the evolution of all the parameters of RHR circadian rhythm (i.e., *MESOR*, *Amplitude* and *Acrophase*) it is possible to make inference about all-cause mortality. As I have shown in [160], there is an association between timing of sleep onset and risk of heart diseases. Hence, as also shown by Yuan et Al. [247], a continuous analysis of sleep habits (sleep onset, duration, and overall effect on resting heart rate) may be used to screen for cardiovascular disease using a simple home device [206] or a wearable device.

The model provided in this study is computationally cheap, depending linearly on the size of the data. The computation of the model does not need the full dataset, but only two surrogates (F and S). This implies that the model can be implemented in a streaming approach, where the data is accumulated on the device, with important consequences for security and privacy of the data, that never leaves the user devices. Moreover, this model is, to the best of my knowledge, the only computational model that can be used to predict users resting heart rate and to analyze the data using the parameters traditionally used to describe the circadian modulation of physiological activity. The model I developed goes beyond fitting circadian activity on RHR, and it can be used to fit arbitrary periodic real valued time series, but also vectorial data (e.g. gesture recognition from the accelerometer), or complex data. With an extension to the mathematical model it could be used to fit non-periodic data using a linear combination of an arbitrary set of functions.

A limitation of this study pertains to the selection of the validation experiment, wherein we forecasted future

values of resting heart rate and compared the prediction error between our model and the single-component cosinor. While this experiment gauges the accuracy of our model in fitting the data, it does not demonstrate the utility of the extracted parameters (θ_0, θ_1, and θ_2). In subsequent studies, it would be prudent to assess the θ parameters against sleep-related measurements, such as melatonin levels.

Chapter 6

Estimation of Sinoatrial activity from semi-continuous HR data

SDNN24, the Standard Deviation of Inter-Beat Intervals over 24 hours, is an important metric of cardiovascular health. Wrist-worn fitness wearable devices record heart beats 24/7 providing a complete overview of users' heart status. Due to motion artefacts affecting heart rate data, and the different nature of the heart rate sensor used on wearable devices compared to ECG, traditionally used to compute SDNN24, the estimation of this important HRV metric has never been performed from wearable data. I propose an innovative approach to estimate SDNN24 that only requires Heart Rate normally available on wearable fitness trackers. SDNN24 and the standard deviation of the Average inter-beats intervals (ANN) derived from the HR (obtained in a time window with defined duration, i.e. 1, 5, 10, 30 and 60 minutes), i.e. $ANN = \frac{60}{HR}$ (SDANN$_{HR}$24), were calculated during 24 hours. Power spectrum analysis using the Lomb-Scargle Peridogram was performed to assess frequency domain HRV parameters (Ultra Low Frequency, Very Low Frequency, Low Frequency, and High Frequency). SDNN24, as it encapsulates the entire power spectrum of HRV, leads to an underestimation of values when estimated from HR measurements (referred to as SDANN$_{HR}$24), primarily due to the absence of high-frequency components. Subjects with low and high cardiovascular risk show different power spectra. Distinct differences become evident, particularly in the Ultra Low and Very Low frequency bands, while resembling outcomes are observed in the Low and High frequency bands. Consequently, I have discerned that HR measurements harbour adequate information for distinguishing cardiovascular risk. The semi-continuous HR measurements collected over a 24-hour period, as commonly provided by wrist-worn fitness wearable devices, should indeed prove sufficient for estimating SDNN24 and, in turn, assessing cardiovascular risk.

6.1 Introduction

As discussed in section 1.1.2, SDNN24 is considered *the gold standard* of HRV features for cardiac health [203]. In particular, it was found that people with SDNN24 values below 50 milliseconds (ms), between 50 and 100 ms and above 100 ms could be considered as unhealthy, compromised health and healthy, respectively [108]. For example, patients with SDNN24 values over 100 ms have been found to have a 5.3 times lower risk of mortality at follow-up than those with values under 50 ms [108]. Values of SDNN24 below 80 ms were found to be predictive of cardiac events [72, 101]. Moreover, people with low HRV values are found to be associated with a high risk of a first cardiovascular event (32–45% more likely than people with normal SDNN24 values). In particular, an increase in SDNN24 of 1% is associated with a lower risk of heart failure of about 1% [86]. It was also found that the alteration in cardiac autonomic function is a predictor of several other individual health problems such as diabetes [222], hypertension [89], and sleep quality [232]. Moreover, one of the last findings about SDNN24 is that changes in autonomic nervous system function that can be estimated by change in HRV, have also been found to be a predictor of infection and in particular for the diagnosis of COVID-19 and its related symptoms [88, 21, 100]. A decline in HRV may be a signal of COVID-19 before common symptoms, e.g. dry cough or fever. Monitoring HRV and HR changes may thus help to evaluate the course of this virus.

The fact that SDNN24 is a strong indicator of cardiac health can be explained studying the effect of Sinoatrial Node (SAN) activity on the HRV spectrum [134, 246, 12, 190]. As also discussed in sections 1.1.1 and 4.4.1, SAN provides the human HRV spectrum a characteristic shape between pink and Brownian noise (with frequency power inversely proportional to the frequency itself). I also explored the reasons behind the effectiveness of SDNN24 as a measure of Sinoatrial Node activity. SDNN24, being a standard deviation, aligns with the HRV spectrum, which is predominantly dominated by Ultra Low Frequency components. The influence of SAN on ULF has been well-documented in previous studies [246, 12, 190]. Given the pivotal role that SAN plays in cardiovascular dynamics, the significance of SDNN24 as the 'gold standard' for assessing cardiac health becomes evident.

SDNN24 is a HRV feature that requires 24 hours of continuous recording Inter-Beat Intervals, traditionally achieved using a Holter monitor[1], that makes the data collection difficult during people's everyday life, therefore not performed routinely.

Due to the fact that instruments able to record inter-beats intervals such as wrist-worn wearable devices are more comfortable to be worn by people during daily life compared to medical devices (e.g., Holter), they could in theory be used to estimate SDNN24. However, inter-beats intervals recorded from these devices suffer from high amount of noise and motion artefact, that propagate to HRV features [153, 149], as discussed in the previous Chapters. Furthermore, wrist-worn devices normally only report HR, as these have been proven to be more reliable than inter-beats intervals data which can only be estimated from these devices in both resting and during physical activity with a small error [123, 153, 149, 209, 240, 98], that seldom report the inter-beats intervals data needed to

[1] `https://en.wikipedia.org/wiki/Holter_monitor`

compute SDNN24.

HR (the reciprocal of the average inter-beats intervals duration over 1 minute) can be considered a low pass filter on the inter-beats intervals signal. A low pass filter that lets frequencies lower than 1 minute pass will not affect ULF and VLF that are the main contributors of SDNN24. Therefore it should be in theory possible to estimate SDNN24 using measures of HR, collected over 24h. If my hypothesis will be verified, this study could show that it is possible to estimate SDNN24 from semi-continuous HR measures and continuous inter-beats intervals data are not strictly necessary. This makes the SDNN24 estimation particularly resilient to noisy data (i.e., unevenly sampled data with huge quantity of missing data induced by motion artefacts) a condition normally found with wrist-worn wearable devices. Furthermore, the second objective of this study is to evaluate whether HR data, such as SDNN24, are effective in distinguishing between healthy and unhealthy individuals. If this hypothesis will be verified, the HR data, that are less sensible to missing values [153], could be used as predictor of cardiovascular disease and the lack of usable inter-beats intervals data (more difficult to be reliably measured by wrist-worn fitness wearable devices with heart rate sensors) would not necessarily prevent continuous cardiac health estimation.

This Chapter draws upon a paper published in *Sensors* [152] in 2021. It was developed in collaboration with Rossi, A. from the University of Pisa, as well as Bartoloni, L. and Cairo, M. from BioBeats, all of whom significantly contributed to the analysis.

6.2 Methods

6.2.1 Dataset and feature engineering

In the study presented in this section, I used three PhysioNet datasets providing about 23 hours electrocardiograms (ECG) data of 105 healthy and unhealthy subjects:

nsr2db: Normal Sinus Rhythm RR Interval Database[2], described in section 3.2.1, comprising healthy subjects.

chfdb: Congestive Heart Failure RR Interval [3] Database PhysioNet dataset [78]. This database includes beat annotation files for 29 long-term ECG recordings of subjects aged 34 to 79, with congestive heart failure (New York Heart Association classes I, II, and III). Subjects included 8 men and 2 women; gender is not known for the remaining subjects.

mmash: Multilevel Monitoring of Activity and Sleep in Healthy people dataset described in section 4.2.1

These data are used to simulate the HR data extracted from wrist-worn wearable devices equipped with heart rate sensors. Typically, these devices still only measure the average heart rate in semi-continuous way during the

[2] `https://physionet.org/content/nsr2db/1.0.0/`
[3] `https://physionet.org/content/chfdb/1.0.0/`

day [9] with a different granularity in accordance with battery saving policies arbitrarily decided by the device manufacturers. In section 6.2.3 I estimate the error of HR values recorded by using PPG devices.

6.2.2 Data pre-processing

The python *hrv-analysis* library [4] is used to remove outliers and ectopic beats from signal in each long-term ECG time-series as showed in [193]. 2.46±4.29% of inter-beats intervals in the dataset were ectopic beats (i.e., disturbance of the cardiac rhythm frequently related to the electrical conduction system of the heart) or missing values induced by motion artefacts. These missing values were reconstructed via quadratic interpolation applied on the time domain (i.e. the heartbeat timestamps) as suggested in Chapter 3.

6.2.3 PPG error estimation

As previously discussed, PPG devices are highly affected by motion artefacts that could induce error in HR estimation. However, due to the fact that HR provided by these devices is the mean of HR values during a time window arbitrarily defined by device brand, it is equivalent to applying a low pass filter on instant HR data that permits to remove most of the noise in the time series. In order to assess this error, I compare the average of the instant HR in different time window length (i.e., 1, 5, 10, 30 and 60 minutes) between ECG (Polar H7 heart rate monitor - Polar Electro Inc., Bethpage, NY, USA) and PPG (BioBeam - BioBeats group Ltd, London, `www.biobeats.com`) data from MMASH dataset. Unfortunately, for some technical issues during data recording, I have HR data for both the devices only for a small, but still significant, part of the day (only a few hours per participants, not enough to be used as the main dataset for the analysis, but enough to characterise PPG sensor noise vs ECG ground truth). The mean and standard deviation of HR estimation error for different window lengths (i.e., 1, 5, 10, 30 and 60 minutes) are provided in Table 6.1. It can be seen that the higher is the time window, the lower is the HR bias variability. This result is due to the fact calculating the average heart rate is equivalent to applying a low pass filter on instant HR data, therefore limiting the influence of outliers caused by motion artefacts on the average HR.

Table 6.1: HR estimation error. The estimation errors are expressed as mean±standard deviation.

Time window	**error**
1 minute	0.03±5.91
5 minutes	0.33±5.09
10 minutes	-0.16±4.71
30 minutes	-0.60±4.01
60 minutes	-0.35±3.95

[4] `https://pypi.org/project/hrv-analysis`

6.2.4 SDNN24 and its approximations

From each user SDNN24 were computed and approximated using multiple approaches: time domain, frequency domain, circadian rhythm.

Approaches based on time domain analysis

SDNN24: was computed as described in section 1.1.2. This is the ground truth of this experiments.

SDNNi24: was also computed as described in section 1.1.2.

SDANN24: The standard deviation of the means of NN intervals calculated at segments of a defined duration, during 24 hours (Equation 6.1).

$$\text{SDANN24} = \sqrt{\frac{\sum_{j=1}^{N_{\text{segments}}} (\overline{\text{NN}_j} - \frac{\sum_{j=1}^{N_{\text{segments}}} \overline{\text{NN}_j}}{N_{\text{segments}}})^2}{N_{\text{segments}}}} \tag{6.1}$$

where $\overline{\text{NN}_j}$ refers to the mean of the NN-intervals in the j^{th} segment. N_{segments} is the number of NN-intervals time windows recorded during 24 hours. From a digital signal processing perspective SDANN24 can be considered similar to SDNN24 applied on a inter-beats intervals dataset after a low pass filter that dampened signals with a shorter period than the duration of the measured segments. To compute SDNN24 it is necessary to have the inter-beats intervals data of each segment, typically not available from wearable devices.

SDANN$_{\text{HR}}$24: The standard deviation of the Average NN intervals (ANN) derived from the HR, i.e. ANN $= \frac{60}{\text{HR}}$, calculated on segments with defined duration, during 24 hours (Equation 6.2).

$$\text{SDANN}_{\text{HR}}\text{24} = \sqrt{\frac{\sum_{j=1}^{N_{\text{segments}}} (\frac{60}{\text{HR}_j + \epsilon_j} - \frac{\sum_{j=1}^{N_{\text{segments}}} \frac{60}{\text{HR}_j + \epsilon_j}}{N_{\text{segments}}})^2}{N_{\text{segments}}}} \tag{6.2}$$

where HR_j is computed as $HR_j = \frac{60}{\overline{\text{NN}_j}}$, and $\overline{\text{NN}_j}$ is the average of the NN intervals that belong to the j^{th} window of 5 minutes of data. SDANN$_{\text{HR}}$24 can be computed from data commonly collected by wrist worn fitness wearable devices. In order to simulate error induced by wrist-worn devices I randomly add a Gaussian error ϵ in each HR obtained during each NN-intervals time window. The bias for each time windows length is provided in Table 6.1.

Approaches based on frequency domain analysis

The power spectrum $S_{xx}(f)$ of a time series $x(t)$ describes the distribution of power into frequency components (ω) composing that signal. Due to the fact that the inter-beats intervals are not uniformly distributed (unevenly sampled time-series) the PSD is computed using the Lomb–Scargle Periodogram [130, 199] instead of the Fourier

Transformation that requires uniformly sampled data. The PSD can be used to compute the variance (net power) of a process by integrating over frequencies as showed in Equation 4.1.

Consequently, it is possible to estimate SDNN24 by PSD derived from a 24 hours time-series by the square of its net power (VAR($x(t)$)). Due to the fact that SDANN$_{\text{HR}}$24 reflects frequencies after a low pass filter that corresponds to the segments length, missing frequencies induce an underestimation of SDNN24. For example, if the 60/HR time-series provide values over 5 minutes, it is possible to assess only frequencies lower than 0.0033. For these reasons I computed SDNN24 in 2 other ways by adding the estimated variance of the missing frequency on the variance of 60/HR time-series with a defined segments length over the 24 hours. To this aim, I computed *a priori* PSD on *nsr2db* and *chfdb* and test the validity of this approach on *mmash* dataset.

I evaluate two ways of correcting the bias of discarding high frequencies when using any approach based on HR measures:

SDANN$_{\text{HR}}$24$_{\text{adjMean}}$: The root square of the sum of the square of SDANN$_{\text{HR}}$24 and the mean of *a priori* missing frequency variance (Equation 6.3).

$$\text{SDANN}_{\text{HR}}24_{\text{adjMean}} = \sqrt{(\text{SDANN}_{\text{HR}}24)^2 + \frac{1}{\pi}\int_{freq}^{\infty} S_{xx}(\omega)d\omega} \tag{6.3}$$

where:

- $freq$ refers to the frequency corresponding to the length of the segment in $\frac{60}{\text{HR}}$ time-series (e.g., for 1, 5 and 10 minutes segments the $freq$ are equal to 0.016, 0.0033 and 0.0016, respectively);
- and $\frac{1}{\pi}\int_{freq}^{\infty} S_{xx}(\omega)d\omega$ is the power of the spectrum for frequencies higher than $freq$, of the general population.

With this approach I attempt at removing the underestimation of SDANN$_{\text{HR}}$24 by adding the portion of spectrum lost by using HR measures instead of inter-beats intervals data, simply adding the average power of the HRV spectrum above $freq$ to the measured variance. The corrective factor is fixed for all subjects.

SDANN$_{\text{HR}}$24$_{\text{adjW}}$: The root square of the sum of the square of SDANN$_{\text{HR}}$24 and the power of the missing frequencies predicted in accordance with *a priori* PSD (Equation 6.4).

$$\text{SDANN}_{\text{HR}}24_{\text{adjW}} = \sqrt{(\text{SDANN}_{\text{HR}}24)^2 + \frac{1}{\pi}\frac{\int_0^{freq} S_{xx}(\omega)d\omega_{ANN} * \int_{freq}^{\infty} S_{xx}(\omega)d\omega}{\int_0^{freq} S_{xx}(\omega)d\omega}} \tag{6.4}$$

where:

- $freq$ refers to the frequency corresponding to the segment in $\frac{60}{\text{HR}}$ time-series (e.g., for 1, 5 and 10 minutes segments the $freq$ are equal to 0.016, 0.0033 and 0.0016, respectively);

- $\frac{1}{\pi}\int_{freq}^{\infty} S_{xx}(\omega)d\omega$ is the power of the spectrum for frequencies higher than $freq$, of the general population;
- $\frac{1}{\pi}\int_{0}^{freq} S_{xx}(\omega)d\omega_{ANN}$ is the power of the frequencies lower than $freq$ for the specific subject;
- and $\frac{1}{\pi}\int_{0}^{freq} S_{xx}(\omega)d\omega$ is the power of the frequencies lower than $freq$ for the general population.

With this approach I correct the underestimation by assuming that the missing high frequencies perfectly correlate with the measured low frequencies.

Approach based on circadian rhythm.

A multiple component cosinor model (Equation 6.5) was fitted on HR values recorded during 24 hours in order to assess the HR circadian rhythm [150].

$$\mathrm{HR}_t = M + A * cos(2\pi * (\frac{t}{\mathrm{freq}} + \phi)) + \epsilon_t \tag{6.5}$$

where M is the MESOR (Midline Estimating Statistic Of Rhythm, a rhythm-adjusted mean), A is the amplitude (a measure of half the extent of predictable variation within a cycle), ϕ is the acrophase (time of the day when the high HR values recurs in each cycle), t is the period (duration of one cycle), freq is the fixed length of the cycle (i.e., 24 hours) and ϵ_t is the error term. The circadian parameters were then used as features for Machine Learning models described in the remaining of this Chapter.

Validation

The statistical differences between SDNN24 (used in this experiment as the gold standard) and SDNNi24, SDANN24, and SDANN$_{HR}$24 was assessed by paired t-test, while their relationship was evaluated by Pearson correlation coefficient. All the Physionet datasets were used in this analysis.

The PSD of the 24 hours RR time-series was estimated using the Lomb-scargle Periodogram [130, 199]. *A Priori* PSD was created on $chfdb$ and $nsr2db$ and was used to estimate $\mathrm{SDANN_{HR}24_{adjMean}}$ and $\mathrm{SDANN_{HR}24_{adjW}}$ on $mmash$ dataset. This approach was used in order to avoid any overfitting problem caused by assessing missing frequency on the same data were was used to predict SDNN24.

Paired t-test, Pearson correlation coefficient and Root Mean Squared Error (RMSE) was assessed to evaluate statistical difference, relationship and error between gold standard and estimated features. All the analysis was conducted on 1, 5, 10, 30 and 60 minutes segments time length in order to assess the maximal time window that provide reliable results.

6.2.5 Healthy vs unhealthy subjects

Statistical analysis.

Following the methodology outlined in [244], the *nsr2db* dataset was designated as the control group for *healthy* individuals. In contrast, the *chfdb* dataset was selected for the *disease* group, encompassing subjects diagnosed with congestive heart failure (New York Heart Association classes II and above) and those at risk (class I). HRV features (i.e., ULF, VLF, LF and HF), SDNN24 and $\text{SDNN}_{\text{HR}}24$ (computed on 5-minute time window differences) between people with low and high cardiovascular risk were assessed by unpaired t-test. The T-score derived from t-test analysis was used to assess the magnitude of the difference between healthy vs unhealthy subjects.

Approaches based on Machine Learning

In order to discriminate between healthy and unhealthy subjects grouped as shown in section 6.2.5 the following Machine Learning (ML) models were evaluated.

LR : Logistic Regression was performed using: only SDNN24 ($\text{LR}_{\text{SDNN24}}$); only $\text{SDNN}_{\text{HR}}24$ ($LR_{\text{SDNN}_{\text{HR}}24}$); and all of the HR features as predictors, i.e. $\text{SDNN}_{\text{HR}}24$, MESOR and Amplitude ($LR_{\text{HR}}$);

RF_{HR} : Random Forest Classifiers (RF) were also performed using all the HR features as predictors;

NN : Fully connected feed forward Neural Network, using all the HR features as predictors. I used *keras* with the *TensorFlow* back-end. I trained my neural networks on the Azure cloud, using bayesian sampling. The only explored topology was fully connected, with a single hidden layer, *leaky relu* activation function for the neurons of the hidden layer, a single output neuron with *sigmoid* activation function, and a dropout layer after the hidden layer. The tuned hyper-parameters were:

- the number of neurons in the hidden layer (between 1 and 8);
- alpha value for the *leaky relu* activation function of the neurons in the hidden layer (between 0.0 and 1.0);
- the dropout rate (between 0% and 99%);
- the batch size (between 1 and 32).

The training set was split in train and validation using the *train_test_split* function from the *sklearn* python package, using a 80-20 split on subjects, ensuring stratification on the predicted class. The validation data was not used during hyper-parameter tuning. A total of 400 combinations of hyper-parameters were tested.

In order to assess the validity of the classifiers I compared my predictive models with two baselines. Baseline B1 randomly assigned a class to an example by respecting the distribution of classes, while Baseline B2 always assigned the majority class.

Table 6.2: Descriptive and statistics analysis of SDNN24 estimations. All the values are expressed in milliseconds (ms).

Segment	SDNN24	SDNNi24	SDANN24	SDANN_{HR}24
1 minute		56.28±18.51*	119.38±49.51*	120.96±48.91*
5 minutes		59.47±21.52*	117.33±48.03*	118.62±47.00*
10 minutes	133.73± 45.98	60.77±23.02*	116.72±47.37*	117.94±46.30*
30 minutes		63.11±26.26*	115.75±46.69*	117.12±45.96*
60 minutes		64.61±28.52*	114.87±46.23*	116.29±45.85*

*p-value < 0.01.

The models were trained on 80% of the dataset and tested on the remains 20%, and the goodness of the predictions were assessed using Precision, Recall and F1-score.

6.3 Results

6.3.1 SDNN24 estimation

Time domain analysis.

Descriptive statistics of all the SDNN24 estimations computed on time domain are provided in Table 6.2. Statistical differences ($p<0.001$) were detected for each estimated feature on each segment length. The longer is the segment length the higher is the difference among SDNN24, SDANN24 and $\text{SDANN}_{\text{HR}}24$. Strong positive correlations were detected between SDNN24, SDANN24 (r=[0.95-0.96]), and $\text{SDANN}_{\text{HR}}24$ (r=[0.93-0.94]) in all of the segments length. Differently, the mean of the standard deviations of the NN intervals calculated on defined segments during 24 hours shows an opposite trend. The higher is the segment length the lower is the difference between SDNN24 and SDNN_i24. Moreover, moderate relationship was found between SDNN24 and SDNN_i24 (r=[0.67-0.71]).

Frequency domain analysis.

Figure 6.1 provides the power spectrum of the 24 hours RR-intervals time series on users in $nsr2db$ and $chfdb$ datasets. SDNN24 measures the total power of the spectrum of the analysed data, therefore including both ultra low (ULF), very low (VLF), low (LF), and high frequencies (HF), as well as the frequencies above HF [203]. When analysing human HRV, ULF are defined as the fluctuations with a period between 5 minutes and 24 hours, therefore capturing mainly circadian effects; VLF measures oscillations between 25 seconds and 5 minutes; LF measures oscillations between 7 and 25 seconds, is known to be influenced by the activity of both the Parasympathetic Nervous System and the Sympathetic Nervous System; HF measures oscillations between 7 and 2 seconds, and is known to be influenced by Sympathetic Nervous System and the respiratory rate. The strong linear relationship (r=0.98, p-value<0.001) detected between ULF and Total Power (i.e., integration of the spectral components) suggests that it is possible to accurately predict Total Power and consequently SDNN24 from ULF.

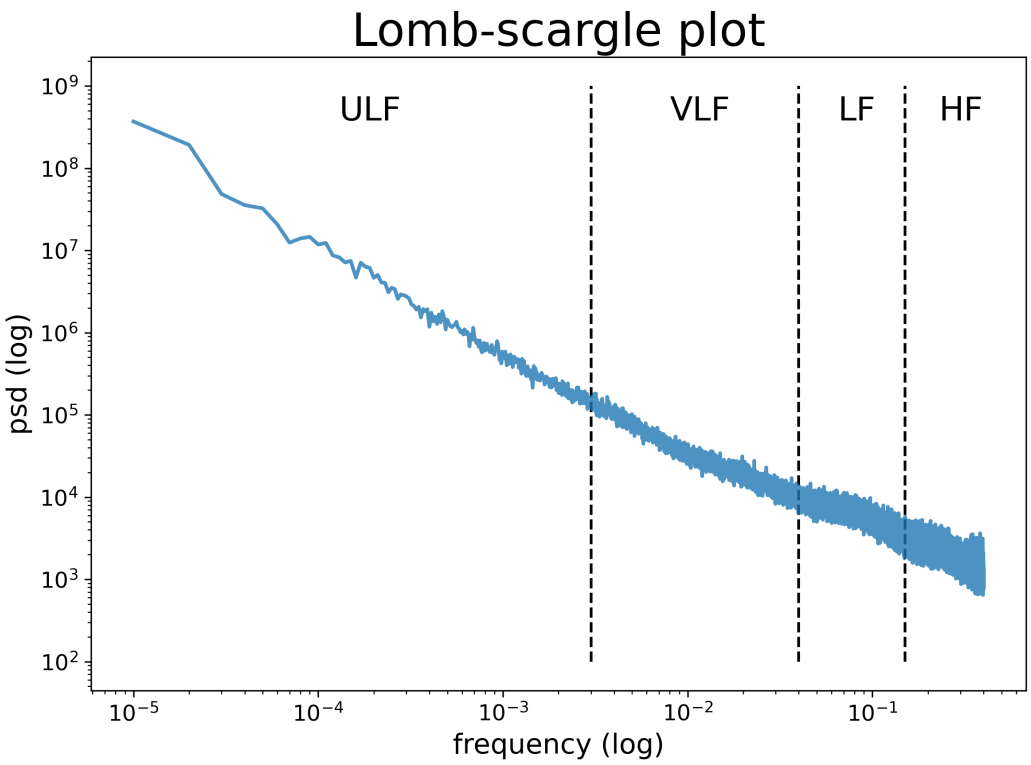


Figure 6.1: Power spectrum density plot obtained by Lomb-Scargle Periodogram.

Lower values obtained from SDANN24 and $\text{SDANN}_{\text{HR}}24$ compared to SDNN24 (see Table 6.2) is due to the fact that only a subset of the spectrum are measured because of the segment length. For example, in 5-minute segments, it is possible to assess frequencies lower than 0.0033 (i.e., ULF), while higher frequencies are missing. Even if the low frequency reflects a huge part of the total power, the missing frequency do not permit to accurately estimate SDNN24, underestimating it.

Table 6.4 shows that $\text{SDANN}_{\text{HR}}24_{\text{adjW}}$ is able to accurately estimate SDNN24 with all the segments length with a bias from 2.18±21.06 ms (RMSE = 21.17 ms) to 15.23±26.86 ms (RMSE = 28.40 ms) for 1 minute and 60 minutes segments, respectively. The strong correlation (r>0.97) found between low frequencies and total power in all of the segments length allows to accurately estimate the total power and consequently SDNN24. Differently, only 1 minute segments show reliable results for $\text{SDNN}_{\text{HR}}24$ (bias=9.54±23.04 ms, RMSE=24.94 ms, p-value=0.07), but any segments show similar result for $\text{SDANN}_{\text{HR}}24_{\text{adjMean}}$.

6.3.2 Healthy vs unhealthy subjects.

Figure 6.2 shows that the differences between people with low and high risk of cardiovascular disease in PSD are found on ULF (p-value<0.001) and VLF (p-value<0.001). In accordance with the hypothesis that low frequency (ULF and VLF) are affected by cardiovascular health more than other frequency bands (LF and HF), Table 6.5 shows that $\text{SDNN}_{\text{HR}}24$ provides more differences (t-score) between people with low and high cardiovascular risk

Table 6.3: Descriptive of power spectrum analysis expressed in ms^2 in all the segments length. In this table it is reported the power of frequencies that are lower and higher than the maximal frequency that can be evaluated because of the segment length. The results are reported as the mean±standard deviation.

Segment	Lower Frequency	Higher Frequency	Total Power
1 minute max frequency = 1.67e-2 hz	1.17e+4±8.68e+3 (20.44%)	4.57e+4±2.86e+4 (79.56%)	5.74e+4±3.66e+4
5 minute max frequency = 3.33e-3 hz	1.12e+4±8.47e+3 (19.41%)	4.62e+4±2.86e+4 (80.59%)	
10 minute max frequency = 1.67e-3 hz	1.09e+4±m8.35e+3 (18.88%)	4.66e+4±2.87e+4 (81.22%)	
30 minute max frequency = 5.55e-4 hz	1.03e+4±8.09e+3 (17.85%)	4.72e+4±2.90e+4 (82.25%)	
60 minute max frequency = 2.78e-4 hz	9.77e+3±7.90e+3 (17.01%)	4.77e+4±2.92e+4 (82.98%)	

Table 6.4: Descriptive and statistics analysis of SDNN24 and SDNN24 derived from HR data on *mmash* dataset. All the values are expressed in milliseconds (ms).

Segment	SDNN24	$\mathbf{SDNN_{HR}24}$	$\mathbf{SDANN_{HR}24_{adjMean}}$	$\mathbf{SDANN_{HR}24_{adjW}}$
1 minute	173.53±25.76	164.00±32.52	213.63±20.81*	175.65±25.65
5 minutes		153.02±32.54*	209.23±19.79*	165.64±31.59
10 minutes		149.84±31.00*	198.92±23.90*	161.51±30.96
30 minutes		146.02±32.47*	197.60±24.55*	164.21±35.98
60 minutes		142.11±35.02*	198.38±31.10*	155.23±40.00

*p-value < 0.01.

compared to SDNN24. Similar results are detected from others segments length. Hence, $SDNN_{HR}24$ is more sensible to discriminate between healthy and non-healthy people due to the fact that it assess only the low frequencies - that is more sensible to cardiovascular risk level as showed in Figure 6.2 - instead of the total power such as SDNN24.

The ML models demonstrate that HR parameters permit to accurately predict SDNN24. In particular, the forward stepwise linear regression shows that it is possible to predict SDNN24 by using HR circadian rhythm parameters and $SDANN_{HR}24$ as showed in Equation 6.6. A very low error is detected between SDNN24 predicted and observed showing a difference of about 0.22±11.47 (RMSE = 53.81 and $r^2 = 0.97$).

$$\text{SDNN24} = 47.248 + 951.590 * \text{SDANN}_{HR}24 - 0.347 * \text{MESOR} \tag{6.6}$$

Moreover, HR is found to be also slightly more informative than SDNN24 to discriminate healthy and unhealthy

Table 6.5: Difference between healthy subjects and subjects with cardiovascular disease in all the segments length.

Segment	Features	Disease	Healthy	t-score
—	**SDNN24 (ms)**	86.54±43.29	142.14±31.05	6.66*
1 minute	$\mathbf{SDNN_{HR}24}$ **(ms)**	67.86±37.23	132.61±30.79	8.28*
5 minute	$\mathbf{SDNN_{HR}24}$ **(ms)**	64.46±37.23	128.31±30.63	8.27*
10 minute	$\mathbf{SDNN_{HR}24}$ **(ms)**	62.55±36.67	126.02±30.52	8.29*
30 minute	$\mathbf{SDNN_{HR}24}$ **(ms)**	58.37±35.45	121.96±30.44	8.44*
60 minute	$\mathbf{SDNN_{HR}24}$ **(ms)**	55.75±34.60	118.81±30.26	8.49*

*p-value < 0.01.

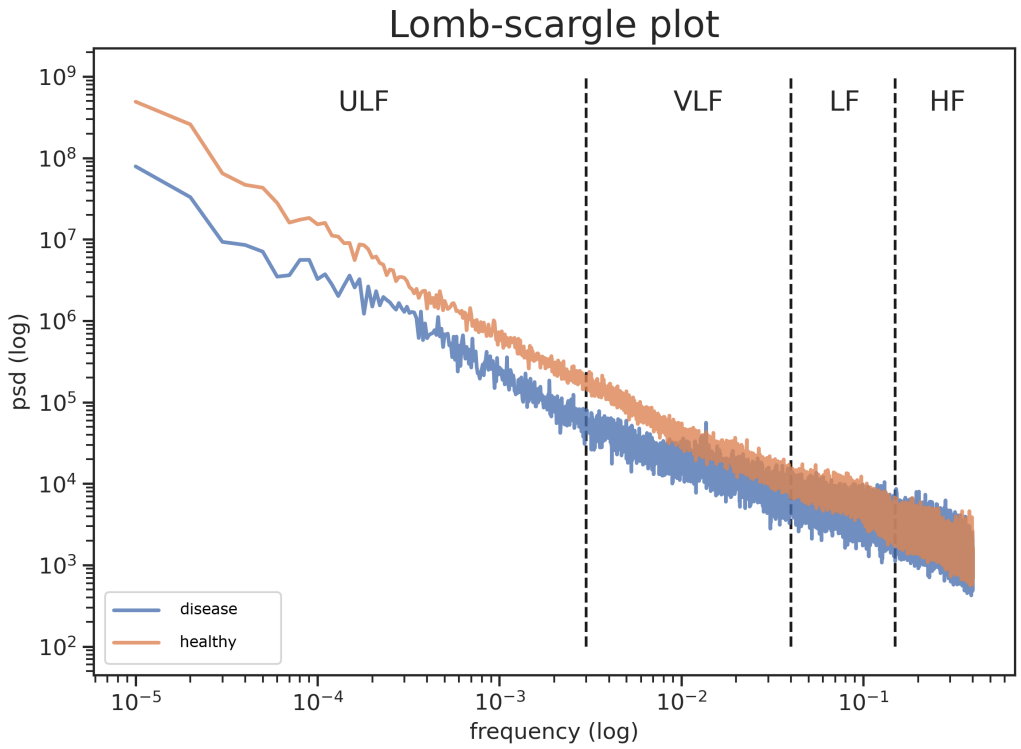


Figure 6.2: Power spectrum density plot obtained by Lomb-Scargle Periodogram in both healthy subjects, and subjects with cardiovascular disease.

people improving the prediction performance of about 9%. In particular, the highest prediction performance is detected by LR_{HR} (Table 6.6). Equation 6.7 shows the LR function that best discriminates subjects with high risk of cardiovascular diseases. As expected, the fully connected neural network also reaches high predictive capabilities. However, since the predictive performance of LR_{HR} and NN are comparable, LR_{HR} should be preferred over NN, due to its simplicity and interpretability.

$$f(x) = \frac{1}{1 + e^{0.121 + 0.283 * \text{SDANN}_{\text{HR}}24 - 0.293 * \text{MESOR} + 0.225 * \text{Amplitude}}} \tag{6.7}$$

6.4 Discussion

This study has shown that SDNN24 may possibly be estimated from HR data, without the need to have 24 hour inter-beats intervals data available. Continuous or semi-continuous HR measures are nowadays affordable via wearable devices equipped with a PPG sensor, such as fitness trackers. Estimating HRV features from wearable devices is known to be problematic because of noise induced by motion artefacts that affect the frequencies normally used to assess the activity of the Autonomic Nervous System, such as RMSSD and the Sympathovagal Balance Index (SVI) [149, 153]. However, HR is calculated taking the average of the duration of the inter-beats intervals time-series that is equivalent to applying a low pass filter that permits to filter out most of the noise.

Table 6.6: Model performance. The model with highest performance is reported in bold.

Model	Class	Precision	Recall	F1-score
LR_{SDNN24}	Low	0.84	1.00	0.91
	High	1.00	0.50	0.67
$LR_{SDNNi24}$	Low	0.76	1.00	0.86
	High	1.00	0.17	0.29
$LR_{SDANN24}$	Low	0.80	1.00	0.89
	High	1.00	0.33	0.50
$LR_{SDANN_{HR}24}$	Low	0.80	1.00	0.89
	High	1.00	0.33	0.50
LR_{HR}	Low	**0.94**	**1.00**	**0.97**
	High	**1.00**	**0.83**	**0.91**
RF_{HR}	Low	0.80	1.00	0.89
	High	1.00	0.33	0.50
NN	Low	0.94	1.00	0.97
	High	0.80	1.00	0.89
$B1$	Low	0.72	0.81	0.76
	High	0.25	0.17	0.20
$B2$	Low	0.73	1.00	0.84
	High	0.00	0.00	0.00

In theory, to truly estimate SDNN24 I need precise information from the whole power spectrum, i.e. from ULF to HF and above. However, analysing the normative values of the human HRV power spectrum (see Table 6.3 and Figure 6.1), I notice that the power of ULV is larger than the power of VLF by several orders of magnitude. In turn, the power of VLF is larger than LF, which in turn is larger than the power of HF. The contribution of HF and LF to SDNN24 is therefore a residual fraction of SDNN24, compared to the importance of ULF and VLF [22]. As discussed in the introduction, the activity of the Sinoatrial Node is indicative of cardiac health [135] and responsible for the characteristic $1/f$ shape of the HRV spectrum [190]. Therefore ULF and VLF will yield more information about Sinoatrial Node activity than LF and HF.

Because semi-continuous HR measures are equivalent to measuring the heart activity after a low pass filter, it is a stable proxy for low HRV frequencies. Due to the fact that SDNN24 reflects the total power of the spectrum, the values estimated by splitting into segments the inter-beats intervals time-series (i.e., SDNNi24, SDANN24 and $SDANN_{HR}24$) underestimate the real values because of the missing high frequencies (Table 6.2 and Table 6.3). However, the higher percentage of the total power comes from ultra and very low frequencies (Table 6.3), it is therefore possible to estimate SDNN24 with a bias. Table 6.2 shows that the length of the segments affect the magnitude of the bias, the shorter are the segments where HR are obtained, the lower are the bias between SDNN24 and estimated values.

By adding the expected power spectrum of the missing frequencies to the variance estimated via HR measures (i.e., $SDANN_{HR}24_{adjMean}$ and $SDANN_{HR}24_{adjW}$), it is possible to obtain accurate estimates of SDNN24. In particular, Table 6.4 shows that no statistical differences were detected between SDNN24 and $SDANN_{HR}24_{adjW}$ in all the segments length, while no statistical difference was only detected on 1 minute segments for $SDANN_{HR}24_{adjMean}$. The

strong linear relationship detected between the power of the low frequencies and the total power (r>0.98) permits to accurately estimate the power of the total spectrum by assessing only a small part of that ($SDANN_{HR}24_{adjW}$). Differently, adding constant power of the missing frequency observed in *a priori* PSD permits at the variance of ANN computed with different time window length ($SDANN_{HR}24_{adjMean}$) is not a good proxy for SDNN24 because the power of the missing frequency are not weighted in accordance with the magnitude of the low frequencies observed. Obviously, the shorter is the time window, the lower is the bias between SDNN24 and $SDANN_{HR}24_{adjMean}$ due to less and less missing frequencies.

Figure 6.2 shows that HRV of people who are healthy and those with cardiovascular disease exhibit different power spectra. In particular, higher ULF and VLF were detected for healthy individuals compared to those with cardiovascular disease, while similar results are shown for LF and HF. For this reason, $SDNN_{HR}24$, which reflects ultra and very low frequencies, showed a higher difference between healthy and unhealthy individuals (Table 6.5) for all the segment lengths. This results is corroborated from the fact that Machine Learning models were able to classify cardiovascular disease from HR measures (Table 6.6). These results indicate that semi-continuous HR measures contain enough information to assess cardiac health.

6.5 Conclusion

HR data permits to accurately estimate SDNN24 and it is found to be also slightly more informative to discriminate healthy and unhealthy people. The information contained in HR measures throughout 24 hours (ultra and very low frequencies), should be sufficient to estimate SDNN24 and the people health status from wrist-warn fitness wearable devices that provide only fragmentary HR data throughout the day. This result makes the SDNN24 estimation particularly resilient to noisy data and consequently could be easily and accurately estimated with wrist-worn wearable devices. This result indicates that HR fitness trackers have the potential to implement continuous monitoring of cardiovascular health from passively collected data, that could enable targeted interventions at early signs of deterioration of Sinoatrial Node activity.

A limitation of this study is the assumption that measured SDNN24 values were error-free, thereby treating them as the gold standard. In reality, these measurements will likely contain some degree of error, leading to a conflation of measurement and estimation error. In cases like ours, where the reference measure cannot be assumed to be error-free, a more suitable approach would be to employ Bland-Altman plots. This method is particularly indicated for such situations as it allows for a more nuanced analysis of agreement between two quantitative measurements, accounting for the potential errors in both the method under investigation and the reference measure.

Chapter 7

Health outcomes predictions from HRV during exercise

The goal of this Chapter is to explore for which health outcomes HRV should be considered as a valuable biomarker. This is achieved by studying when HRV can provide information not already captured by Electronic Health Records (EHR). I build a representation based on Convolutional Neural Networks for Heart Rate Variability during exercise data, because there is no standard HRV feature that can effectively describe the dynamics of heart rate during exercise. To avoid confounding effects and to study if HRV can provide additional information to what's available to GPs during routine visits, I also build a representation of the whole medical history of the subjects using BERT (described in section 7.4.1). I then perform an extensive study on a set of identified health outcomes that showed potential in a review of clinical literature of the relationship between HRV and health outcomes. I then find the best hyper parameters and train predictive models for all possible combinations of inputs: HRV data, Electronic Health Record, and basic information about the user (including age, sex, ethnicity, diet, alcohol, body composition, and physical activity). I then discuss the results obtained. I outline potential future work enabled by this research.

7.1 Introduction

In section 7.2 I define the health outcomes that will be used as target labels for the models we will train in this study. In sections 7.5, 7.4, and 7.6 I describe the experiment I designed and conducted to find in which cases HR during exercise can provide useful information, not already captured by the patients medical history, for the health outcomes previously defined. The value of this research is to identify potential cases in which the analysis of HR dynamics during exercise provides information that is unlikely to be explainable with confounding variables.

In section 7.5 I describe a model capable of describing the heart rate dynamics in response to physical exercise, using the *ECG during exercise* data in UKB. In the tables and graphs I call this model *HRV*. In section 7.4 I

describe a model capable of generating a representation of the patient's medical history as a tensor of fixed size, generating an embedding of the patient's Electronic Health Records present in UKB. In the tables and graphs I call this model *EHR*. I also selected a small number of basic information about the subjects, like age, gender, diet, smoking status, and a few other, representing simple information that can be obtained from the subjects by asking questions. In the tables and graphs I call this model *BASIC*.

Finally, I trained models to predict the health outcomes of interest looking at all the possible inputs and their combinations (*BASIC*; *HRV*; *EHR*; *BASIC+HRV*; *BASIC+EHR*; *EHR+HRV*; *BASIC+HRV+EHR*). I then analysed the results and the difference in performance of all the models for all health outcomes. I used both binary predictions and survival analysis. Section 7.6 reports the results in a binary classification setting, and section 8.2.3 in a survival analysis setting.

7.2 Definition of the Health Outcomes of interest

To establish the list of health outcomes we aim to predict using HRV, my approach involved an analysis of the existing research detailed in the literature review in section 1.1.3. The primary criterion for selection was the identification of outcomes where a significant disparity in HRV has been observed between individuals in healthy states and those with specific conditions. I purposefully omitted scenarios where the observed HRV changes could be attributed to confounding factors. For instance, mental health conditions often coincide with medication or substance use that impacts the autonomic nervous system, and consequently, HRV. In such instances, the link between HRV and the health outcome is not causal, leading to their exclusion from this study.

We chose to examine Coronary Heart Disease (CHD) as a distinct category from other Cardiovascular Disease (CVD) outcomes. This decision was based on literature suggesting that Heart Rate Recovery (HRR) is a particularly strong predictor of CHD [175]. Our study will explore the effectiveness of Heart Rate Variability (HRV) during exercise, which we hypothesise is related to HRR, in predicting both CVD and CHD outcomes. The conclusions will be drawn based on the comparative analysis of these predictions.

Angina pectoris is known to be linked with physical exercise, as reported in Routledge et al. (2010)[194]. Recent studies, including Zhang et al. (2023) [250], have demonstrated that resting Heart Rate Variability (HRV) can specifically and accurately predict angina pectoris. Consequently, we have opted to investigate angina pectoris as a separate outcome from other cardiovascular outcomes, such as Cardiovascular Disease (CVD), Coronary Heart Disease (CHD), and Hypertension.

Heart Rate Variability (HRV) has been extensively studied as a predictor of hypertension [201, 212], and physiological pathways causally linking HRV to hypertension have been identified [248]. Given that hypertension is a known risk factor and precedes the development of more complex cardiovascular diseases, as outlined in the cardiovascular disease continuum [37], we focused on studying High Blood Pressure (HBP) in isolation from other

Target	Inclusion criteria	ICD10 codes
CVD	Myocardial Infarctions	I21, I22, I23
	Syncopal Events	R55
	Coronary Revascularization	Y84.0, Z98.61, Z95.5
	Subarachnoid Haemorrhage	I60
	Intracerebral Haemorrhage	I61
	Cerebral Infarction	I63
	Stroke, not specified as Haemorrhage or Infarction	I64, G45
	Transient Ischemic Attack	I24
CHD	Coronary Heart Disease (predicted by HRR)	I25
Angina	Unstable Angina	I20.0
HBP	Any Hypertensive disease	I10-I15

Table 7.1: Definition of the targets for the experiments presented in this Chapter.

cardiovascular diseases (CVD).

I therefore define my outcomes as shown in table 7.1.

From the analysis of literature review I also identified the data that is known to affect HRV and the health outcomes of interest, that should therefore be included in the experiments:

- Information available in UK Biobank fields (these fields will be explicitly added): Age; Smoking; Waist to hip ratio; Activity level (MET);
- Information present in the EHR data (therefore already present in the patients representation): Hypertension; Diabetes; Atherosclerosis; hypertensive cardiac hypertrophy; Cancer.
- Medications (we will use this as exclusion criteria): β blocking agents, Calcium antagonists, ACE inhibitors: BNF code 114086.

7.3 Dataset

The *ECG during exercise* data present in UK Biobank is ECG data collected during a fitness test of the duration of approximately 7 minutes. The study was conducted using a 4-lead ECG (CAM-USB 6.5, Cardiosoft v6.51), during cycle ergometry on a stationary bike (eBike, Firmware v1.7) as shown in figure 7.1 [1] The data collected from the cycle ergometry task includes ECG and workload (W), has a total duration of the exercise of 440 seconds, and it includes 3 phases:

warm-up the subject is sitting still on the bicycle for 20 seconds before the start of the physical exercise;

exercise the subject is asked to cycle at varying slopes for 6 minutes;

recovery the subject stops cycling and sits on the stationary bike for 1 minute after the physical exercise.

[1] Details extracted from the UK Biobank website `https://biobank.ndph.ox.ac.uk/ukb/label.cgi?id=100012`

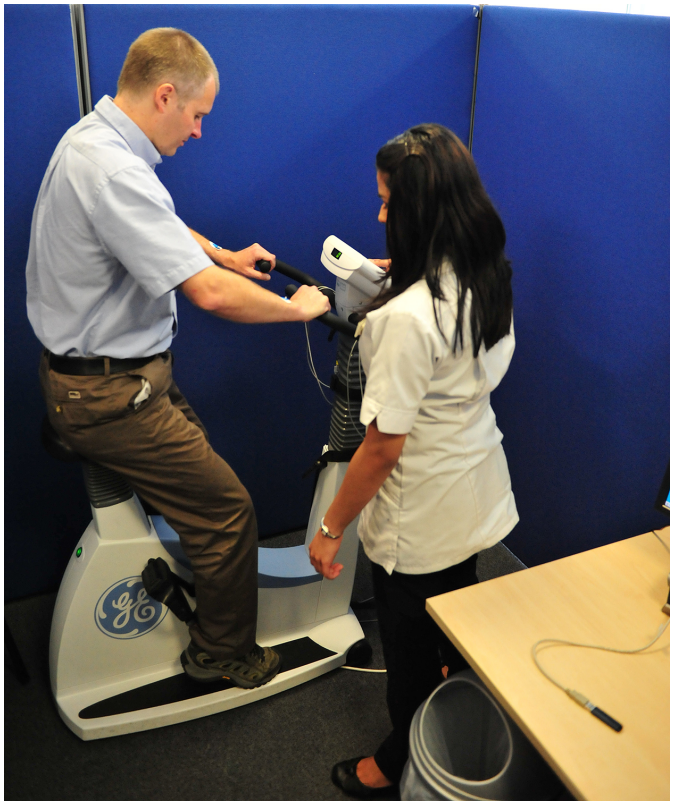


Figure 7.1: UK Biobank participants during the collection of the *ECG during exercise* dataset. Picture downloaded from the UK Biobank website

This dataset contains more than 75,000 subjects. Figures 7.2 and 7.3 show the data collected from 10 random participants. It can be seen that some participants had shorter sessions (due to their deteriorated physiological condition the exercise was ended before reaching the end of the protocol). It can also be seen that the slope is not the same for all participants: the workload was adjusted based on the participant age, gender, and weight.

I defined the following inclusion criteria:

1. the subject has completed the exercise protocol, has all 3 phases, and their duration is at least what's defined in the protocol;

2. in the exercise phase there is no HR less than 50 bpm (that would indicate a task incorrectly performed, or a faulty ECG sensor)

I therefore excluded participants with a total exercise duration shorter than the expected duration. This has the effect of removing subjects with known cardiovascular diagnoses, that I would exclude in any case from the health outcome prediction task. After applying the exclusion criteria the dataset had 47,917 subjects.

I resampled the data to have one datapoint for both HR and workload every 5 seconds using *Pandas resample* function of the *DataFrame* class. Each participant in my dataset has therefore two arrays of 88 values, one for HR and one for workload, for a total of 88x2 values.

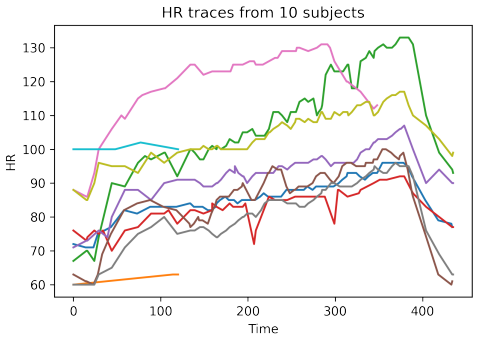


Figure 7.2: HR data for 10 random participants

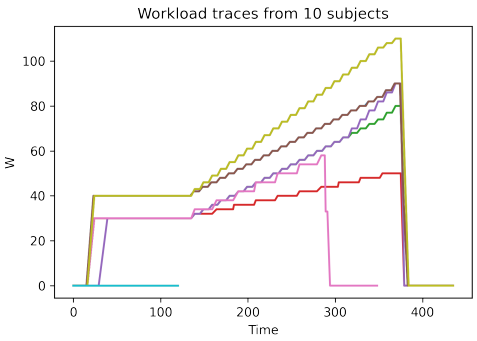


Figure 7.3: Workload data for 10 random participants

Field name	UK Biobank Field ID	Type	Preprocessing
Sex	31	Binary	
Alcohol Intake Frequency	1558	Categorical value	one hot encoded
Smoking Status	20116	Categorical value	one hot encoded
Current Tobacco Smoking	1239	Categorical value	one hot encoded
Ethnic Background	21000	Categorical value	one hot encoded
Body Mass Index	21001	Continuous value	normalised
Age At Recruitment	21022	Continuous value	normalised

Table 7.2: Fields used for the BASIC fields set. The table shows the field name, the UK Biobank field ID, the field type and the field pre-processing

For the basic information I used the fields shown in table 7.2. Continuous values (Body Mass Index and Age At Recruitment) have been normalised using min-max normalisation, missing values were imputed using mean imputation. Binary and Categorical values (Sex, Alcohol Intake Frequency, Current Tobacco Smoking, and Ethnic Background) were one hot encoded, missing data and non informative answers (i.e. *Prefer not to answer*) were assigned to a special category. For Ethnic Background we only considered the top-most categories: White, Mixed, Asian, Black, Chinese, Other, NA.

7.4 Electronic Health Records representation model

In this section I illustrate how I built the *EHR* model, that produces a fixed size representation (embedding) of the patients medical history. Models that can build embeddings of words can be used, such as *word2vec* can be applied to medical concepts [34] to learn a representation of each diagnostic, procedure or medication code that has a fixed size, i.e. a point in a continuous space, as shown in figure 7.4. These embeddings can be created in such a way that related concepts will be close in the embedding space (e.g. diagnoses for CVD and medications typically prescribed to CVD patients), and unrelated concepts far apart, as shown in figure 7.5.

However, I am interested in obtaining a representation of the whole medical history of each participant, not

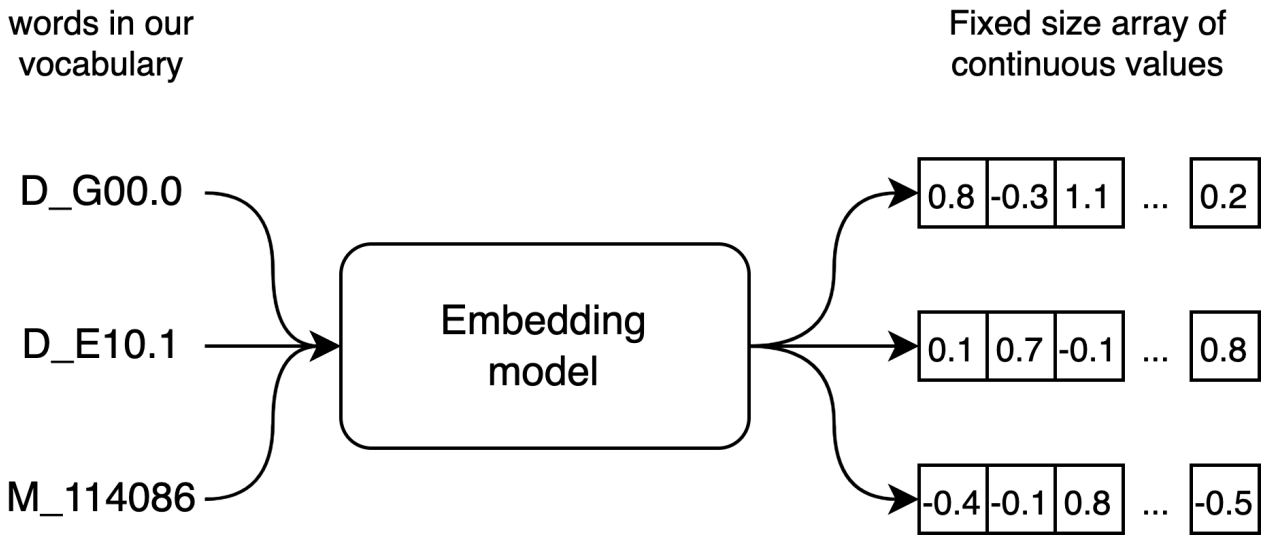


Figure 7.4: Diagram of word embedding: discrete words are mapped to points in a continuous space.

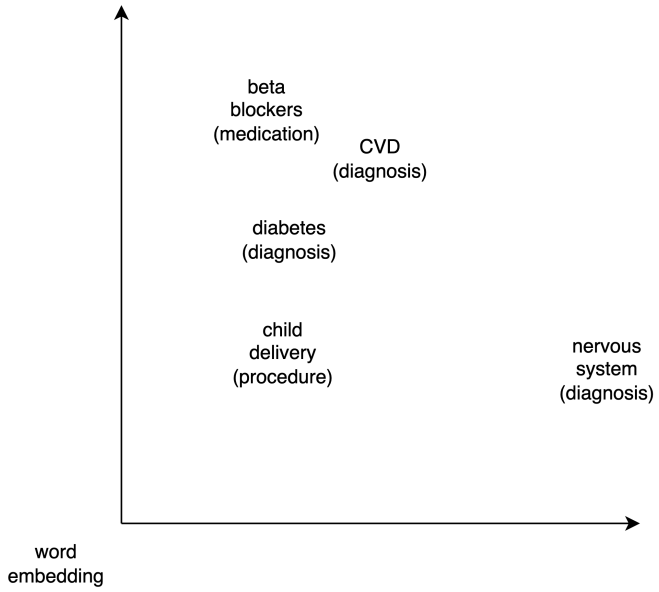


Figure 7.5: Example of how words should be mapped on the embedding space (here represented as a 2D space): medical concepts that refer to similar health status should be close together, unrelated concepts should be far from each other

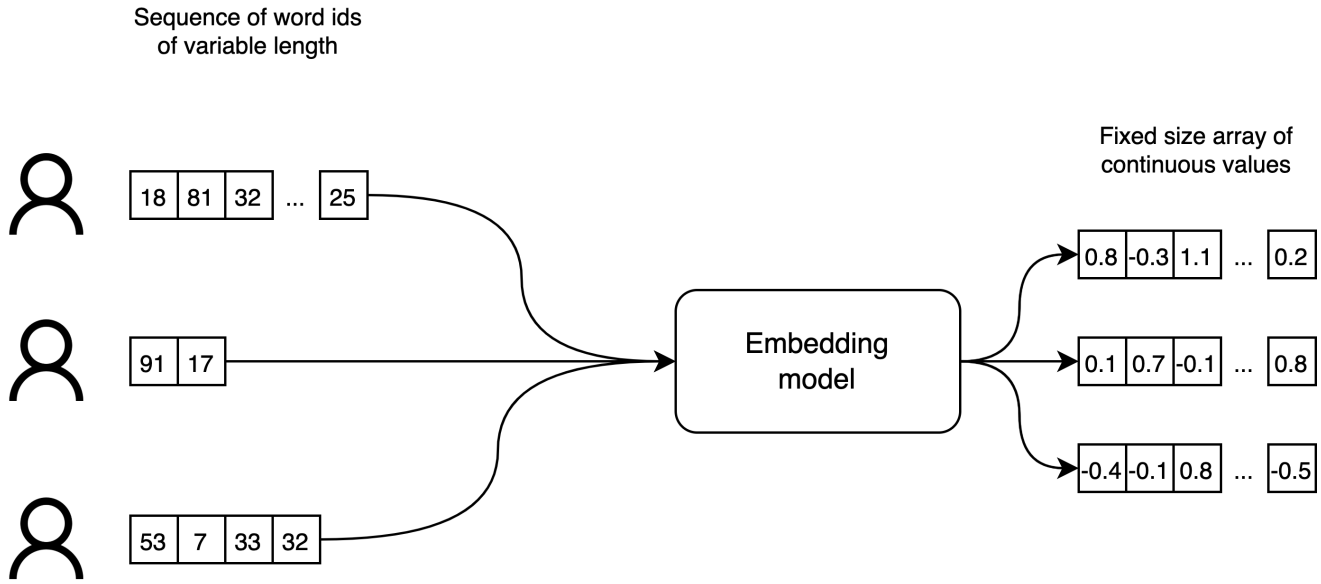


Figure 7.6: Diagram of EHR embedding: sequences of medical concepts of arbitrary length are mapped to points in a continuous space

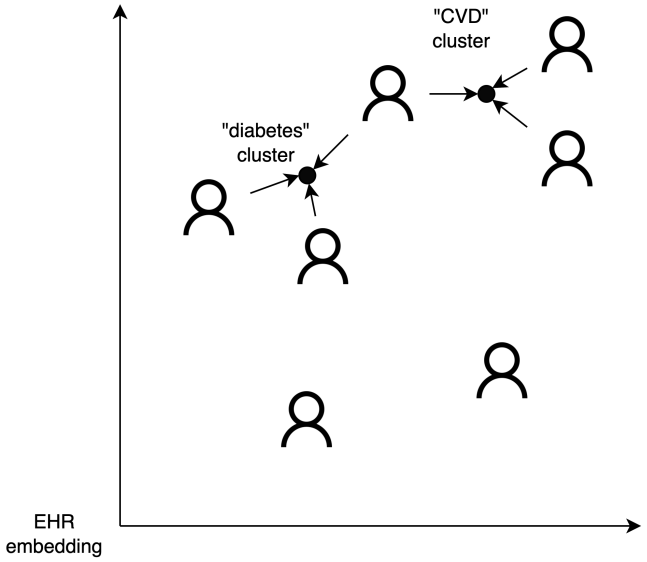


Figure 7.7: Example of how EHR should be mapped on the embedding space (here represented as a 2D space): patients with similar health conditions should be mapped to similar sub-regions of the embedding space, i.e. they should be close, patients with unrelated health conditions and medical history should be far from each other

of individual ICD10 codes. The challenge of this task is to create a model capable of ingesting a variable number of medical concepts (each subjects will have a different number of hospital visits, diagnoses, prescription, etc), but still produce representations of the same size, as shown in figure 7.6. Moreover, I want my model to produce representations that capture the semantic context of the health condition of the subjects, placing them in the same sub-region of the embedding space if they have similar health status 7.7. I expect this characteristic to constitute a good feature to be used in subsequent models, to build predictors of future health outcomes.

7.4.1 The choice of BERT as the patients representation model

As described in the previous section, my goal is to build a model that takes two paired lists of codes indicating medical concepts (diagnoses, procedures, and medications) and ages associated to the medical concepts, and outputs a dense vector of fixed size that represents the patients. The representation should convey information that is semantically significant. In particular participants with similar medical history should be close in the patient representation space.

Bidirectional Encoder Representations from Transformers (BERT) is a language models introduced in 2018 [51], that quickly became ubiquitous for many tasks. BERT has caused a stir in the Machine Learning community by presenting state-of-the-art results in a wide variety of NLP tasks, including Question Answering (SQuAD v1.1), Natural Language Inference (MNLI), and others [51]. BERT is an open source machine learning framework for natural language processing (NLP) that is designed to understand the meaning of ambiguous language in text by using surrounding text to establish context. The BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with question and answer datasets. BERT is based on the transformer architecture. Specifically, BERT is composed of Transformer encoder layers.

The Transformer architecture was introduced in the paper "Attention Is All You Need" [235]. It is used for natural language processing tasks such as machine translation, text summarisation, and question answering. As shown in figure 7.8, the Transformer architecture consists of two main components: the encoder and the decoder, and it does not rely on recurrence and convolutions to generate an output. The encoder maps an input sequence to a series of continuous representations. The decoder receives the encoder's output and the decoder's output at a previous time step, and generates an output sequence. The Transformer architecture has several elements including: Multi-head attention; Feed-forward neural network; Residual connections; Layer normalisation. These elements are used in both the encoder and decoder.

In the multi-head attention mechanism of the Transformer architecture, the input is split into multiple heads. Each head computes attention on its own set of *queries*, *keys*, and *values*. The *queries*, *keys*, and *values* are linear projections of the input. The attention mechanism computes a weighted sum of the values based on the similarity between the queries and keys. The output of each head is concatenated and linearly transformed to obtain the final output. The *queries*, *keys*, and *values* are used to compute the attention scores between the input and the output.

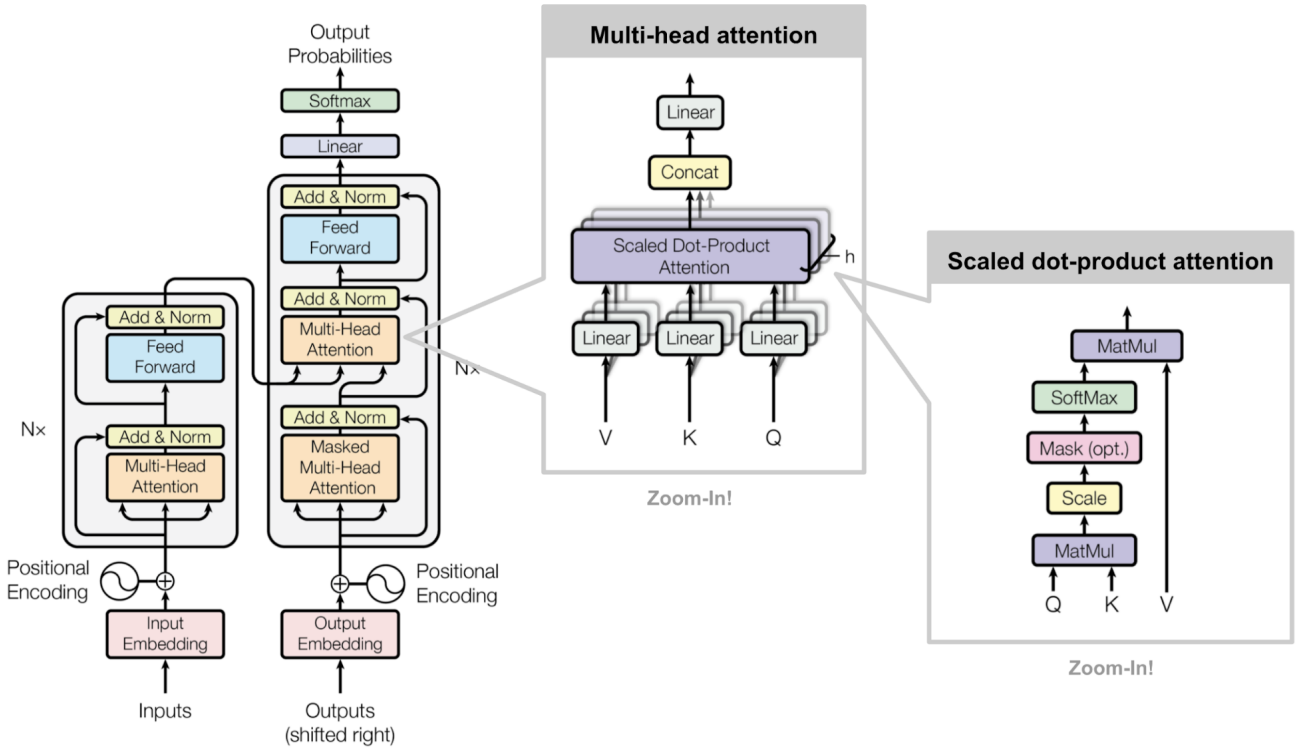


Figure 7.8: The full model architecture of the transformer, with the details of the multi head attention and the scaled dot product. (Image source: Fig 1 and 2 in Vaswani, et al., 2017 [235].)

The *queries* are used to compute the similarity between the input and the output. The *keys* are used to compute the similarity between the input and the output. The *values* are used to compute the weighted sum of the input based on the similarity between the input and the output. The scaled dot product of the multi head attention is used to compute the dot products of the query with all keys, divide each by the square root of the dimension of the key vectors, and apply a *softmax* function to obtain the weights on the values. The dot product attention is scaled by dividing it by the square root of the dimension of the key vectors to avoid large dot products that can cause numerical problems when computing gradients.

BERT converts each English word into an integer code. Its vocabulary has size 30,000. Any token not appearing in its vocabulary is replaced by [UNK] for "unknown". BERT was pre-trained simultaneously on two tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). These models are also used in practice to fine-tune text when doing natural language processing with BERT. In transformers, each output element is connected to each input component and the weightings between them are continuously determined based on their relationship.

BERT has been used to automatically assign ICD10 codes from free text [251]. Darabi et Al. apply BERT to the MIMIC-III dataset, including physiological data and clinician notes, that include both free text and codes, to obtain a unified representation of the patient [46]. BERT has been previously used to derive patients representations from diagnoses, medications and procedures codes [125, 184]. In my study I adopt a similar approach, but using a larger vocabulary: Li et Al. only use 301 codes for the diagnoses[125], I use a much larger vocabulary including

every ICD-10 code with more than 1,000 occurrences in UKB; moreover, they only consider the diagnoses, I include also procedures and medications. Shang et Al. [207] combine Graph Neural Networks and BERT, they use visit embedding instead of age like I do, they use a specific embedding for the ontology of ICD10 codes, I rely on the word embedding mechanism for the positioning of medical concepts in a meaningful way.

Choi et Al. [34] use *word2vec* to create embeddings of medical concepts. I need to obtain embeddings for whole EHR (that includes a varying number of medical concepts). Codes embedding can be summed or averaged [35], or other measures of distances over sequences can be applied [158] to obtain patients representation, but I believe that this approach will force simple metrics over a complex concept like *similarity* over *patients representations*, which I believe to be complex in nature, a complexity similar to semantic context in natural language processing, for which models based on BERT have proven to have superior performance. Alhasoun et Al. [6] use an approach based on Hierarchical Agglomerative Clustering to group ICD10 codes and studied the age distribution of groups. Other approaches included clustering of ICD10 codes [239] [188]. [236] used an even simpler approach based on non-hierarchical cluster analysis, comprising two separate steps: first, they used Multiple Correspondence Analysis (MCA); then K-means. The main interest was the definition of a comorbidity network built through the measure of relative risk between conditions. Before BERT, Stacked Denoising Autoencoders were used to obtain a representation of the patient [147], Recurrent Neural Networks based on Long-Short Term Memory (LSTM) [172] or GRU [33], focusing on prediction of disease rather than interpretability. I am interested in disease prediction but also in the creation of an explicit patient representation. However, recently proposed BERT based models proved to have better accuracy at downstream tasks such as health outcomes prediction, so I decided to use BERT for my patient representation model. I apply some of the ideas presented by Nguyen et Al. [159]: EMR are sequenced as sentences; ICD10 are the words; time gaps are handled as words (i.e. like a special ICD10 code) after discretisation, words then go through Embedding; infrequent words are coded as "RAREWORD". However, there are also noticeable differences: in their model a sentence is a single visit, and codes within a sentence are randomised. In my model the whole EHR is a single sentence, and I add words that indicate the beginning of a visit (with the indication if the admission was planned, or an emergency). They then use a CNN after embedding to find motifs (comorbidities, disease progression), with max pooling to get a single vector for the whole patient. Finally, they have a classifier (e.g. logistic regression). Instead, I build a dense network on top of the patient representation.

From Nguyen et Al. [159] I borrow two ideas:

- I add tokens to express time-gaps between entries in the EHR, to give the model an opportunity to explicitly use this token to create the patient representation: *1 week*; *1 month*; *1 month*; and *1 year*;

- I replace rare items, defined as occurring less than 30 times in all UK Biobank EHRs, with blanket tokens: *diag_RARE* for diagnoses; *op_RARE* for procedures; and *bnf_RARE* for medications.

7.4.2 Age embedding

In the seminal paper that proposed the transformer architecture [235] another key concept is proposed: positional embedding. In the original formulation the position of each word in the sentence was transformed into a vector of d dimensions by alternating sines and cosines at different frequencies, as shown in equations 7.1 and 7.2.

$$\mathrm{PE}_{\mathrm{pos},2i} = \sin(\mathrm{pos}/10000^{2i/d}) \tag{7.1}$$

$$\mathrm{PE}_{\mathrm{pos},2i+1} = \cos(\mathrm{pos}/10000^{2i/d}) \tag{7.2}$$

Most EHR models based on BERT do not use age to create embedding and only encode visit and positional [184, 207, 46]. However, research indicates that access to primary and secondary care is significantly influenced by factors beyond direct health concerns, such as area-level social deprivation. A study using the Rijnmond Primary Care Database [66] showed that children in socially deprived areas consult general practitioners more frequently for various symptoms, highlighting disparities in health-seeking behaviours influenced by socioeconomic status. This trend is particularly pronounced in metropolitan areas, where factors like age and socioeconomic background play a crucial role in healthcare access. These findings suggest that using hospital visits as health outcome predictors could be confounded by non-health-related factors like deprivation, behaviour, culture, and ethnicity, which impact care access.

Consequently, I opted to solely include age information corresponding to each medical code in the participants' EHRs, while omitting details about hospital visits. For encoding age, I employed the methodology outlined by Li et al.[125]. However, unlike their approach, which integrates the age data with visit embeddings and positional embeddings within visits, my analysis is confined to the use of age alone.

As shown in figure 7.9, the age of each medical concept in each patients' EHR is mapped to a vector of sines and cosines at different frequencies. Slower frequencies will express concepts like at what age, or in which year of the decade, and fast frequencies will express concepts like which day of the week. The input format of the age embedding module for each patient is a vector of n ages, and the output is a matrix $n \times d$ where d is the age embedding size. Like Vaswani et Al.[235] I sum the age embedding to the medical concepts (i.e. diagnoses, procedures, and medications), I use the same size for both embeddings.

The denominator of the embedding formula (that determines the range of frequencies captured by the age embedding, and their position across the embedding dimensions) is a hyper-parameter that I fit during the hyper-parameter search (called *TIME_ENCODING_DEN*).

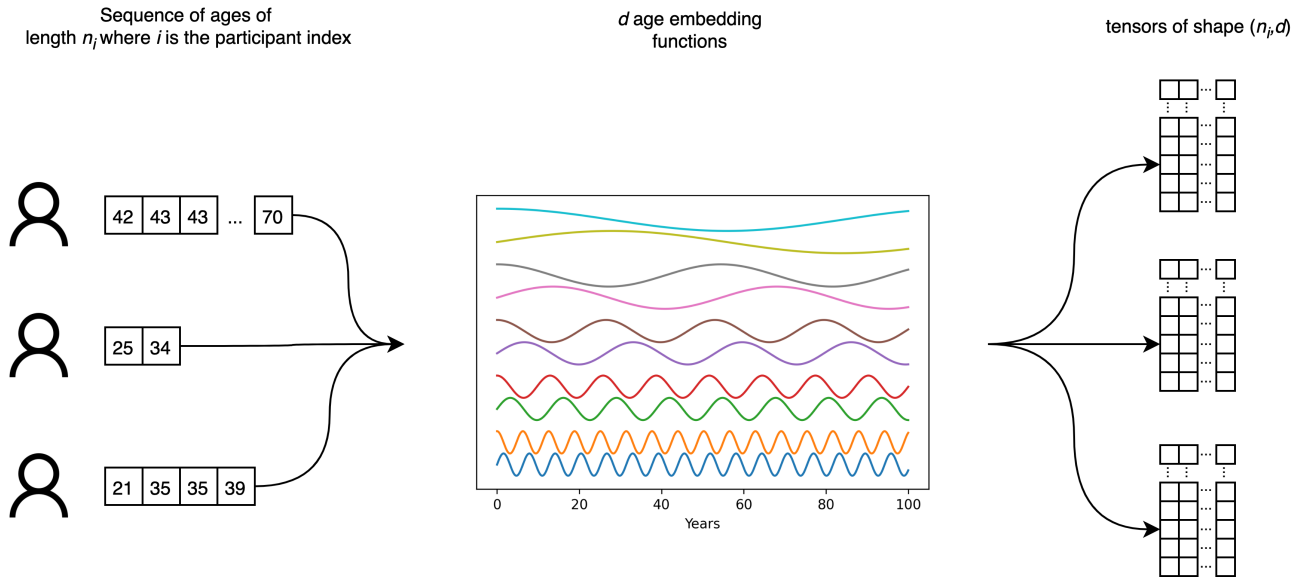


Figure 7.9: A diagram of age embedding: vectors of ages associated with medical concepts (of size n) are mapped to matrices (of size $d \times n$), where each age is mapped using sines (for even dimensions of the embedding) and cosines (for odd positions) at different frequencies

Country	GP Computer System	Clinical coding classification	Prescription coding classification
Scotland	EMIS / Vision	Read v2	Read v2 BNF
Wales	EMIS / Vision	Read v2	Read v2
	TPP	Clinical Terms Version 3 CTV3 or Read v3)	BNF
England	Vision	Read v2	Read v2 dm+d

Table 7.3: GP Systems and Clinical codings used

7.4.3 Description of the EHR data

EHR data in UK Biobank include primary and secondary care data collected from different regions in the UK, England Wales and Scotland. Different regions use different systems, including Albasoft [2]; EMIS Health [3]; Vision Health [4]; SAIL Databank [5]; TPP from SystmOne [6]; the Data Access Request Service (DARS) [7]; the Secure Anonymised Information Linkage (SAIL) Databank [8]; the electronic Data Research and Innovation Service (eDRIS) [9]. The EHR systems use different data formats, table 7.3 shows the formats of primary care data present in the dataset, and table 7.4 shows the formats used for secondary care data.

[2] http://www.albasoft.co.uk/
[3] https://www.emishealth.com/
[4] https://www.visionhealth.co.uk/
[5] https://saildatabank.com/)
[6] https://www.tpp-uk.com/)
[7] https://digital.nhs.uk/services/data-access-request-service-dars
[8] https://saildatabank.com/
[9] https://www.isdscotland.org/Products-and-Services/EDRIS/

Data provider	Dataset	Clinical Coding
England NHS Digital	Hospital Episode Statistics (HES)	ICD-10 OPCS-4 ICD-9
Scotland ISD	Scottish Morbidity Record (SMR)	ICD-10 OPCS-3 OPCS-4
Wales SAIL	Patient Episode Database for Wales (PEDW)	ICD-10 OPCS-4

Table 7.4: Hospital inpatient Systems and Clinical codings used

Medical Concept	Destination coding format
Diagnoses	ICD-10
Medications	BNF
Operations	OPCS-4

Table 7.5: Destination coding formats

7.4.4 Harmonisation of the EHR data in UK Biobank

As described in the previous paragraph different coding systems are present in the dataset. My goal is to build a model that understands the semantic relationships between medical concepts identified by codes. To avoid dealing with multiple codes indicating the same diagnoses (or medication, or procedure) I decided to harmonise the different coding systems. Table 7.5 reports the chosen data format for each type of medical concept. For Diagnoses, I map READ and ICD-9 codes to ICD-10 codes; for Operations, I map READ and OPCS-3 to OPCS-4 codes; for medications, I map READ to BNF codes. To map source to destination codes I used the tables provided by UK Biobank [10].

After the mapping of all formats of diagnoses to ICD10, all procedures to OPCS4, and all medications to BNF, I end up with a list of codes. Because ICD10 and OPCS4 have the same structure, to be able to distinguish them I prepend *diag_* to all diagnoses, *op_* to all procedures, and *BNF_* to all medications, and I merge all them in the same table. Because I believe that the number and type of hospital admissions (planned, or emergency) will provide useful information to the model about the health status of the subjects, I also add custom codes to the start of a hospital admission, to encode the following alternatives: *Elective admission*; *Emergency admission*; *Maternity admission*; *Psychiatric admission*; and *Transfer*. As I described in section 7.4.1 I also added tokens for time-gaps, and removed concepts that occurred less than 30 times and replaced them with *diag_RARE*; *op_RARE*; and *bnf_RARE* for medications.

I also calculate the age that the subjects had at the moment the medical concepts was added to their EHR. The

[10] `https://biobank.ndph.ox.ac.uk/ukb/coding.cgi?id=1834` to map Read2 to ICD10; `https://biobank.ndph.ox.ac.uk/ukb/coding.cgi?id=1835` to map Read3 to ICD10; `https://biobank.ndph.ox.ac.uk/ukb/coding.cgi?id=1836` to map ICD9 to ICD10; also mappings from READ to OPCS and from READ to BNF are available on the UK Biobank website

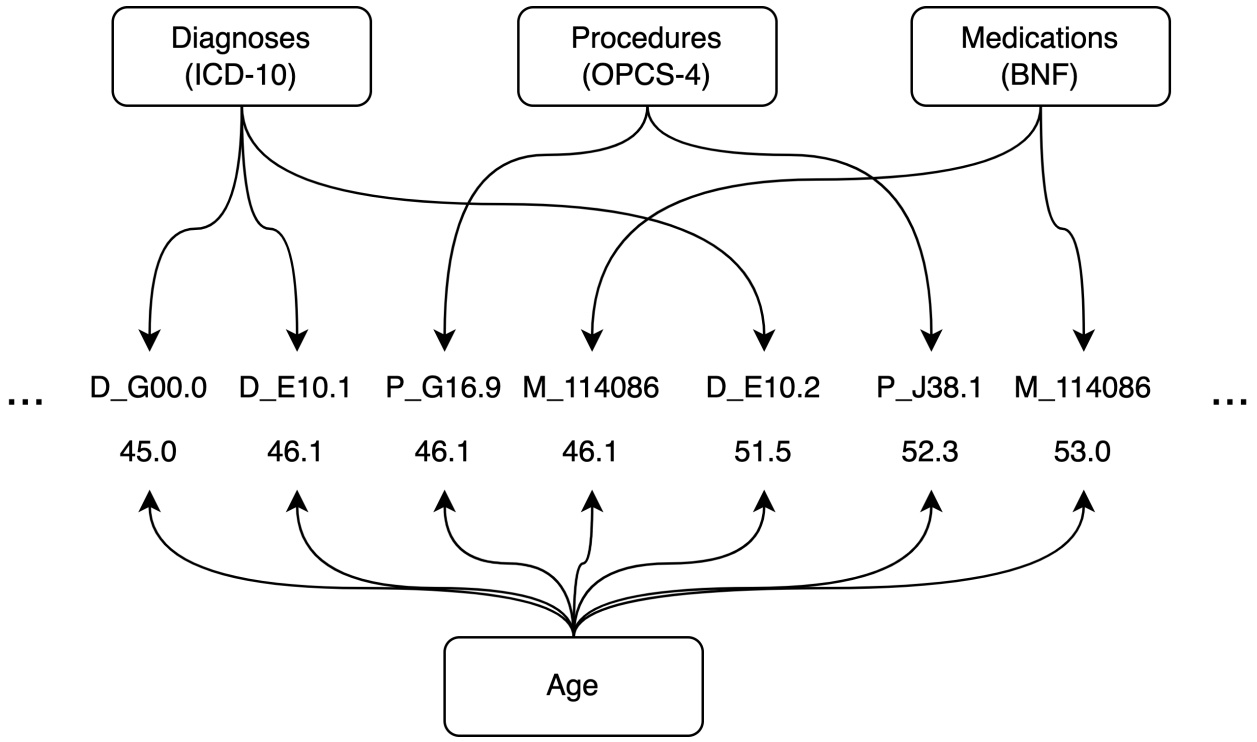


Figure 7.10: Medical concepts (diagnoses, procedures, and medications) and merged in a single sequence of tokens. Every token is also associated with the age of the patient at the moment the token was added to the patient's EHR.

final format of the EHR data is shown in figure 7.10: the EHR of each participant is expressed as two paired lists of codes and ages.

I have set the maximum sequence length to 256 tokens, I filled the data of participants with EHR shorter than 256 with the padding token (index 0), and applied a sliding window to the data of participants with EHR longer than 256, generating multiple sequences for each participants (each with a different starting point). To counterbalance for the over-exposure of the participants with long EHR (therefore present multiple times in the dataset, once per possible starting point with a length of 256) over participants with short EHR I also calculated a weight to be used in training to be applied to sequences, setting the weight to $1/n_i$ where n_i is the number of sequences generated for participant i. To train the EHR model I only used UKB subjects not included in the *ECG during exercise* dataset, to avoid using the same subjects for training and testing (in the downstream health prediction tasks). The final dataset contains 38,821,376 sequences of length 256.

To train the EHR model I excluded the EHR of the subjects in the test set defined for the HRV representation task, also used to test the downstream health prediction tasks. This ensures that the test data was never used in any stage of the training.

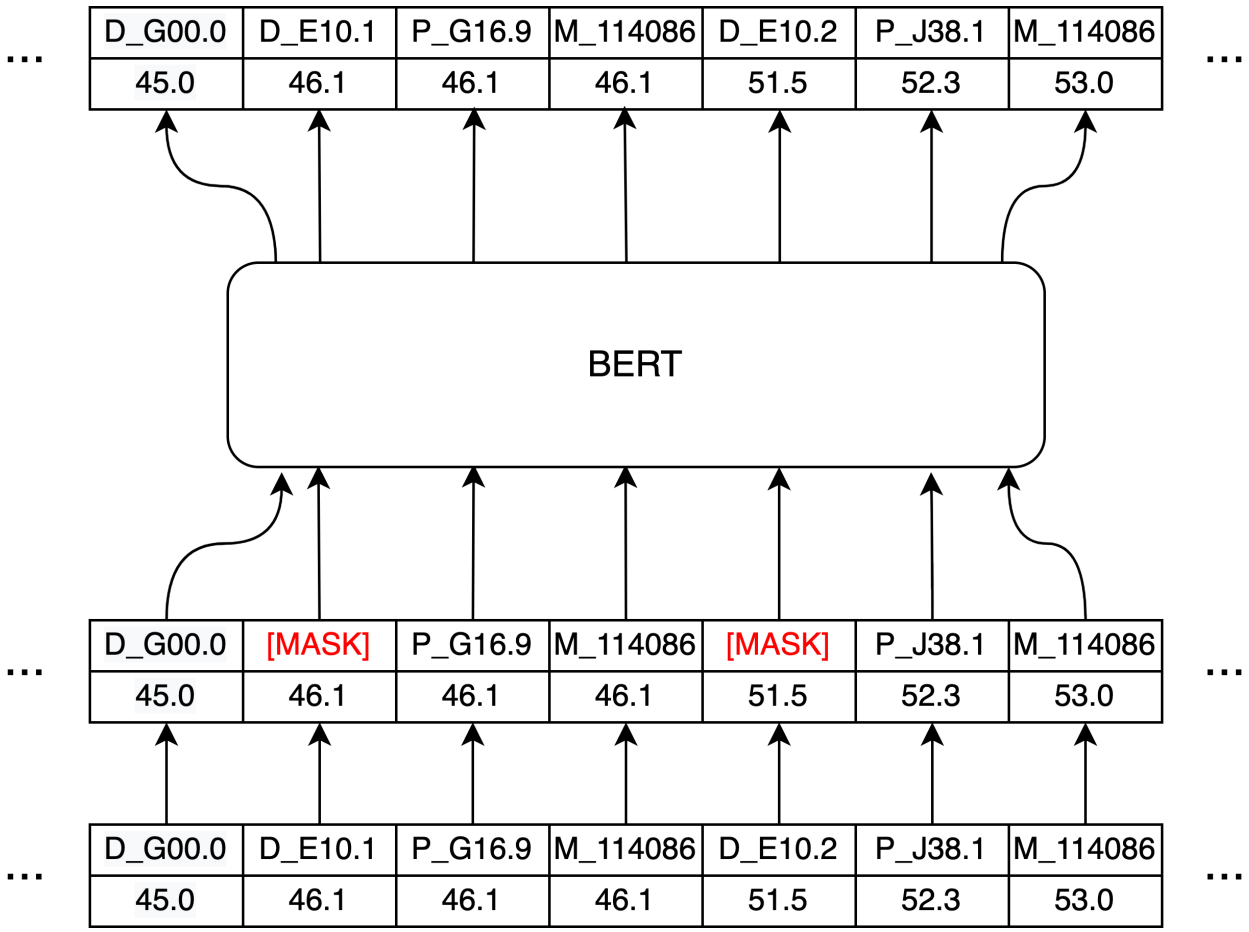


Figure 7.11: Diagram depicting the Masked Language Model used to train the BERT model: some codes were randomly masked and the model was asked to predict the masked codes.

7.4.5 Masked Language Model task: Hyper-parameter search and training

To train the BERT model I used the standard Masked Language Model task: I added a *[MASK]* token to the vocabulary; randomly selected 15% of the tokens and assigned 90% of them with *[MASK]*, and assigning the remaining 10% to random tokens. I trained the model to rebuild the original sequence (replacing *[MASK]* with the original tokens). A diagram of the Masked Language Model task is shown in figure 7.11. Hyper-parameter search was done on epochs of 131,072 sequences, allowing up to 1,024 epochs, patience 10 epochs on validation loss. I used the Bayesian Optimisation functionality of the python library *optuna*[3], allowing 60 trials (some failed because of too many parameters/data that didn't fit into the GPU memory), so had a total of 53 valid trials. Figure 7.12 shows the loss achieved throughout the attempted trials, where it can be appreciated that after an initial exploration of approximately 20 trials, the hyperparameter search could consistently achieve values close to the best trial, indicating that the search was able to find an area of the hyperparameter space where it was possible to achieve stable results, lowering the risk of overfitting.

Figures 7.13 and 7.14 show the loss achieved for the different values of each parameter in the search space. These graphs effectively illustrate the impact of various parameters on the loss. Notably, batch size appears to have an inversely proportional relationship with loss: larger batch sizes tend to result in lower loss. However, it was not possible to attempt a larger batch size than 2^8 given the limitations of the available GPU memory (a Tesla A100), as batch size values larger than 2^8 resulted in out of memory errors (see appendix D). The number of dimensions used for the embedding vector achieve minimum loss for a number of dimensions equal to $2^7 = 128$, a smaller embedding size consistently resulted in higher loss, probably because a smaller embedding vector was not able to express the necessary information, but larger embedding size also resulted in higher loss, probably because my dataset is not large enough to avoid overfitting using too large embedding spaces. The number of stacked BERT modules that achieved best results in 3, also in this case probably because a lower number of layers did not allow to model the semantic complexity of EHR (hence lower loss for less than 3 layers), but the dataset is not large enough to allow model deeper than 3 layers before overfitting. The number of heads used in BERT modules don't show a strong trend but seem to indicate the more heads usually help: a low number of heads can achieve comparable results to the best model (14 heads), but overall low losses were more easily achieved with a large number of heads (12 to 16), indicating that a large number of heads was helpful to capture different parallel semantic aspects of EHR. The dimension of the Feed-Forward layer, the Learning Rate, and the value used in the denominator for the age embedding formula don't show clear trends with loss, at least in this uni-dimensional representation. Many values, spread across the admissible range, resulted in values close to the minimum.

I then picked the best model and changed the data pipeline: I used the whole train dataset of 38,821,376 sequences for each epoch. Each epoch took 8 hours on my machine (see Appendix D for details of the system used). I allowed up to 16 epochs, with patience 1 on validation loss.

Validation was implemented with a holdout set 1% of the size of the whole dataset (approx 400,000 sequences).

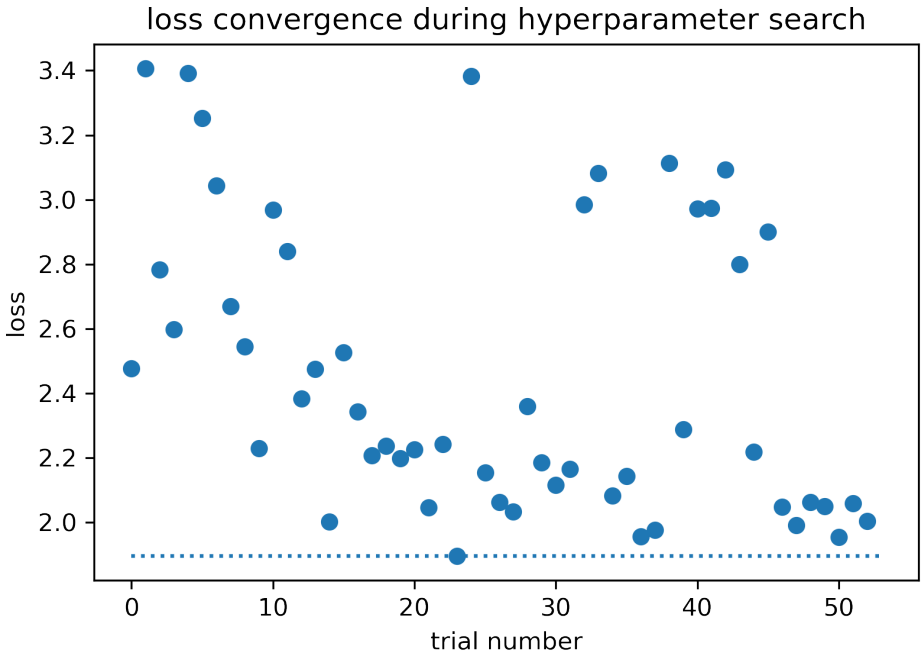


Figure 7.12: Loss of the trials during the hyper-parameter search. The dotted line shows the lowest loss.

The best model had a validation loss of 1.8953, with the parameters shown in table 7.6

param_name	param_value
BATCH_SIZE_exponent	8
EMBED_DIM_exponent	7
FF_DIM_exponent	7
LR (Learning Rate)	0.0006022
NUM_HEAD	14
NUM_LAYERS	3
TIME_ENCODING_DEN	3,411.0891

Table 7.6: Selected hyperparameters for EHR

In training I obtained loss: 1.6052 and validation loss: 1.6222.

7.5 HRV representation model

7.5.1 The choice of model

As discussed in section 1.1.3, there are very few standard metrics to assess the dynamics of the heart in response to physical exercise. Common metrics are:

Maximal heart rate This metric is the most commonly used in sport. It's known to change with age (a simple formula to estimate it is *subtract your age from 220* [221]);

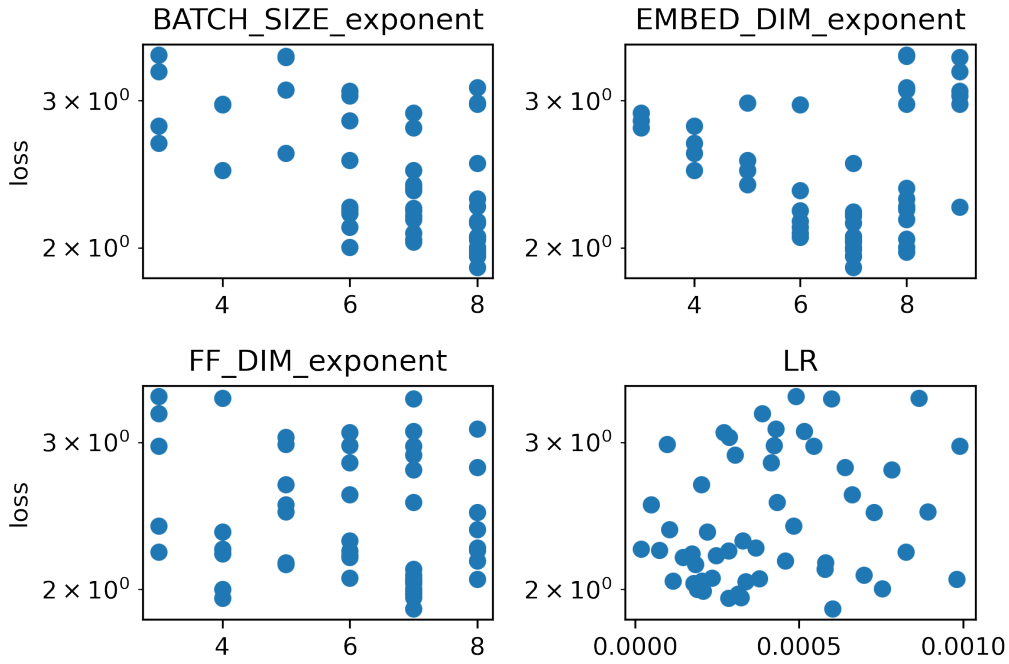


Figure 7.13: The loss obtained for each attempted value of the parameters. *BATCH_SIZE_exponent* is the exponent (base 2) of the batch size; *EMBED_DIM_exponent* is the exponent (base 2) of the size of the embedding vector used for the patient representation; *FF_DIM_exponent* is the exponent (base 2) of the Feed-Forward layer in the BERT modules; *LR* is the learning rate used for the Adam optimiser

.

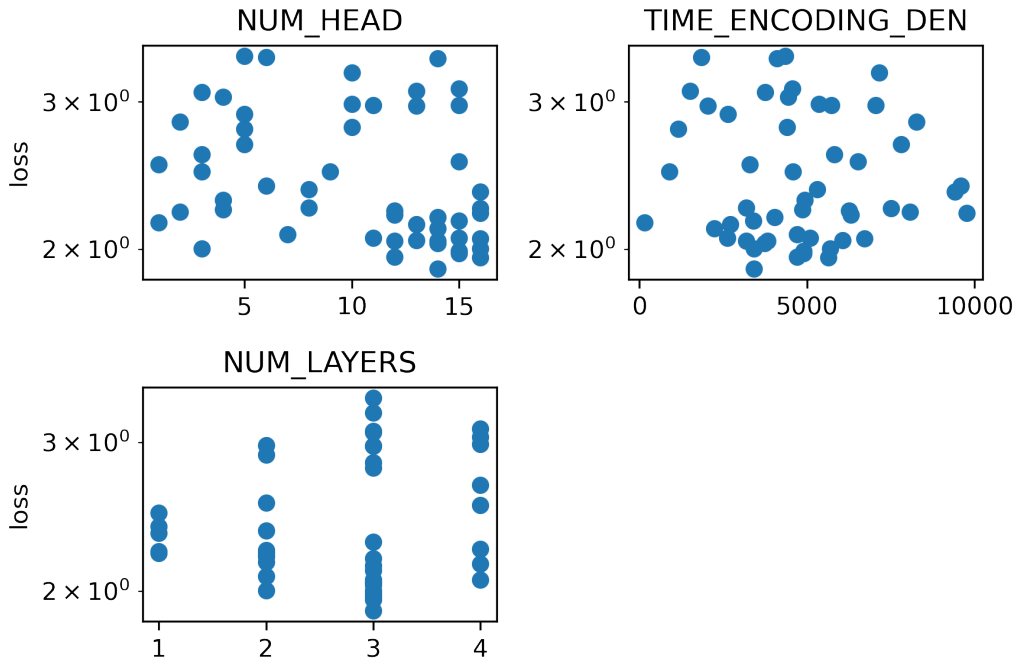


Figure 7.14: The loss obtained for each attempted value of the parameters. *NUM_HEAD* is the number of heads used in the BERT modules; *TIME_ENCODING_DEN* is value used for the denominator in the age embedding formula; *NUM_LAYERS* is the number of stacked BERT modules used

.

Heart Rate Recovery it's defined as *the difference between your heart rate at the end of your workout and one minute after* [210].

These two metrics are well accepted and commonly used, but it's evident that many other aspects of the heart response to physical exercise are not captured, such as how quickly the heart rate accelerates in response to a certain workload; how such measure changes as the workload increases; in what way the heart rate stabilises as the workload reaches a stable point and how it changes as the workload grows; etc.

For this reason I decided to create a model that would attempt to capture all the information in a compact representation. I decided then to build a Neural Network, an AutoEncoder (AE), that will attempt to rebuild the original input: the sequence of HR and work load resampled each every 5 seconds (described in section 7.3). The AE will have a bottleneck (the representation layer) that will force the model to learn a compact representation of the data. I set the size of the representation layer to 44, arbitrarily chosen to be able to design a convolutional neural network topology compatible (after several *max-pooling* layers) with the shape of the input data (a tensor of shape 88x2). For the encoder and decoder, I tested a Convolutional 2-Dimensional (CovAE) topology, using a Fully Connected (FC) topology as the baseline model. The AutoEncoder were trained using Mean Squared Error (MSE) as the loss function, minimising the distance between the reconstructed and the actual data.

I defined the holdout set (test set) as 10% of the ECG during exercise dataset. The test set was never used to train the HRV representation task, and was only used to estimate the performance of downstream health prediction tasks that use the EHR and HRV models. The train set (90% of the whole dataset) was then split in train and validation (10%) sets, and have been used for the hyper-parameter search. To train the best model found by hyper-parameter search the train a validation sets have been merged and used to fit the final models and have been tested on the test set.

7.5.2 Data processing

I used the dataset and pre-processing described in section 7.3 To normalise the data I built an Empirical Cumulative Distribution Function (ECDF) for HR data and an ECDF for workload data, and mapped each datapoint to its corresponding cumulative probability. Figure 7.15 shows original and ECDF normalised HR and workload of 5 random subjects.

7.5.3 Hyper-parameter search and training

To find the best model for a representation of the *HR during exercise* data I tested two different architectures, comparing the losses achieved on the test set:

- a Dense Neural Network: I tested the number of layers (1 to 3); the number of neurons for each layer (up to 128); the activation function (*selu*; *sigmoid*; *tanh*; and *relu*); and the dropout after each layer.

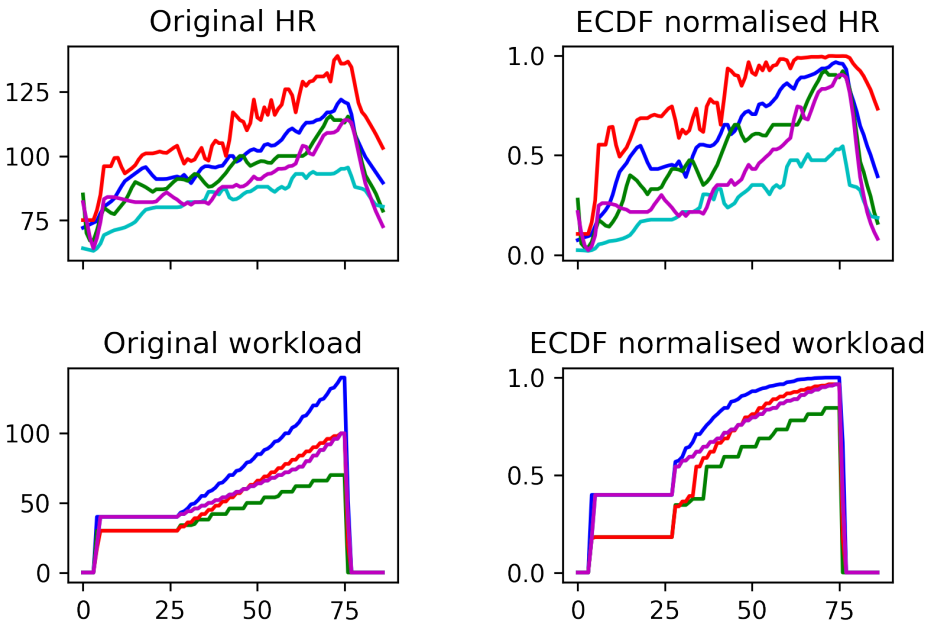


Figure 7.15: Original and ECDF normalised HR and workload of 5 random subjects

- a 2D Convolutional Neural Network: the dimension of the latent space was fixed to 44 units, but I tested different shapes ($11 \times 1 \times 4$; $22 \times 1 \times 2$, $44 \times 1 \times 1$). The number of convolutional layers was determined by the shape of the latent space: each convolutional layer halves the first dimension of the tensor, the input tensor is $88 \times 2 \times 1$, therefore to obtain a latent space e.g. of $11 \times 1 \times 4$ I needed 3 convolutional layers. I experimented with various numbers of filters for each convolutional layer and explored different options for the layer where the tensor's second dimension is reduced from 2 to 1. Consequently, the filter sizes became a hyper-parameter, optimised through the search process (the best performing combination of filters is reported in table 7.7). The kernel size was consistently set to (3, 3), with padding configured to *same*.

To run the search I used the training set for CVD, as later defined in section 7.6.1. Figure 7.16 shows the loss obtained by the trials during the hyper-parameter search. The search could consistently replicate the results of the best model, indicating that it was not a fortuitous result, but it was possible to find an area in the hyperparameter space where loss is minimised. Figures 7.17, 7.18, and 7.19 show the loss obtained by the trials for all the possible values of each parameter. The activation function that performed best was the Rectified Linear Unit (RELU), closely followed by Hyperbolic Tangent (TANH). The size of the batch size correlates with loss, with larger batch size resulting in higher losses. The convolutional topology achieved better results compared to the fully connected topology, and was therefore selected more often, as it can be noticed by the large amount of trials present in the graphs of the parameters of the convolutional networks (number of convolutional filter, layer where stride reduction was performed), compared to the small number of trials for the fully connected network (dropout probability,

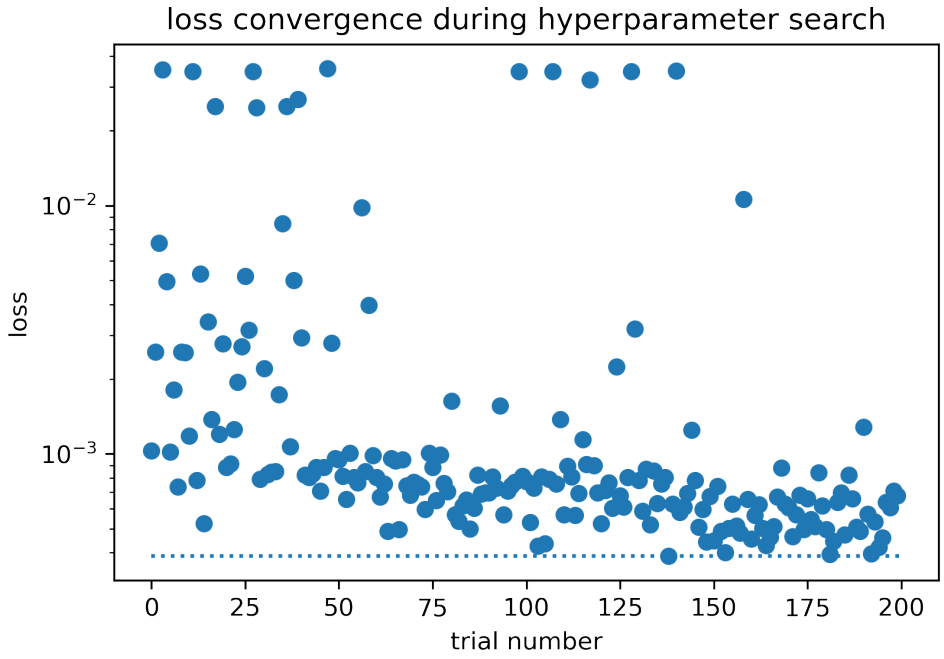


Figure 7.16: Loss (in logarithmic scale) of the trials attempted during the hyper-parameter optimisation. The dotted line shows the lowest loss obtained.

number of units in the 3 layers). This indicates that the hyperparameter search was confident of the superiority of the convolutional topology over the fully connected one. The learning rate did not show a clear best option, with low losses achieved by the upper half of the admissible range (with low losses for learning rates between 0.0005 and 0.001).

For the convolutional topology, the number of filters used in the convolutional layers show different behaviour: in the first layer the larger the number of filters the better the loss, with best results with 30, 31, and 32 filters; in the second layer I found a minimum with 17 filters. The best layer to perform the stride reduction was found to be the last available convolutional layer, and the optimal number of layers were found to be 3, that was the largest admitted value. This indicates that the apparent simplicity of the task still required a relatively deep architecture.

For the Fully Connected topology the dropout probability was clearly directly proportional to the loss (lower dropout probability resulted in lower loss), and the number of unites for each layer had minima at intermediate values. However, the number of trials attempted for this topology was too small to be able to obtain reliable graphs of the relationship between number of units and loss.

The search was conducted on a total of 200 trials The best trial was number 138, had loss 0.0003883 with the parameters shown in table 7.7.

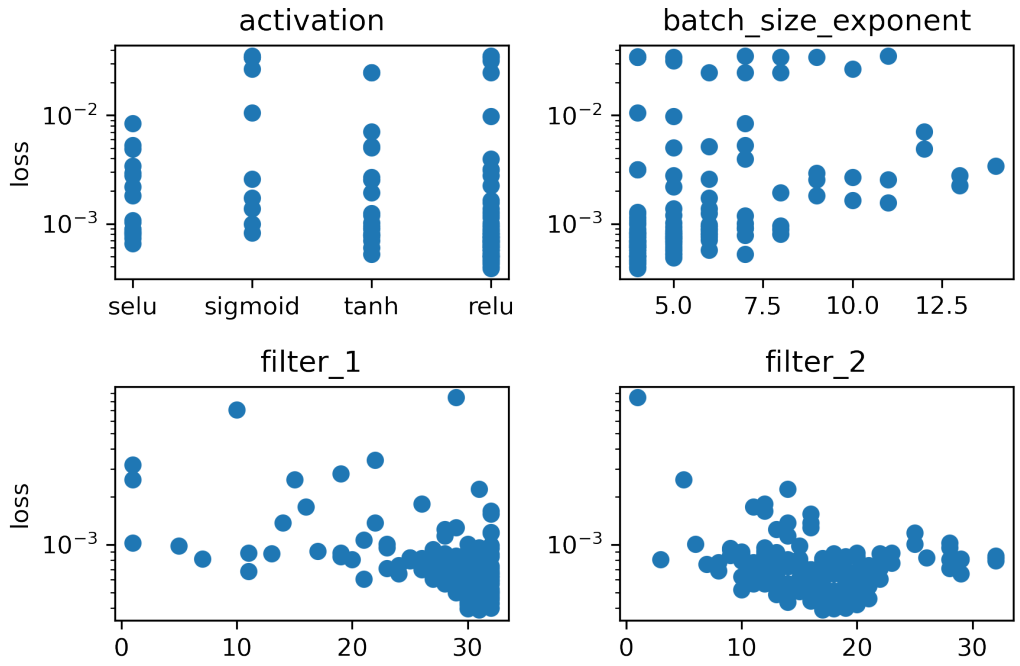


Figure 7.17: Loss (in logarithmic scale) for the different values attempted for each parameter during the hyper-parameter search. *Activation* is the activation function used in the neural network (except the last layer); *batch_size_exponent* is the exponent (base 2) used to define the batch size; *filter_1* is the number of filters used in the first convolutional layer of the network (only if using a *Conv* topology); *filter_2* is the number of filters used in the second convolutional layer of the network (only if using a *Conv* topology).

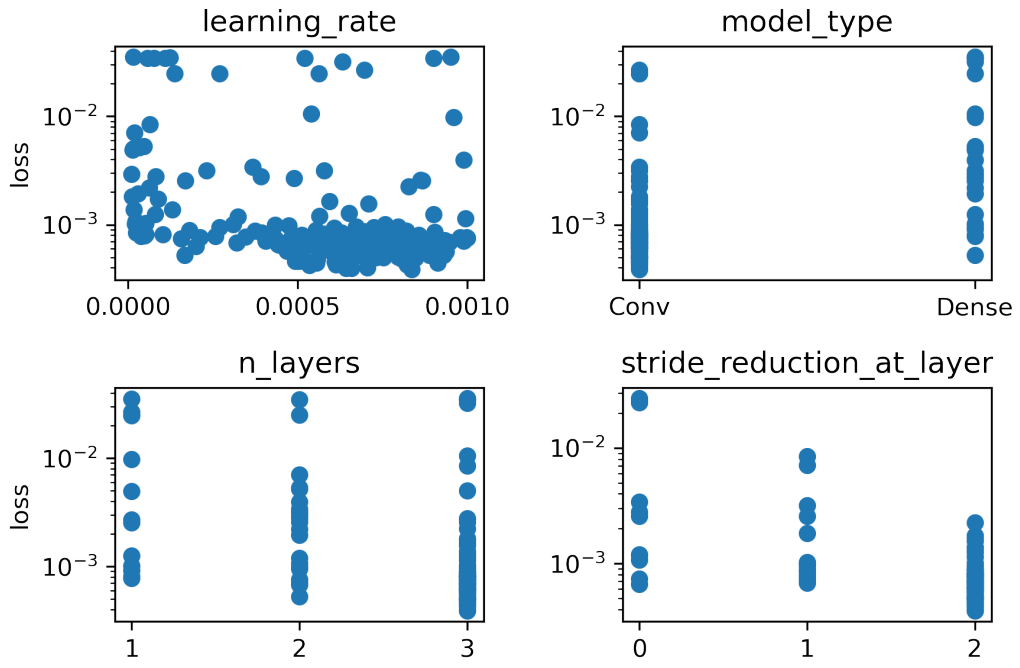


Figure 7.18: Loss (in logarithmic scale) for the different values attempted for each parameter during the hyper-parameter search. *learning_rate* is the learning rate assigned to the *Adam* optimiser; *model_type* is overall choice of topology; *n_layers* is the number of layers in the neural network, *stride_reduction_at_layer* indicates at which layer of the network the 2 inputs (HR and workload) are merged into 1 single dimension (only if using a *Conv* topology).

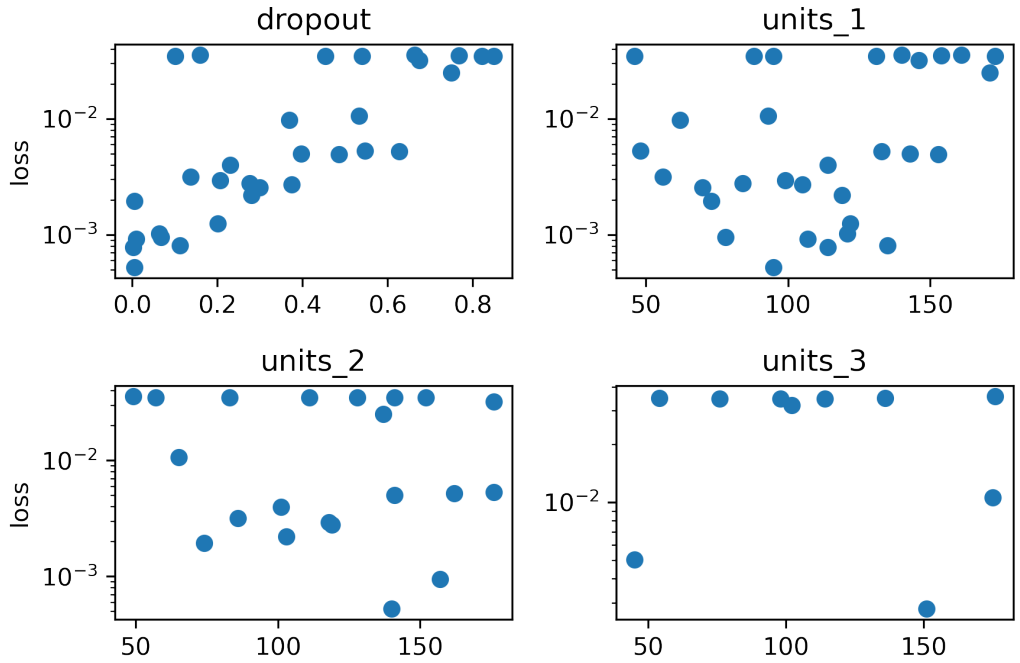


Figure 7.19: Loss (in logarithmic scale) for the different values attempted for each parameter during the hyperparameter search. *dropout* the dropout used after each layer in the *Dense* topology; *units_1* the number of neurons used for the first layer (only if using a *Dense* topology); *units_2* the number of neurons used for the second layer (only if using a *Dense* topology); *units_3* the number of neurons used for the third layer (only if using a *Dense* topology).

param_name	param_value
activation	relu
BATCH_SIZE_EXPONENT	4
FILTER_1	31
FILTER_2	17
Learning Rate	0.0008353
model_type	Conv
n_layers	3
stride_reduction_at_layer	2

Table 7.7: Selected hyperparameters for HRV

Figure 7.20 shows the topology of the best trial.

7.5.4 Quality of reconstructed HR data

Figures 7.21 and 7.22 show how the HRV representation model rebuilds the data of two subjects with different cardiovascular response to physical activity. The participant in figure 7.21 started from a resting heart rate in the lowest percentile and the maximum heart rate reached was just above 60th percentile even if the workload reached the highest percentile, and the heart rate recovery fully and quickly recovered once the exercise stopped, indicating

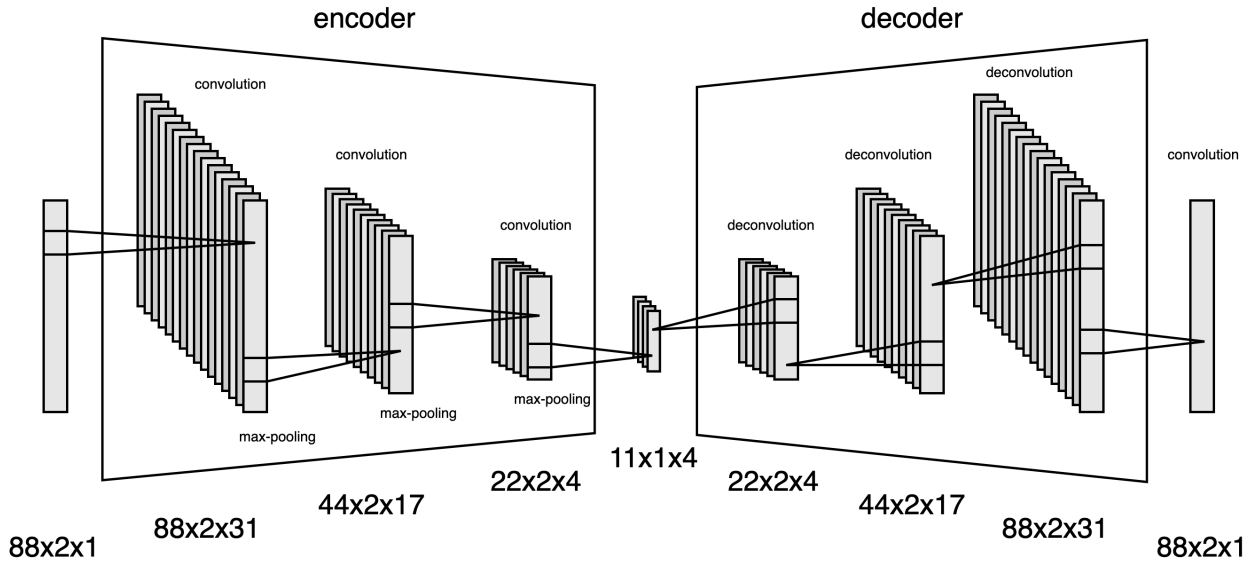


Figure 7.20: The model found by the hyper-parameter search for the *HR during exercise* representation

a fit cardiovascular system. The participant in figure 7.22 shows a very different response to exercise, starts from a much higher resting heart rate, and does not fully recover from the exercise, indicating a much less fit cardiovascular system.

It can be seen that in both participants, and both for HR and workload, the reconstructed data is very close to the original data. The reconstructed data seems to be smoothed, so the model acts as a low pass filter over the data.

7.6 Health Outcomes prediction from HRV

I then attempted to build models to predict the health outcomes discussed in section 7.2. My goal is to assess for which health outcomes HRV provides independent predictive information. To achieve this goal, for each heath target, I prepared 3 sets of inputs:

BASIC This data includes simple demographic, behavioural, and phenotypical data. Specifically, it includes:

Sex [11] Binary field, encoded as 0 or 1;

Alcohol Intake Frequency [12] Categorical value, one hot encoded;

Smoking Status [13] Categorical value, one hot encoded

Current Tobacco Smoking [14] Categorical value, one hot encoded

[11] UBK Field 31, see `https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=31`
[12] UBK Field 1558, see `https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=1558`
[13] UBK Field 20116, see `https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=20116`
[14] UBK Field 1239, see `https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=1239`

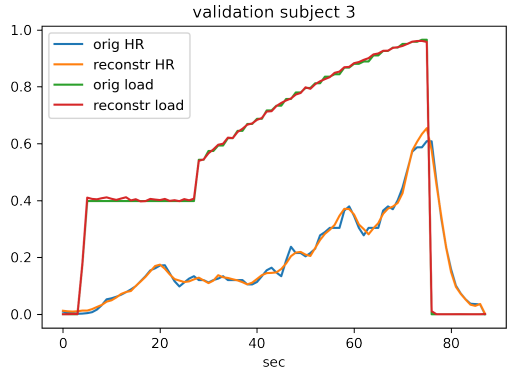


Figure 7.21: Example of HR data reconstructed by the Convolutional Auto-encoder. This subject from the validation dataset does not reach the top percentile for load but the HR goes close to the top percentile even at low loads, indicating a sub-optimal cardiovascular status.

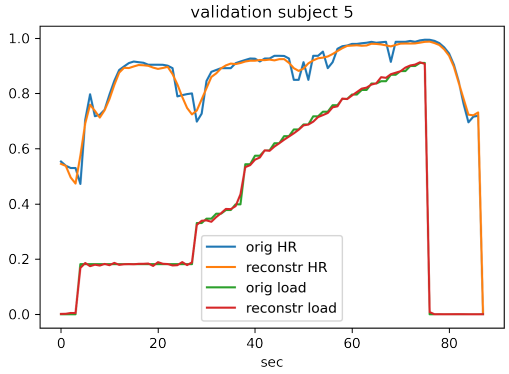


Figure 7.22: Example of HR data reconstructed by the Convolutional Auto-encoder. This subject from the validation dataset does not reach the top percentile for load but the HR goes close to the top percentile even at low loads, indicating a sub-optimal cardiovascular status.

Ethnic Background [15] Categorical value, one hot encoded

Body Mass Index [16] Continuous value, normalised

Age At Recruitment [17] Continuous value, normalised

EHR The patient representation from Electronic Health Records as described in section 7.4.4 [18]. This includes primary and secondary care date, diagnoses, medications, and operative codes, but I did not use lab tests and other measurements;

HRV The HRV representation from the *ECG during exercise* task [19], as described in section 7.5.2. This includes HR (beats per minute) and workload (W) collected during the exercise, as described in section 7.3.

I then trained seven separated models, one for each possible combination of inputs: BASIC; HRV; EHR; BASIC+HRV; BASIC+EHR; HRV+EHR; and BASIC+HRV+EHR. I then calculated the Area Under the Curve (AUC) for each model to study for which health outcomes HRV provides unique independent information that can not be explained via counfounders.

7.6.1 Validation strategy

For each target I applied the following processing:

1. I removed the subjects who had the target health outcome at assessment. For example when predicting CVD I removed all subject who were already diagnosed with CVD at assessment time;

[15] UBK Field 21000, see `https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=21000`
[16] UBK Field 21001, see `https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=21001`
[17] UBK Field 21022, see `https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=21022`
[18] UKB Category 3000, see `https://biobank.ndph.ox.ac.uk/ukb/label.cgi?id=3000`, and Category 2000, see `https://biobank.ndph.ox.ac.uk/ukb/label.cgi?id=2000`
[19] UKB Category 100012, see `https://biobank.ndph.ox.ac.uk/ukb/label.cgi?id=100012`

Target	Complete dataset	Training set	Test set
CVD	35,695 (1,270/34,425)	28,556 (1,016/27,540)	7,139 (254/6,885)
CHD	37,038 (785/36,253)	29,631 (628/29,003)	7,407 (157/7,250)
Angina	36,876 (775/36,101)	29,501 (620/28,881)	7,375 (155/7,220)
HBP	35,191 (286/34,905)		

Table 7.8: Total number of cases (with the detail of positives and negatives) in the train and test sets for all targets after inclusion and exclusion criteria, with a prediction window of 5 years

2. Since we are conducting a classification task (as opposed to survival analysis), I excluded subjects who either died or withdrew from the study before completing five years from the initial assessment. This exclusion is necessary because some subjects who withdrew might have developed the target health outcome after their departure. If included, they would be incorrectly assigned a label of 0 in our dataset, leading to potential inaccuracies;

3. I labelled all the subjects who were diagnosed with the target health outcomes within 5 years from assessment date (but only after assessment date) as a 1. All the other subjects were labelled as 0;

4. The resulting dataset was split in 5 folds, stratified on the target outcome;

5. if fold 4 contained less than 100 subjects labelled as 1, the target was excluded (the dataset after the exclusion criteria does not contain enough subjects with the target health outcome);

6. folds 0, 1 and 2 were used for the hyper-parameter search (each trial was performed with a 3-fold approach, using one fold for testing and the other two for training, then reported the average of the obtained AUCs);

7. once the hyper-parameters were obtained, folds 0, 1, and 2 were used as the training set and fold 3 was used as the validation set for early stopping (with patience 10 epocs on AUC of the validation set);

8. once the model was trained I calculated AUC on fold 4.

Table 7.8 shows the number of positive and negative cases in the train (folds 0, 1, 2, and 3) and test (fold 4) datasets after applying the exclusion criteria, for all targets, when setting a prediction window of 5 years. HBP (Any Hypertensive disease) did not have enough positive subjects after applying the exclusion criteria (we have set an arbitrary limit of 500 positive cases in the dataset, to ensure I have at least 100 cases in the test set), and has therefore been excluded from the experiments.

I then repeated this analysis for a shorter prediction window, set at 2 years. I expect to have a lower number of cases in the dataset because the criteria to include subjects in the positive case set is stricter, but possibly easier for the model to predict, because the health status of the subjects would be closer to the transition to the outcome of interest. Table 7.9 shows the number of cases when setting a prediction window of 2 years. It can be seen that

Target	After exclusion	Train set	Test set
CVD	35,146 (454/34,692)	28,117 (363/27,754)	7,029 (91/6,938)
CHD	36,853 (259/36,594)	29,483 (207/29,276)	7,370 (52/7,318)
Angina	36,741 (295/36,446)	29,393 (236/29,157)	7,348 (59/7,289)
HBP	33,654 (239/33,415)		

Table 7.9: Total number of cases (with the detail of positives and negatives) in the train and test sets for all targets after inclusion and exclusion criteria, with a prediction window of 2 years

Target	After exclusion	Train set	Test set
CVD	36,967 (3,188/33,779)	29,574 (2,550/27,024)	7,393 (638/6,755)
CHD	37,351 (1,918/35,433)	29,881 (1,534/28,347)	7,470 (384/7,086)
Angina	36,546 (1,306/35,240)	29,237 (1,045/28,192)	7,309 (261/7,048)
HBP	38,167 (355/37,812)		

Table 7.10: Total number of cases (with the detail of positives and negatives) in the train and test sets for all targets after inclusion and exclusion criteria, with a prediction window of 10 years

there never are enough cases. I have therefore excluded this prediction window from the analysis. The results of the analysis on a 2 years prediction window are reported in appendix C.

I repeated the analysis for a longer prediction window, 10 years. As expected, the number of positive cases increased compared to 5 years, because the inclusion criteria for the positive class become easier to be met. On the other hand, the inclusion criteria for the negative case became more strict: all subjects that dropped off the study (or died) before 10 years after assessment had to be removed from the study. In can be seen that there are less subjects overall, but more positive cases, for each target. Table 7.10 shows the number of cases with a prediction window of 10 years. In this case it can be seen that most targets have enough positive cases, so I have included this prediction window in the analysis.

7.6.2 Hyper parameter search and training

I opted for a simple fully connected topology, with *relu* activation in the hidden layers, and *sigmoid* activation in the last layer, binary cross-entropy as the loss function, maximum patience 5 epochs for early stopping. To find the optimal parameters for each of the seven models I ran an hyper-parameter search for each one of them, testing:

BATCH_SIZE_exponent the exponent (base 2) of the batch size. I tested values between 2 and 8, therefore with batch size between $2^2 = 4$ and $2^8 = 256$;

units_hidden_exponent the exponent (base 2) of the number of neurons in the hidden layer. I tested values between 1 and 7, therefore with a number of units between $2^1 = 2$ and $2^7 = 128$;

learning_rate the learning rate used with the Adam optimiser. I tested values between 0.00001 and 0.01;

layers the number of hidden layers. I tested values 1, 2, and 3. At each layer the number of neurons halves. The number of neurons at hidden layer 1 is set by the units_hidden_exponent parameter.

Given the heavy imbalance of the positive and negative cases in all outcomes in my dataset I used *class weight* in the fit *procedure*. I have set class weights to $w_c = N/n_c$ with n_c being the number of samples in class c ($c = 1$ if the subject had the health outcome of interest after assessment), N the total number of training samples, and w_c the resulting class weight.

I performed hyper-parameter search for all the seven combinations of inputs using Bayesian optimisation using the python package *optuna*. For each model (BASIC, HRV, EHR, HRV+EHR, BASIC+EHR, BASIC+HRV, BASIC+EHR+HRV) I allowed 50 trials or 3 days of running time. 3-fold cross validation was used, so each trial was performed 3 times on a different combination of train and test data, and the average AUC was returned as the trial value.

To reduce the hyper parameter optimisation time I decided to optimise only for CVD, trusting that the selected hyper parameters will also yield the best results for the other targets, that share similar class imbalance and characteristics with CVD. The training sets contain 3 of the 5 folds. For each model I used folds 0, 1, and 2 to run the hyper-parameter search. Folds 3 and 4 were left as validation and test sets. Each trial of the hyper-parameter search runs the selected configuration on 3 folds, using one of the fold as test set and the other folds as training set. The 3 obtained AUCs are averaged to compute the overall AUC of the trial. For each model I allow 50 trials, using an implementation of Bayesian Optimisation offered by the *optuna* python package [3].

The selected hyperparameters, the plots of the history of AUC of trials for each model, as well as the AUC obtained for each value of the parameters attempted in each trial, is available in appendix A.

7.6.3 Results for all targets

In this section I report AUC for all models on all targets, with prediction windows of 5 and 10 years (AUC for 2 years prediction window is available in appendix C). All the reported results have been obtained running the model on the test sets. The models were trained on the training set. See section 7.6.1 for details of the training and test sets for each target. Table 7.11 shows the AUC and the best performing models for all targets for 5 years prediction windows, and table 7.12 for 10 years prediction window.

Target	BASIC	HRV	EHR	BASIC+HRV	BASIC+EHR	HRV+EHR	BASIC+HRV+EHR
CVD	0.62	0.69	0.65	0.69	0.68	**0.71**	0.68
CHD	0.69	0.70	0.69	0.72	**0.74**	0.69	**0.74**
Angina	0.59	0.64	0.67	0.64	0.66	**0.73**	0.71

Table 7.11: AUC of all targets for the 5 years prediction windows. Best performing models are in bold.

Target	BASIC	HRV	EHR	BASIC+HRV	BASIC+EHR	HRV+EHR	BASIC+HRV+EHR
CVD	0.65	0.66	0.65	0.67	**0.68**	0.66	0.67
CHD	0.67	0.71	0.66	0.72	0.71	0.72	**0.73**
Angina	0.61	0.66	0.66	0.66	0.65	**0.70**	0.66

Table 7.12: AUC of all targets for the 10 years prediction windows. Best performing models are in bold.

As expected, BASIC is never the best performing model. For both the 5 years and the 10 years prediction windows, HRV, in combination with BASIC, EHR, or both BASIC and EHR, is almost always present in the best performing models. This indicates that HRV during exercise contains additional information that is valuable in combination with other data sources (BASIC, EHR). The ROC curves for all models and targets at both 5 and 10 years prediction windows are reported in appendix B.

CVD

HRV during exercise has previously been shown to be associated with CVD [52]. The results of my experiment confirm that heart dynamics during exercise provide valuable information to predict first occurrence of CVD. However, HRV is present (in combination with EHR) only in the best performing model for 5 years prediction window (AUC=0.71), but not for the 10 years window, potentially indicating that is useful only for short prediction window, and is not helpful when trying to predict CVD far in the future.

On a 10 years prediction window all models have similar performance, with BASIC still achieving the worst AUC (0.65) and BASIC+EHR achieving the best (0.68). Accuracy on the 10 years prediction window is lower than on 5 years. This is an expected result, because longer prediction windows inherently entail more uncertainty. However, it's important to notice that the dataset for the 10 years prediction window has more participants compared to the dataset for the 5 years prediction window. Therefore the performance of the models on the two prediction windows can not be used to draw conclusions on their relative predictive capability.

Chronic Ischemic heart disease

For CHD, for the 5 years prediction window, all the models achieve similar, relatively good results (between AUC 0.69 and 0.74). The models with the worst AUC are EHR and HRV+EHR (with AUC=0.69). The models with the best AUC are BASIC+EHR and BASIC+HRV+EHR (AUC=0.74).

HRV has slightly higher AUC than BASIC and EHR (0.69 vs 0.70), but the best AUC is achieved both with and without HRV, indicating that HRV is probably not needed if BASIC and EHR are available.

The models for the 10 years prediction window achieve similar results, and also in this case the best model is BASIC+HRV+EHR, with a AUC of 0.73, very close to the AUC achieved on the 5 years prediction window. In this case the models with HRV perform slightly better than models without: HRV alone has higher AUC than BASIC or EHR alone, and BASIC+HRV and HRV+EHR have higher AUC than BASIC+EHR. This might indicate that

for long prediction windows HRV becomes a valuable independent predictor of Ischemic heart disease.

Angina pectoris

For Angina, on the 5 years prediction window there is a wider range of accuracy values between models (AUC ranging from 0.59 to 0.73), the range shrinks in the 10 years prediction window (AUC ranging between 0.61 and 0.70). The results are are consistent across different prediction windows (5 and 10 years): the model that obtained the worst AUC is BASIC (0.59 in the 5 years prediction window, 0.61 in the 10 years prediction window), and the model that obtained the best AUC is HRV+EHR (0.73 in the 5 years prediction window, 0.70 in the 10 years prediction window).

For this target the EHR model performs better than HRV (AUC 0.67 vs 0.64) in the 5 years prediction window, and performs at the same level of accuracy in the 10 years prediction window. Combining EHR and HRV I obtain the best model (0.73 on 5 years, 0.70 on 10 years). Note that adding BASIC I do not increase accuracy, indeed BASIC seems to always worsen the AUC.

HRV in patients with angina pectoris changes significantly depending on the presence of coronary stenosis [143]. This finding suggests that models relying solely on HRV signatures may underperform due to the presence of diverse sub-populations with varying HRV characteristics. My experimental results confirm this observation: the HRV-only model achieves an AUC of 0.64, the EHR model achieves an AUC of 0.67, and the model incorporating both HRV and EHR data achieves an AUC of 0.73. Notably, the combined usage of EHR and HRV data confers an advantage, enabling the model to attain significantly higher precision compared to using HRV data alone.

7.6.4 Discussion

The Framingham Risk Score has exhibited variable accuracy, with the area under the curve (AUC) ranging from 0.50 to 0.83 (median 0.74) [231]. A recent investigation utilising the UK Biobank data revealed an AUC of 0.72 [5]. Comparatively, the BASIC and EHR models demonstrate lower performance, with AUCs of 0.62 and 0.65 for cardiovascular disease (CVD), respectively. For coronary heart disease (CHD), both models achieved an AUC of 0.69, while for angina, the AUCs were 0.59 and 0.67. An enhancement in AUC for CHD and angina was observed only when the models were combined; this did not extend to CVD. This discrepancy might be attributed to the broader definition of outcomes employed by Alaa et al. [5] as opposed to my approach, which involves subdividing the outcome into multiple targets to isolate the predictive efficacy of heart rate variability (HRV). Furthermore, the Framingham Risk Score includes measures that typically remain unknown to healthy individuals in the absence of health examinations and CVD screening. In contrast, the BASIC model was purposefully designed to integrate data readily accessible to the general public, such as information that could be gathered through a straightforward questionnaire distributed via a smartphone application.

The predictive capabilities of the BASIC, EHR, and HRV models relative to the state-of-the-art [5] were sat-

isfactory, with my model attaining an AUC of 0.74, marginally below the AUC of 0.76 reported by Alaa [5], who utilised 470 input variables. Nonetheless, it was anticipated that the EHR model would parallel the state-of-the-art performance, specifically reaching an AUC of at least 0.76. This can be caused by several factors, including:

- My EHR model does not include important data: I included diagnoses and medications, but I did not include lab tests. Important information such as cholesterol, triglycerides, and other blood tests are not included;
- The training tasks I designed for my EHR model were not effective in training the model to predict relationships over extended time periods. I employed the Masked Language Model training task, which teaches the network to learn relationships that are temporally proximal, as it reconstructs masked tokens by examining the surrounding tokens. To mitigate this issue, a different training task could be considered. A potential candidate is the Next Sentence Prediction task, suitably adapted for EHR data;
- EHR data actually predicts poorly most health outcomes.

I believe that all but the latter to be possible explanations. In the future I will include lab tests in the dataset, and I will design a different training tasks for the EHR model.

It is also important to mention that the results reported in this study use a relatively small dataset (see section 7.6.1). The AUC reported in the previous section might therefore be subject to instability due to small numbers of positive subjects in the test set. Further experiments should be run by using other datasets, and by using different validation strategies.

HRV models generally perform well. On the 5 years prediction window HRV is part of the best performing model for CVD (HRV+EHR); CHD (BASIC+HRV+EHR); and Angina (HRV+EHR). On the 10 years prediction window HRV is part of the best performing models for CHD (BASIC+HRV+EHR); and Angina (HRV+EHR). This was expected because all the chosen targets are health outcomes related to cardiopulmonary health (CVD; CHD; Angina), and HRV is known to be affected by the fitness of the a compromised cardiopulmonary system, so it was expected that HRV would prove to be a useful predictor.

Our experiments indicate that

$$p(\mathrm{O} \mid \mathrm{HRV}, \mathrm{EHR}, \mathrm{BASIC}) \neq p(\mathrm{O} \mid \mathrm{EHR}, \mathrm{BASIC}) \implies \mathrm{O} \not\perp\!\!\!\perp \mathrm{HRV} \mid \mathrm{EHR}, \mathrm{BASIC} \tag{7.3}$$

where O is the health outcomes I have considered. O is not conditionally independent of HRV, given EHR and BASIC. In other words, I can conclude that HRV provides additional information, independent of BASIC (age and behavioural), and EHR. HRV during exercise could therefore be used for health outcomes prediction in many cases.

This study shows that not only HRV at rest, but also HRV during exercise holds promise as a predictor for many health outcomes. Commercial grade fitness trackers collect both HR and physical activity data, so the findings of this study might suggest that such devices might be used to build screening models based on passively collected

data. In Celik et Al.[27] the most informative HRV feature to predict of CAD was SDNN24, that requires 24 hours of continuous ECG monitoring. I have shown in Chapter 6 how this valuable HRV feature can be estimated from consumer grade fitness trackers (as opposed to expensive and uncomfortable clinical grade Holter devices). This experiment shows the potential of passive HR tracking (during daily activities and during physical exercise) to predict CVD.

7.7 Conclusions

In this Chapter I have presented a study based on UK Biobank data, specifically on the *HRV during exercise* dataset, to analyse for which health outcomes HRV data collected during physical activity might provide useful predictive information. I have trained a model that can turn the participants EHR of arbitrary length into a patient representation of fixed size, using BERT. I have trained a model that turns HR and activity data during physical exercise into a dense representation. I then trained multiple models (BASIC, HRV, EHR, and all their combinations) for each health outcome. I performed hyper-parameter optimisation, training, and testing on separate datasets.

I then analysed for which outcomes HRV provided useful information, finding that HRV is almost always present in the model with highest accuracy, and seems to often provide useful additional information to EHR and BASIC data. This indicates that there is potential for HR during exercise to be used as a predictor for several health outcomes. Given that HR during exercise can be collected by commercial grade fitness trackers, this can enable valuable proactive tracking from passively collected data, without requiring expensive measurement devices, nor active involvement of the patients. This study illustrates that combining Electronic Health Record data with Heart Rate Variability metrics in response to physical activity bolsters the predictive accuracy for health outcomes. While HRV during exercise and EHR data have been independently examined in prior research, their concurrent utilisation as predictive variables allows for an analysis of their respective contributions. The findings support the concurrent use of EHR and HRV data to achieve a more comprehensive prediction model.

In the final chapter of this thesis, I delve into the limitations of the HRV and EHR models that have been developed. A particular limitation of the BASIC model is its exclusion of an area deprivation index, such as the Townsend Index, which is recognised for its predictive capability regarding health outcomes and risk factors. In future work, it will be essential to revisit the analysis and incorporate this index to enhance the model's predictive accuracy and comprehensiveness.

Chapter 8

Conclusions and future work

8.1 Limitations

The present thesis has made significant contributions to the understanding of various aspects of HRV, how it is affected by circadian rhythm, error propagation in wearable data, the effect of physical exercise, and the unique information provided by HRV in conjunction with Electronic Health Record when trying to predict a variety of health outcomes. However, several limitations exist within these areas of research, which are discussed below.

8.1.1 Error Propagation from PPG to HRV

The investigation of error propagation from PPG to heart rate variability in wearable data is a crucial aspect of this thesis. However, the study suffers from limitations due to the relatively small datasets used and the absence of longitudinal data. To comprehensively understand the evolution of error propagation over time, future research should incorporate larger datasets and longitudinal studies. These additions will enable a more comprehensive examination of the influence of various factors on the propagation of errors from PPG to HRV in wearable devices.

8.1.2 Circadian Model Validation

The circadian model proposed in this thesis has been validated on a dataset (MMASH) with a limited number of subjects and approximately 24 hours of duration. While the validation provides valuable insights into the model's performance, it lacks generalisability due to the small sample size and limited duration of recordings. To enhance the validity and robustness of the circadian model, future studies should include a larger number of subjects and longer recordings to assess how effectively the model captures changes in circadian parameters. Additionally, exploring non-parametric models beyond the outputs of single-component cosinor rhythmometry can further enhance the accuracy of the circadian model.

8.1.3 HRV Modelling during Exercise

The current model of HRV during exercise presented in this thesis focuses primarily on heart rate dynamics during and after maximal HR, activity achievable only during lab tests or professional training. However, a limitation arises from the lack of investigation into the effect of sub-maximal physical activity, which is more representative of everyday activity levels. To develop a more comprehensive understanding of HRV dynamics during exercise, future experiments should encompass sub-maximal physical activities that align with daily routines, providing insights into the impact of different intensities on HRV parameters.

8.1.4 EHR Model Design

The exploration of Electronic Health Record model design in this thesis offers valuable insights; however, there are opportunities for further exploration and improvement. Firstly, alternative ways of structuring EHR data, such as encoding visits in different formats, could be explored to determine the most effective approach. Additionally, the performance evaluation of the EHR model should extend to larger datasets to ensure scalability and generalisability. Furthermore, considering alternative models beyond BERT may enhance the performance and capabilities of the EHR model. Investigating and comparing the performance of different models will contribute to a more comprehensive understanding of EHR data processing and analysis.

In conclusion, the research presented in this thesis is subject to several limitations. Overcoming these limitations by including larger and more diverse datasets, longitudinal studies, investigating different activity levels, and exploring alternative modelling approaches will lead to more robust and reliable findings in these respective areas of study.

8.2 Future work

8.2.1 Physical activity and heart dynamics

The experiment that led to the creation of the *HRV during exercise dataset* involved physical activity following a controlled protocol, with fixed stages and duration, designed to bring the HR to maximal levels. I have seen that heart dynamics in response to physical activity offer valuable insight into the prediction of many health outcomes. I speculate that heart dynamics during sub-maximal physical activity, e.g. normal every day activity, walking to the train station to commute, might hold similar or analogous information, and could therefore be used to develop predictive models that only require everyday life, without asking users to follow specific instructions.

A series of experiments should be conducted to verify that the predictive capability of HR data collected during exercise holds if we remove two assumptions that are present in the *HRV during exercise dataset*:

- In *HRV during exercise dataset* HR reaches Maximal levels. However, in every day life most activities result

in sub-maximal HR, and relying on having reached maximal HR limits the applicability of my model. How quickly HR rises in response to a certain amount of physical activity, and how quickly HR returns to baseline levels after the physical activity has stopped will change as a function of activity intensity and duration. Models that describe heart rate dynamics in this setting should be developed.

- The *HRV during exercise* data I used in this study only consist in frequent HR and physical movement measurements during exercise. This kind of data is collected by most commercial-grade fitness trackers, and need no change to the way wearable devices normally operate. However, the dataset is collected in a controlled environment, a lab setting, with specific protocols and timing for each phase of the exercise. It would be valuable to collect data during daily activities of multiple subjects. In such data would be present moments where subjects perform mild physical activity, e.g. when they walk to the train station to commute, and the combination of HR and activity data should hold information similar to the information used in this study. There is therefore the potential to build a model that could be used to track users health from passively collected data, without needing them to follow any special protocol. Extending the HRV model I presented in this study to support long sequences of HR and activity data, would allow the creation of such proactive prognostic tool.

8.2.2 Probabilistic models

Patients similarity

The EHR model I have presented in section 7.4 outputs a dense representation of patients medical history: they are represented as points in the embedding space, a vector of fixed size (in my model 128, given by the EMBED_DIM_exponent hyper parameter). As discussed in section 7.4.1, having translated the patients medical history from a sequence of varying length of medical codes to points in the embedding space allows to measure similarity of patients, i.e. using cosine similarity.

Recent developments on Deep Learning methods for Density Estimation [165, 73] allow to estimate the distribution of the population on the embedding space. This might be used for example to sample participants in a Randomised Controlled trial ensuring that the Control and Intervention groups are truly stratified taking into account the whole medical history of participants (as opposed to only a limited set of criteria).

More generally, the patients embedding space combined with Deep Density Estimation methods allow the estimations of the probability of the data.

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \tag{8.1}$$

Equation 8.1 shows the well known Bayes theorem, where

- $p(\theta)$ is the prior, the probability of hypothesis before considering the evidence;

- $p(X|\theta)$ is called the likelihood, the probability of evidence when hypothesis is true;
- $p(\theta|X)$ is the posterior, the probability of the hypothesis considering the data;
- $p(X)$ is the evidence, the probability of the data.

Most models in ML are discriminative, i.e. they focus on finding the most likely value of θ given the data available. However, they usually don't calculate the full posterior, because they just select the value of θ that maximises $p(X|\theta)p(\theta)$. This happens because calculating $p(X)$ is in general intractable.

If estimating $p(X)$ were possible, it would enable the calculation of the full posterior of my model, rather than just the Maximum a Posteriori. This capability would facilitate the inclusion of numerous desirable features in my models. Among these benefits, I would gain a comprehensive probabilistic understanding of the hypothesis; I could easily combine probabilistic models by chaining them using Bayes' formula; and I would obtain an estimation of the uncertainty inherent in my model.

Causal analysis of HRV and outcomes

Equation 7.3 in section 7.6.4 indicates that HRV is an independent predictor of the health outcomes I have studied (health outcomes O are not independent of HRV, given EHR and BASIC).



Figure 8.1: Causal graph of BASIC, EHR, HRV, and O (outcomes). Graph 8.1a shows the observational graph; Graph 8.1b shows the do(O) interventional graph.

Figure 8.1 shows a probable causal graph of BASIC, EHR, HRV, and O (outcomes). The causal graph has the following assumptions:

- HRV does not directly affect any other node, but is affected by all other nodes;
- EHR captures potential comorbidity and general health status, that will affect HRV and the health outcome (e.g. having diabetes will have an influence on HRV, and changes the probability of being diagnosed with CVD);
- BASIC (age, and behavioural elements like diet) has a direct effect on every other node (e.g. age correlates with probability of developing diabetes, HRV changes with age independently of other comorbidities, and age correlates with CVD);

- O (the outcome) might have a direct effect on HRV (e.g. CVD causes changes in HRV).

Figure 8.1a shows the observational graph (the graph that generated the data as observed); Figure 8.1b shows the do(O) interventional graph where I try to quantify the direct effect of O on HRV. The do(O) interventional graph can be estimated by conditioning on BASIC and EHR, as shown in equation 8.2 [166].

$$p(\text{HRV}|\text{do}(\text{O})) = p(\text{HRV}|\text{O}, \text{BASIC}, \text{EHR}) \tag{8.2}$$

This potentially allows me to quantify the expected effect on each health outcome on HRV, normalised by comorbidities, sex, age, BMI, and all other data used in this study.

8.2.3 Health Outcomes survival analysis

To facilitate a comprehensive comparison of the outcomes of my current experiments with my prior research on Cardiovascular Disease (CVD) [55], it may be beneficial to expand the scope of the analysis delineated in the previous Chapter of the thesis. This expansion would involve transitioning to a survival analysis framework, which inherently supports right censoring in a time-to-event methodology. The primary objective within this extended analysis would be to estimate risk using a ranking mechanism, with the effectiveness of such estimation measured by the concordance index. This approach would not only enhance the depth of the comparative analysis but also enrich the methodological robustness of the study. In Appendix F, I thoroughly explore various strategies and methodologies that could potentially be adopted to seamlessly integrate our work into a survival analysis paradigm. This discussion is aimed at laying the groundwork for this methodological shift, detailing the conceptual underpinnings and the practical considerations that would inform such an extension. By doing so, the appendix serves as a blueprint for transitioning the current research findings into a more advanced analytical framework, thereby broadening the horizon of its applicability and relevance in the context of cardiovascular health research.

8.3 Contributions of this Thesis and Conclusions

In this thesis I have studied Heart Rate Variability and how passively collected PPG data could be used to monitor the health status of subjects and proactively predict many health outcomes. Wrist worn wearable devices equipped with PPG sensors (i.e. smartwatches) are widely adopted by the general population. This provides the opportunity for inexpensive and unobtrusive continuous passive monitoring.

In Chapter 2, I addressed a notable gap in the literature concerning the impact of errors in PPG data on Heart Rate Variability analysis, particularly focusing on noise affecting wearable devices. PPG data, as gathered from these devices, are prone to motion artefacts, which can significantly distort HRV measurements. An example of this is seen in the analysis of the SVI, a frequency-domain HRV metric commonly used to assess stress levels through

the LF:HF ratio. The investigation into the noise characteristics revealed that SVI's reliability is compromised, primarily due to the signal-to-noise ratio in the HF band falling below SNR = 1.0, a critical threshold for accuracy.

Moreover, the research demonstrated that while most HRV metrics can remain accurate when the user is stationary, even during non-stationary cardiovascular states such as post-exercise recovery, the same cannot be said under motion. This finding is significant in assessing the robustness of HRV features in various physiological conditions. To enhance the reliability of HRV data extracted from wearables, I employed accelerometry data to evaluate the quality of the PPG signals. This approach enables the potential for automatic filtering of unreliable PPG data, which could otherwise lead to erroneous HRV metrics, ensuring that only data with a high degree of reliability is used for analysis.

Throughout the entirety of HRV literature and practice, the standard approach to managing missing values has been to interpolate them within the duration domain of the inter-beat-interval (IBI) time-series. However, as explored in Chapter 3, this method introduces artificial signals into the original IBI time-series, potentially distorting the derived HRV features. Despite the critical nature of this issue, the impact of various interpolation methods on HRV analysis has not been comprehensively studied. To address this gap, I conducted an in-depth investigation into the error propagation caused by missing values in IBI time-series and their effect on HRV metrics.

Using a Burst model to simulate missing data, my research examined how errors in HRV features escalate with increasing amounts of missing data. A key finding of this study is that the optimal strategy for pre-processing involves working with the timestamp time-series, rather than the duration time-series. This conclusion may appear intuitive, yet it stands in stark contrast to the prevailing practices and literature in HRV pre-processing. By shifting the focus to the time domain for interpolation, this research challenges established methodologies and offers a novel perspective on handling missing values in HRV data analysis.

Wearable devices need to ensure battery life is as long as possible to allow users to wear the device without having to recharge it often. The PPG sensor is therefore often used parsimoniously to avoid wasting battery with leds, sensors, and computation. Users expect wearable to collect data throughout the day, so the usual compromise is to collect heart activity frequently, but for short amount of time. HRV analysis usually requires 5 minutes of continuously collected RR intervals. Most wearable devices focus on data collection windows of 1 minute or less, called ultra-short measurements. In Chapter 4 I have studied which HRV features can be reliably estimated from ultra-short measurement windows. I conclude that it is possible to accurately estimate RMSSD and SDNN by using 30 seconds and 1 minute time windows, respectively. I find that SDNN can not be reliably estimated from measures shorter than 1 minute, because such short measurement windows do not allow for good estimate of heart activity in the VLF band, inducing a bias in SDNN that, as discussed in section 1.1.2 is mainly affected by low frequencies in human HRV. This finding is of particular interest, given that SDNN is the main HRV feature reported by commercial grade wearable devices, e.g. the Apple watch. To the best of my knowledge, Chapter 4 presents the first study of the effect of missing values on ultra-short HRV data.

Chapter 5 introduces an innovative mathematical model tailored to analyze the influence of circadian rhythms on resting heart rate, addressing a crucial limitation in existing literature. Traditional circadian analysis models either prioritise simplicity of interpretation, as seen in the cosinor model, at the expense of fitting the data accurately, or they employ complex machine learning models that sacrifice the explainability of resulting parameters. Our model bridges this gap, offering both the expressiveness of a multi-component cosinor for precise data fitting and the straightforward simplicity of interpretation of a single-component cosinor, with just three easily understandable parameters: *MESOR*, *Amplitude*, and *Acrophase*. This model is designed to support HR data collected from wearable devices, because it supports unevenly sampled data (wearable devices will not always sample heart activity at regular intervals), and allows to assign weights to each HR data point individually (e.g. data collected when mild movement is detected, therefore possibly corrupted by motion artefacts, could be assigned lower weight than data collected at rest). The model is designed to be computationally cheap, with a complexity of $O(n)$ where n is the number of HR measurements. The model can be updated with a streaming approach: when updating the model with new data, the old data is not used, and the surrogates are used in combination with new data, allowing for inexpensive computation. The model allows assessing the evolution over time of parameters of circadian rhythm (i.e., *MESOR*, *Amplitude* and *Acrophase*), making it possible to detect changes in heart activity from passively collected data from wearable devices.

In Chapter 6, I have illustrated that SDNN24, a crucial metric for heart rate variability, can be effectively estimated from heart rate data collected by wrist-worn wearable devices. This finding diverges from the traditional reliance on continuous 24-hour electrocardiogram (ECG) data, commonly used in HRV analysis. Before this work, it was not widely recognised in HRV literature that accurate SDNN24 estimation could be achieved from semi-continuous heart rate readings. The research underscores that heart rate data from fitness trackers, despite not being continuous, retains valuable information about the ultra-low frequency (ULF) components of the HRV spectrum. These components are significant as they are indicative of the activity of the Sinoatrial Node (SAN), the heart's primary pacemaker. Compromised SAN activity is usually evidenced by a decrease in power in the ULF region, which correlates with reduced SDNN values. By showing that SDNN24 can be estimated from fitness tracker data, this work provides a new perspective on heart health monitoring. It suggests a potential for broader, more accessible screening for cardiovascular diseases (CVD) using widely available wearable devices, making continuous tracking of heart health less obtrusive and more feasible on a larger scale.

Finally, in section 7, I have discussed how heart rate data collected during exercise dataset holds valuable independent information that can be used, together with demographic, behavioural, and medical (EHR) data, to predict many health outcomes. Compared to HRV data at rest, HRV data during exercise has seldom been explored, analysed, and used to predict health outcomes. Wearable devices are normally worn during the day, and often recharged during the night. Therefore, HRV is seldom collected during the night, when the absence of movement would ensure HRV is free of motion artefacts. Moreover, is physiological plausible to expect that heart

dynamics in response to physical activity holds valuable insight on cardiopulmonary fitness and health. Yet, even though traditional HRV features developed for HRV at rest can not be applied to HRV during exercise, a very reduced number of features have been proposed specifically to HRV during exercise. I presented a non-parametric model that can learn a representation of HRV during exercise using an Auto-Encoder architecture. I have also shown a large language model (BERT) to be able to represent the full medical history of subjects in an embedding space. I then used both models to predict several health outcomes, and I concluded that HRV during exercise is a useful predictor for most outcomes. This indicates that HRV during exercise should be considered as a potential target for the development of digital biomarkers. Since most fitness trackers can be used to continuously record this data during physical activity, there is the potential for low cost unobtrusive monitoring for many outcomes, without requiring uncomfortable and expensive clinical equipment, but only commercial grade fitness trackers.

This study represents a novel approach by combining Electronic Health Record data with Heart Rate Variability metrics during physical activity, a methodology not previously explored in medical research. The analysis indicates that HRV consistently plays a significant role in the most accurate models, often enhancing the information provided by EHR and BASIC data. This finding underscores the potential of HRV, particularly during exercise, as a valuable predictor for a range of health outcomes. The integration of EHR and HRV marks a significant advancement in predictive modelling, suggesting a more comprehensive approach to predicting health outcomes than has been achieved with either data type alone.

Bibliography

[1] Agelink, M., Majewski, T., Wurthmann, C., Postert, T., Linka, T., and Rotterdam, S. Autonomic neurocardiac function in patients with major depression and effects of antidepressive treatment with nefazodone. *Journal of Affective Disorders 62* (2001), 187–198.

[2] Agliari, E., Barra, A., Barra, O. A., Fachechi, A., Franceschi Vento, L., and Moretti, L. Detecting cardiac pathologies via machine learning on heart-rate variability time series and related markers. *Scientific Reports 10*, 1 (2020), 1–18.

[3] Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (2019), pp. 2623–2631.

[4] Aktypi, A., Nurse, J. R., and Goldsmith, M. Unwinding Ariadne's identity thread: Privacy risks with fitness trackers and online social networks. In *Proceedings of the 2017 on Multimedia Privacy and Security* (2017), ACM, pp. 1–11.

[5] Alaa, A. M., Bolton, T., Angelantonio, E. D., Rudd, J. H. F., and Schaar, M. v. d. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PloS one 14*, 5 (May 2019), e0213653.

[6] Alhasoun, F., Aleissa, F., Alhazzani, M., Moyano, L. G., Pinhanez, C., and González, M. C. Age density patterns in patients medical conditions: A clustering approach. *PLOS Computational Biology 14*, 6 (2018), e1006115.

[7] Allen, J. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement 28*, 3 (Feb. 2007), R1.

[8] Bacciu, D., Morelli, D., and Pandelea, V. Modeling mood polarity and declaration occurrence by neural temporal point processes. *IEEE Transactions on Neural Networks and Learning Systems 34*, 4 (2022), 1800–1807.

[9] Baek, H., and Cho, J. Novel heart rate variability index for wrist-worn wearable devices subject to motion artifacts that complicate measurement of the continuous pulse interval. *Physiological Measurement 40* (2019).

[10] Beak, H., Cho, C., and Cho, J adn Woo, J. Reliability of Ultra-Short-Term Analysis as a Surrogate of Standard 5-Min Analysis of Heart Rate Variability. *Telemedicine and e-Health 21(5)* (2015), 404–414.

[11] Benichou, T., Pereira, B., Mermillod, M., Tauveron, I., Pfabigan, D., Maqdasy, S., and Dutheil, F. Heart rate variability in type 2 diabetes mellitus: A systematic review and meta–analysis. *PloS one 13*, 4 (Apr. 2018), e0195166.

[12] Bergfeldt, L., and Haga, Y. Power spectral and Poincaré plot characteristics in sinus node dysfunction. *Journal of Applied Physiology 94* (2003).

[13] Bernardi, L., Wdowczyk-Szulc, J., Valenti, C., Castoldi, S., Passino, C., Spadacini, G., and Sleight, P. Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability. *Journal of the American College of Cardiology 35*, 6 (2000), 1462–1469.

[14] Berntson, G. G., Thomas Bigger, J., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., Nagaraja, H. N., Porges, S. W., Saul, J. P., Stone, P. H., and VanDerMolem, M. W. Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology 34*, 6 (1997), 623–648.

[15] Bielecka-Dabrowa, A., Aronow, W. S., Rysz, J., and Banach, M. The Rise and Fall of Hypertension: Lessons Learned from Eastern Europe. *Current Cardiovascular Risk Reports 5*, 2 (Apr. 2011), 174–179.

[16] Billman, G. Heart Rate Variability – A Historical Perspective. *Frontiers in Physiology 2* (2011).

[17] Bolanos, M., Nazeran, H., and Haltiwanger, E. Comparison of Heart Rate Variability Signal Features Derived from Electrocardiography and Photoplethysmography in Healthy Individuals. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE* (2006), IEEE, pp. 4289–4294.

[18] Bolea, J., Laguna, P., Remartínez, J. M., Rovira, E., Navarro, A., and Bailón, R. Methodological Framework for Estimating the Correlation Dimension in HRV Signals. *Computational and Mathematical Methods in Medicine 2014* (Jan. 2014), e129248.

[19] Boonnithi, S., and Phongsuphap, S. Comparison of heart rate variability measures for mental stress detection. In *Computing in Cardiology, 2011* (2011), IEEE, pp. 85–88.

[20] Brosschot, J. F., Van Dijk, E., and Thayer, J. F. Daily worry is related to low heart rate variability during waking and the subsequent nocturnal sleep period. *International Journal of Psychophysiology 63*, 1 (2007), 39–47.

[21] Buchhorn, R., Baumann, C., and Willaschek, C. Heart Rate Variability in a Patient with Coronavirus Disease. *International Cardiovascular Forum Journal* (2020).

[22] Camm, A. J. e. a. Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *Circulation* (1996), 927–934.

[23] Cao, R., Azimi, I., Sarhaddi, F., Niela-Vilen, H., Axelin, A., Liljeberg, P., and Rahmani, A. M. Accuracy Assessment of Oura Ring Nocturnal Heart Rate and Heart Rate Variability in Comparison With Electrocardiography in Time and Frequency Domains: Comprehensive Analysis. *Journal of Medical Internet Research 24*, 1 (Jan. 2022), e27487.

[24] Castaldo, R., Melillo, P., Bracale, U., Caserta, M., Triassi, M., and Pecchia, L. Acute mental stress assessment via short term hrv analysis in healthy adults: A systematic review with meta-analysis. *Biomedical Signal Processing and Control 18* (2015), 370–377.

[25] Castaldo, R., Montesinos, P., and Melillo, C adn Pecchia, L. Ultra-short term HRV features as surrogates of short term HRV: a case study on mental stress detection in real life. *BMC Medical Informatics and Decision Making 19* (2019).

[26] Cavoukian, A., et al. Privacy by Design: The 7 Foundational Principles. *Information and Privacy Commissioner of Ontario, Canada 5* (2009).

[27] Celik, A., Ozturk, A., Ozbek, K., Kadi, H., Koc, F., Ceyhan, K., and Erkorkmaz, U. Heart rate variability and turbulence to determine true coronary artery disease in patients with ST segment depression without angina during exercise stress testing. *Clinical and Investigative Medicine* (Dec. 2011), E349–E357.

[28] Cha, S.-A., Park, Y.-M., Yun, J.-S., Lee, S.-H., Ahn, Y.-B., Kim, S.-R., and Ko, S.-H. Time- and frequency-domain measures of heart rate variability predict cardiovascular outcome in patients with type 2 diabetes. *Diabetes Research and Clinical Practice 143* (Sept. 2018), 159–169.

[29] Chandra, P., Sands, R. L., Gillespie, B. W., Levin, N. W., Kotanko, P., Kiser, M., Finkelstein, F., Hinderliter, A., Pop-Busui, R., Rajagopalan, S., and Saran, R. Predictors of heart rate variability and its prognostic significance in chronic kidney disease. *Nephrology Dialysis Transplantation 27*, 2 (Feb. 2012), 700–709.

[30] Chelidoni, O., Plans, D., Ponzo, S., Morelli, D., Cropley, M., et al. Exploring the effects of a brief biofeedback breathing session delivered through the biobase app in facilitating employee stress recovery: randomized experimental study. *JMIR mHealth and uHealth 8*, 10 (2020), e19412.

[31] Chen, J., Yeh, D., Lee, J., Chen, C., Huang, C., and SD, L. Methods of evaluation of autonomic nervous system function. *The Journal of Strength & Conditioning Research 25* (2011), 1546–1552.

[32] Chiew, C. J., Liu, N., Tagami, T., Wong, T. H., Koh, Z. X., and Ong, M. E. Heart rate variability based machine learning models for risk prediction of suspected sepsis patients in the emergency department. *Medicine 98*, 6 (2019).

[33] Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *Machine learning for healthcare conference* (2016), PMLR, pp. 301–318.

[34] Choi, E., Bahadori, M. T., Searles, E., Coffey, C., Thompson, M., Bost, J., Tejedor-Sojo, J., and Sun, J. Multi-layer Representation Learning for Medical Concepts. In *proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016), pp. 1495–1504.

[35] Choi, E., Schuetz, A., Stewart, W. F., and Sun, J. Medical Concept Representation Learning from Electronic Health Records and its Application on Heart Failure Prediction. *arXiv preprint arXiv:1602.03686* (2016).

[36] Choi, J., and Gutierrez-Osuna, R. Using Heart Rate Monitors to Detect Mental Stress. In *Wearable and Implantable Body Sensor Networks, 2009. BSN 2009. Sixth International Workshop on* (2009), IEEE, pp. 219–223.

[37] Chrysant, S. A new paradigm in the treatment of the cardiovascular disease continuum: focus on prevention. *Hippokratia 15*, 1 (2011), 7.

[38] Clifford, G., and Clifton, D. Annual Review: Wireless Technology in Disease State Management and Medicine. *Annual Review of Medicine 63* (2012), 479–492.

[39] Clifford, G., and Tarassenko, L. Quantifying errors in spectral estimates of HRV due to beat replacement and resampling. *IEEE Transactions on Biomedical Engineering 52*, 4 (Apr. 2005), 630–638.

[40] Clifton, D., Pimentel, M., Niehaus, K., Clifton, L., Peta, T., Crook, D., and Watkinson, P. Intelligent Electronic Health Systems. *Telemedicine and Electronic Medicine, CRC Press* (2015), 73–97.

[41] Cohen, S., Janicki-Deverts, D., and Miller, G. E. Psychological Stress and Disease. *JAMA 298*, 14 (Oct. 2007), 1685–1687.

[42] Cohn, J. N., Hoke, L., Whitwam, W., Sommers, P. A., Taylor, A. L., Duprez, D., Roessler, R., and Florea, N. Screening for early detection of cardiovascular disease in asymptomatic individuals. *American Heart Journal 146*, 4 (Oct. 2003), 679–685.

[43] Cornelissen, G. Cosinor-based rhythmometry. *Theoretical Biology & Medical Modelling 11* (2014), 16.

[44] Dabbah, M. A., Reed, A. B., Booth, A. T., Yassaee, A., Despotovic, A., Klasmer, B., Binning, E., Aral, M., Plans, D., Morelli, D., et al. Machine learning approach to dynamic risk modeling of mortality in covid-19: a uk biobank study. *Scientific reports 11*, 1 (2021), 16936.

[45] Dalton, A. R., Bottle, A., Okoro, C., Majeed, A., and Millett, C. Uptake of the NHS Health Checks programme in a deprived, culturally diverse setting: cross-sectional study. *Journal of Public Health 33*, 3 (05 2011), 422–429.

[46] Darabi, S., Kachuee, M., Fazeli, S., and Sarrafzadeh, M. TAPER: Time-Aware Patient EHR Representation. *IEEE Journal of Biomedical and Health Informatics 24*, 11 (Nov. 2020), 3268–3275.

[47] Das, M. K., and Zipes, D. P. Fragmented QRS: A predictor of mortality and sudden cardiac death. *Heart Rhythm 6*, 3 (2009), S8–S14.

[48] De Boor, C., and De Boor, C. *A Practical Guide to Splines*, vol. 27. Springer New York, 1978.

[49] Dekker, J. M., Crow, R. S., Folsom, A. R., Hannan, P. J., Liao, D., Swenne, C. A., and Schouten, E. G. Low Heart Rate Variability in a 2-Minute Rhythm Strip Predicts Risk of Coronary Heart Disease and Mortality From Several Causes. *Circulation 102*, 11 (2000), 1239–1244.

[50] Delaney, J. P. A., and Brodie, D. A. Effects of Short-Term Psychological Stress on the Time and Frequency Domains of Heart-Rate Variability. *Perceptual and Motor Skills 91*, 2 (Oct. 2000), 515–524.

[51] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[52] Dewey, F. E., Freeman, J. V., Engel, G., Oviedo, R., Abrol, N., Ahmed, N., Myers, J., and Froelicher, V. F. Novel predictor of prognosis from exercise stress testing: Heart rate variability response to the exercise treadmill test. *American Heart Journal 153*, 2 (Feb. 2007), 281–288.

[53] Dolezalova, N., Cairo, M., Despotovic, A., Booth, A. T., Reed, A. B., Morelli, D., and Plans, D. Development of a dynamic type 2 diabetes risk prediction tool: a uk biobank study. *arXiv e-prints* (2021), arXiv–2104.

[54] Dolezalova, N., Gkrania-Klotsas, E., Morelli, D., Moore, A., Cunningham, A. C., Booth, A., Plans, D., Reed, A. B., Aral, M., Rennie, K. L., et al. Feasibility of using intermittent active monitoring of vital signs by smartphone users to predict sars-cov-2 pcr positivity. *Scientific Reports 13*, 1 (2023), 10581.

[55] Dolezalova, N., Reed, A. B., Despotovic, A., Obika, B. D., Morelli, D., Aral, M., and Plans, D. Development of an accessible 10-year Digital CArdioVAscular (DiCAVA) risk assessment: a UK Biobank study. *European Heart Journal - Digital Health 2*, 3 (2021), 528–538.

[56] Dzedzickis, A., Kaklauskas, A., and Bucinskas, V. Human Emotion Recognition: Review of Sensors and Methods. *Sensors 20*, 3 (Jan. 2020), 592.

[57] Edelman, A., and Murakami, H. Polynomial roots from companion matrix eigenvalues. *Mathematics of Computation 64*, 210 (1995), 763–776.

[58] Edmonds, R., Sinclair, W., and Leicht, A. Effect of a Training Week on Heart Rate Variability in Elite Youth Rugby League Players. *International Journal of Sports Medicine 34* (2013), 1087–1092.

[59] Eguchi, K., Schwartz, J. E., Pickering, T. G., Hoshide, S., Ishikawa, J., Shimada, K., and Kario, K. Increased heart rate variability during sleep is a predictor for future cardiovascular events in patients with type 2 diabetes. *Hypertension Research 33*, 7 (July 2010), 737–742.

[60] Escobar-Restrepo, B., Torres-Villa, R., and Kyriacou, P. A. Evaluation of the Linear Relationship Between Pulse Arrival Time and Blood Pressure in ICU Patients: Potential and Limitations. *Frontiers in Physiology 9* (2018).

[61] Faria, I., and Drummond, B. Circadian changes in resting heart rate and body temperature, maximal oxygen consumption and perceived exertion. *Ergonomics 25*, 5 (1982), 381–386.

[62] Farivar, S., Abouzahra, M., and Ghasemaghaei, M. Wearable device adoption among older adults: A mixed-methods study. *International Journal of Information Management 55* (Dec. 2020), 102209.

[63] Fereidooni, H., Frassetto, T., Miettinen, M., Sadeghi, A.-R., and Conti, M. Fitness Trackers: Fit for Health but Unfit for Security and Privacy. In *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)* (2017), IEEE, pp. 19–24.

[64] Fernández, J. R., Mojón, A., Hermida, R. C., and Alonso, I. Methods for Comparison of Parameters from Longitudinal Rhythmometric Models with Multiple Components. *Chronobiology International 20*, 3 (Jan. 2003), 495–513.

[65] Floyer, J. *The Physicians Pulse-watch, Or an Essay to Explain the Old Art of Feeling the Pulse, and to Improve it by the Help of a Pulse Watch*, vol. 2. Sam. Smith, 1710.

[66] Fonderson, M., Bindels, P., Bohnen, A., and de Schepper, E. The role of area level social deprivation on childhood and adolescent consultation rate in primary care: a population based, cohort study. *BMC Primary Care 23*, 1 (2022), 1–10.

[67] Fossion, R., Rivera, A., and Estanol, B. A physicist's view of homeostasis: how time series of continuous monitoring reflect the function of physiological variables in regulatory mechanisms. *Physiological measurements 39* (2018).

[68] Fossion, R., Rivera, A. L., Toledo-Roy, J. C., Ellis, J., and Angelova, M. Multiscale adaptive analysis of circadian rhythms and intradaily variability: Application to actigraphy time series in acute insomnia subjects. *PloS one 12*, 7 (2017), e0181762.

[69] Fox, K., Borer, J. S., Camm, A. J., Danchin, N., Ferrari, R., Sendon, J. L. L., Steg, P. G., Tardif, J.-C., Tavazzi, L., and Tendera, M. Resting Heart Rate in Cardiovascular Disease. *Journal of the American College of Cardiology 50*, 9 (2007), 823–830.

[70] Freeman, J. V., Dewey, F. E., Hadley, D. M., Myers, J., and Froelicher, V. F. Autonomic Nervous System Interaction With the Cardiovascular System During Exercise. *Progress in cardiovascular diseases 48*, 5 (2006), 342–362.

[71] Friedman, B. H., and Thayer, J. F. Autonomic balance revisited: Panic anxiety and heart rate variability. *Journal of psychosomatic research 44*, 1 (1998), 133–151.

[72] Gao, L., Chen, Y., Shi, Y., Xue, H., and Wang, J. Value of DC and DRs in prediction of cardiovascular events in acute myocardial infarction patients. *Zhonghua yi xue za zhi 96* (2016).

[73] Germain, M., Gregor, K., Murray, I., and Larochelle, H. MADE: Masked Autoencoder for Distribution Estimation. In *International conference on machine learning* (2015), PMLR, pp. 881–889.

[74] Gil, A., Segura, J., and Temme, N. M. *Numerical methods for special functions*. SIAM, 2007.

[75] Gilbert, E. N. Capacity of a Burst-Noise Channel. *Bell system technical journal 39*, 5 (1960), 1253–1265.

[76] Gnauck, A. Interpolation and approximation of water quality time series and process identification. *Analytical and bioanalytical chemistry 380*, 3 (2004), 484–492.

[77] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., and Stanley, H. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation 101* (2000).

[78] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *circulation 101*, 23 (2000), e215–e220.

[79] Grassberger, P., and Procaccia, I. Measuring the strangeness of strange attractors. In *The theory of chaotic attractors*. Springer, 2004, pp. 170–189.

[80] Guzzetti, S., Piccaluga, E., Casati, R., Cerutti, S., Lombardi, F., Pagani, M., and Malliani, A. Sympathetic predominance in essential hypertension: a study employing spectral analysis of heart rate variability. *Journal of hypertension 6*, 9 (Sept. 1988), 711–717.

[81] Haghi, M., Thurow, K., and Stoll, R. Wearable Devices in Medical Internet of Things: Scientific Research and Commercially Available Devices. *Healthcare Informatics Research 23*, 1 (Jan. 2017), 4–15.

[82] Halberg, E., Halberg, F., and Shankaraiah, K. Plexo-serial linear-nonlinear rhythmometry of blood pressure, pulse and motor activity by a couple in their sixties. *Chronobiologia 8*, 4 (Dec. 1981), 351–366.

[83] Han, H., Kim, M.-J., and Kim, J. Development of real-time motion artifact reduction algorithm for a wearable photoplethysmography. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE* (2007), IEEE, pp. 1538–1541.

[84] Healey, J., Picard, R. W., and others. Detecting stress during real-world driving tasks using physiological sensors. *Intelligent Transportation Systems, IEEE Transactions on 6*, 2 (2005), 156–166.

[85] Hemmings, N. R., Kawadler, J. M., Whatmough, R., Ponzo, S., Rossi, A., Morelli, D., Bird, G., and Plans, D. Development and feasibility of a digital acceptance and commitment therapy–based intervention for generalized anxiety disorder: Pilot acceptability study. *JMIR Formative Research 5*, 2 (2021), e21737.

[86] Hillebrand, S., Gast, K., de Mutsert, R., Swenne, C., Jukema, J., Middeldorp, S., Rosendaal, F., and Dekkers, O. Heart rate variability and first cardiovascular event in populations without known cardiovascular disease: meta-analysis and dose–response meta-regression. *EP Europace 15* (2013).

[87] Hillebrand, S., Gast, K. B., de Mutsert, R., Swenne, C. A., Jukema, J. W., Middeldorp, S., Rosendaal, F. R., and Dekkers, O. M. Heart rate variability and first cardiovascular event in populations without known cardiovascular disease: meta-analysis and dose–response meta-regression. *EP Europace 15*, 5 (2013), 742–749.

[88] Hirten, R., Danieletto, M., Tomalin, L., Choi, K., Zweig, M., Golden, E., Kaur, S., Helmus, D., Biello, A., Pyzik, R., Nabeel, I., Charney, A., Glicksberg, B., Levin, M., Reich, D., Charney, D., Bottinger, E., Keefer, L Suarez-Farinas, M., Nadkarni, G., and Fayad, Z. Longitudinal Physiological Data fromaWearable Device Identifies SARS-CoV-2Infection andSymptoms and Predicts COVID-19 Diagnosis. *MedWxiv* (2020).

[89] Hoshi, R., Santos, I., Dantas, E., Andreão, R., Mill, J., Lotufo, P., and Bensenor, I. Reduced heart-rate variability and increased risk of hypertension: a prospective study of the ELSA-Brasil. *Journal of Human Hypertension 0* (2021).

[90] Huang, J., Sopher, S. M., Leatham, E., Redwood, S., Camm, A. J., and Kaski, J. C. Heart rate variability depression in patients with unstable angina. *American Heart Journal 130*, 4 (Oct. 1995), 772–779.

[91] Huikuri, H. V., Ylitalo, A., Pikkujämsä, S. M., Ikäheimo, M. J., Airaksinen, K. E. J., Rantala, A. O., Lilja, M., and Kesäniemi, Y. A. Heart rate variability in systemic hypertension. *The American Journal of Cardiology 77*, 12 (May 1996), 1073–1077.

[92] Hurst, J. W. Naming of the Waves in the ECG, With a Brief Account of Their Genesis. *Circulation 98*, 18 (1998), 1937–1942.

[93] Imai, H. *Essentials of Error-Control Coding*. Academic Press, 2014.

[94] Jaiswal, M., Urbina, E. M., Wadwa, R. P., Talton, J. W., D'Agostino, R. B., Hamman, R. F., Fingerlin, T. E., Daniels, S., Marcovina, S. M., and Dolan, L. M. Reduced Heart Rate Variability Among Youth With Type 1 Diabetes: The SEARCH CVD study. *Diabetes care 36*, 1 (2013), 157–162.

[95] Javorka, M., Zila, I., Balhárek, T., and Javorka, K. Heart rate recovery after exercise: relations to heart rate variability and complexity. *Brazilian Journal of Medical and Biological Research 35* (Aug. 2002), 991–1000.

[96] Jennings, J., Sheu, L., Kuan, D., Manuck, S., and Gianaros, P. Methods of evaluation of autonomic nervous system function. *Psychophysiology 53* (2016), 444–454.

[97] Jeyhani, V., Mahdiani, S., Peltokangas, M., and Vehkaoja, A. Comparison of HRV parameters derived from photoplethysmography and electrocardiography signals. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2015), IEEE, pp. 5952–5955.

[98] Jo, E., Lewis, K., Directo, D., Kim, M., and Dolezal, B. Validation of Biofeedback Wearables for Photoplethysmographic Heart Rate Tracking. *Journal of Sport Medicine 15* (2016).

[99] Johansson, B. A history of the electrocardiogram. *Dansk Medicinhistorisk Arbog* (2001), 163–176.

[100] Josephine, M., Lakshmanan, L., Resmi, R., Visu, P., Ganesan, R., and Jothikumar, R. Monitoring and sensing covid-19 symptoms as a precaution using electronic wearable devices. *International Journal of Pervasive Computing and Communications 16* (2020).

[101] Karcz, M; Chojnowska, L. W. Z. R. Prognostic significance of heart rate variability in dilated cardiomyopathy. *EP Europace 87* (2003), 75–81.

[102] Karcz, M; Chojnowska, L. W. Z. R. Prognostic significance of heart rate variability in dilated cardiomyopathy. *EP Europace 87* (2003), 75–81.

[103] Karlsson, M., Hörnsten, R., Rydberg, A., and Wiklund, U. Automatic filtering of outliers in RR intervals before analysis of heart rate variability in Holter recordings: a comparison with carefully edited data. *BioMedical Engineering OnLine 11* (2012), 2.

[104] Karthikeyan, P., Murugappan, M., and Yaacob, S. Detection of human stress using short-term ECG and HRV signals. *Journal of Mechanics in Medicine and Biology 13*, 02 (2013), 1350038.

[105] Kawadler, J. M., Hemmings, N. R., Ponzo, S., Morelli, D., Bird, G., and Plans, D. Effectiveness of a smartphone app (biobase) for reducing anxiety and increasing mental well-being: pilot feasibility and acceptability study. *JMIR formative research 4*, 11 (2020), e18067.

[106] Kemp, A., Quintana, D., Gray, M., Felmingham, K., Brown, K., and Gatt, J. Impact of Depression and Antidepressant Treatment on Heart Rate Variability: A Review and Meta-Analysis. *Biological Psychiatry 67* (2010), 1067–1074.

[107] Kim, K. K., Lim, Y. G., Kim, J. S., and Park, K. S. Effect of missing RR-interval data on heart rate variability analysis in the time domain. *Physiological Measurement 28*, 12 (Oct. 2007), 1485.

[108] Kleiger, R., Miller, J., Bigger, J., and AJ, M. Decreased heart rate variability and its association with increased mortality after acute myocardial infarction. *The American Journal of Cardiology 59* (1987), 256–262.

[109] Koukkari, W. L., and Sothern, R. B. *Introducing Biological Rhythms: A Primer on the Temporal Organization of Life, with Implications for Health, Society, Reproduction, and the Natural Environment.* Springer Netherlands, 2006.

[110] Kristal-Boneh, E., Silber, H., Harari, G., and Froom, P. The association of resting heart rate with cardiovascular, cancer and all-cause mortality. eight year follow-up of 3527 male israeli employees (the cordis study). *European Heart Journal 21*, 2 (2000), 116–124.

[111] Kudat, H., Akkaya, V., Sozen, A., Salman, S., Demirel, S., Ozcan, M., Atilgan, D., Yilmaz, M., and Guven, O. Heart Rate Variability in Diabetes Patients. *Journal of International Medical Research 34*, 3 (May 2006), 291–296.

[112] Laborde, S., Mosley, E., and Thayer, J. F. Heart Rate Variability and Cardiac Vagal Tone in Psychophysiological Research – Recommendations for Experiment Planning, Data Analysis, and Data Reporting. *Frontiers in Psychology 8* (2017).

[113] Lanza, G. A., Cianflone, D., Rebuzzi, A. G., Angeloni, G., Sestito, A., Ciriello, G., Torre, G. L., Crea, F., and Maseri, A. Prognostic value of ventricular arrhythmias and heart rate variability in patients with unstable angina. *Heart 92*, 8 (Aug. 2006), 1055–1063.

[114] Larsson, S., Drca, N., Mason, A. M., and Burgess, S. Resting heart rate and cardiovascular disease. *Circulation: Genomic and Precision Medicine 12* (2019).

[115] Lee, C., Yoon, J., and Van Der Schaar, M. Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis With Competing Risks Based on Longitudinal Data. *IEEE Transactions on Biomedical Engineering 67*, 1 (2019), 122–133.

[116] Lee, C., and Zhang, Y. Reduction of motion artifacts from photoplethysmographic recordings using a wavelet denoising approach. In *Biomedical Engineering, 2003. IEEE EMBS Asian-Pacific Conference on* (2003), IEEE, pp. 194–195.

[117] Lee, J., Matsumura, K., Yamakoshi, K.-i., Rolfe, P., Tanaka, S., and Yamakoshi, T. Comparison between red, green and blue light reflection photoplethysmography for heart rate monitoring during motion. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE* (2013), IEEE, pp. 1724–1727.

[118] Lees, T., Shad-Kaneez, F., Simpson, A. M., Nassif, N. T., Lin, Y., and Lal, S. Heart Rate Variability as a Biomarker for Predicting Stroke, Post-stroke Complications and Functionality. *Biomarker Insights 13* (2018), 1177271918786931.

[119] Leise, T. L. Analysis of Nonstationary Time Series for Biological Rhythms Research. *Journal of biological rhythms 32*, 3 (2017), 187–194.

[120] Leise, T. L., and Harrington, M. E. Wavelet-Based Time Series Analysis of Circadian Rhythms. *Journal of biological rhythms 26*, 5 (2011), 454–463.

[121] Lepot, M., Aubin, J.-B., and Clemens, F. H. L. R. Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment. *Water 9*, 10 (Oct. 2017), 796.

[122] Lerma, C., Minzoni, A., Infante, O., and José, M. V. A Mathematical Analysis for the Cardiovascular Control Adaptations in Chronic Renal Failure. *Artificial Organs 28*, 4 (2004), 398–409.

[123] Li, and et al. The Current State of Mobile Phone Apps for Monitoring Heart Rate, Heart Rate Variability, and Atrial Fibrillation: Narrative Review. *JMIR mHealth and uHealth 7* (2019).

[124] Li, F., Sheng, C., Zhang, D., An, D., Huang, J., Li, Y., and Wang, J. Resting heart rate in the supine and sitting positions as predictors of mortality in an elderly Chinese population. *Journal of Hypertension* (2019).

[125] Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., and Salimi-Khorshidi, G. BEHRT: Transformer for Electronic Health Records. *Scientific Reports 10*, 1 (Apr. 2020), 7155.

[126] Liew, W. S., Seera, M., Loo, C. K., Lim, E., and Kubota, N. Classifying Stress From Heart Rate Variability Using Salivary Biomarkers as Reference. *IEEE Transactions on Neural Networks and Learning Systems 27*, 10 (Oct. 2016), 2035–2046.

[127] Lin, W.-H., Wu, D., Li, C., Zhang, H., and Zhang, Y.-T. Comparison of heart rate variability from PPG with that from ECG. In *The International Conference on Health Informatics* (2014), Springer, pp. 213–215.

[128] Liu, N. T., Holcomb, J. B., Wade, C. E., Darrah, M. I., and Salinas, J. Utility of Vital Signs, Heart Rate Variability and Complexity, and Machine Learning for Identifying the Need for Lifesaving Interventions in Trauma Patients. *Shock 42*, 2 (2014), 108–114.

[129] Liu, X., Morelli, D., Littlejohns, T. J., Clifton, D. A., and Clifton, L. Combining machine learning with cox models to identify predictors for incident post-menopausal breast cancer in the uk biobank. *Scientific Reports 13*, 1 (2023), 9221.

[130] Lomb, N. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science 39* (1976).

[131] Lombardi, F. Origin of Heart Rate Variability and Turbulence: An Appraisal of Autonomic Modulation of Cardiovascular Function. *Frontiers in Physiology 2* (2011).

[132] Lu, G., Yang, F., Taylor, J. A., and Stein, J. F. A comparison of photoplethysmography and ECG recording to analyse heart rate variability in healthy subjects. *Journal of medical engineering & technology 33*, 8 (2009), 634–641.

[133] Ludwig, C. Beitrage zur Kenntniss des Einflusses der Respirationsbewegungen auf den Blutlauf im Aortensysteme. *Arch Anat Physiol Leipzig 13* (1847), 242–302.

[134] MacDonald, E., Rose, R., and Quinn, T. Neurohumoral Control of Sinoatrial Node Activity and Heart Rate: Insight From Experimental Models and Findings From Humans. *Frontiers in Physiology 11* (2020).

[135] MacDonald, E. A., Rose, R. A., and Quinn, T. A. Neurohumoral Control of Sinoatrial Node Activity and Heart Rate: Insight From Experimental Models and Findings From Humans. *Frontiers in Physiology 11* (2020).

[136] Malik, M. Heart Rate Variability. *Annals of Noninvasive Electrocardiology 1*, 2 (1996), 151–181.

[137] Malik, M., Farrell, T., and Camm, A. J. Circadian rhythm of heart rate variability after acute myocardial infarction and its influence on the prognostic value of heart rate variability. *American Journal of Cardiology 66*, 15 (Nov. 1990), 1049–1054.

[138] Marked, V. K. Correction of the Heart Rate Variability Signal for Ectopics and Missing Beats. *Heart rate variability* (1995).

[139] McCorry, L. K. Physiology of the Autonomic Nervous System. *American journal of pharmaceutical education 71*, 4 (2007).

[140] McCraty, R., and Shaffer, F. Heart Rate Variability: New Perspectives on Physiological Mechanisms, Assessment of Self-regulatory Capacity, and Health Risk. *Glob Adv Health Med 4:46* (2015).

[141] Mejía-Mejía, E., Allen, J., Budidha, K., El-Hajj, C., Kyriacou, P. A., and Charlton, P. H. Photoplethysmography signal processing and synthesis. In *Photoplethysmography*, J. Allen and P. Kyriacou, Eds. Academic Press, Jan. 2022, pp. 69–146.

[142] Melillo, P., Izzo, R., Orrico, A., Scala, P., Attanasio, M., Mirra, M., Luca, N. D., and Pecchia, L. Automatic Prediction of Cardiovascular and Cerebrovascular Events Using Heart Rate Variability Analysis. *PloS one 10*, 3 (Mar. 2015), e0118504.

[143] Meloni, C., Stazi, F., Ballarotto, C., Margonato, A., and Chierchia, S. L. Heart rate variability in patients with variant angina: effect of the presence of significant coronary stenosis. *Italian heart journal 1*, 7 (July 2000), 470–474.

[144] Miao, X., He, W., Yang, H., and Tail, H.-M. Heart rate variability characterization using correlation dimension. In *The 2002 45th Midwest Symposium on Circuits and Systems, 2002. MWSCAS-2002.* (Aug. 2002), vol. 1, pp. I–447.

[145] Michael, S., Graham, K. S., and Davis, G. M. Cardiac Autonomic Responses during Exercise and Post-exercise Recovery Using Heart Rate Variability and Systolic Time Intervals—A Review. *Frontiers in Physiology 8* (2017).

[146] Millasseau, S. C., Ritter, J. M., Takazawa, K., and Chowienczyk, P. J. Contour analysis of the photoplethysmographic pulse measured at the finger. *Journal of Hypertension 24*, 8 (Aug. 2006), 1449–1456.

[147] Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports 6*, 1 (May 2016), 26094.

[148] Molloy, T. J., Okin, P. M., Devereux, R. B., and Kligfield, P. Electrocardiographic detection of left ventricular hypertrophy by the simple QRS voltage-duration product. *Journal of the American College of Cardiology 20*, 5 (1992), 1180–1186.

[149] Morelli, D., Bartoloni, L., Colombo, M., and Clifton, D. Profiling the propagation of error from PPG to HRV features in a wearable physiological-monitoring device. *Healthcare technology letters 5* (2018), 59–64.

[150] Morelli, D., Bartoloni, L., Rossi, A., and DA, C. A computationally efficient algorithm to obtain an accurate and interpretable model of the effect of circadian rhythm on resting heart rate. *Pysiological Measurements 40* (2019).

[151] Morelli, D., Dolezalova, N., Ponzo, S., Colombo, M., and Plans, D. Development of digitally obtainable 10-year risk scores for depression and anxiety in the general population. *Frontiers in Psychiatry 12* (2021), 689026.

[152] Morelli, D., Rossi, A., Bartoloni, L., Cairo, M., and Clifton, D. A. Sdnn24 estimation from semi-continuous hr measures. *Sensors 21*, 4 (2021), 1463.

[153] Morelli, D., Rossi, A., Cairo, M., and Clifton, D. Analysis of the Impact of Interpolation Methods of Missing RR-Intervals Caused by Motion Artifacts on HRV Features Estimations. *Sensors 19* (2019), 3163.

[154] Munoz, M., Van Roon, A., Riese, H., Thio, C., Oostenbroek, E., Westrik, I., De Geus, EJC adn Gansevoort, R., Lefrandt, J., Nolte, I., and H, S. Validity of (Ultra-)Short Recordings for Heart Rate Variability Measurements. *Plosone 10(9)* (2015), e0138921.

[155] Murukesan, L., Murugappan, M., Iqbal, M., and Saravanan, K. Machine Learning Approach for Sudden Cardiac Arrest Prediction Based on Optimal Heart Rate Variability Features. *Journal of Medical Imaging and Health Informatics 4*, 4 (2014), 521–532.

[156] Mussalo, H., Vanninen, E., Ikäheimo, R., Laitinen, T., Laakso, M., Länsimies, E., and Hartikainen, J. Heart rate variability and its determinants in patients with severe or mild essential hypertension. *Clinical Physiology 21*, 5 (2001), 594–604.

[157] Nakagawa, M., Iwao, T., Ishida, S., Yonemochi, H., Fujino, T., Saikawa, T., and Ito, M. Circadian rhythm of the signal averaged electrocardiogram and its relation to heart rate variability in healthy subjects. *Heart 79*, 5 (May 1998), 493–496.

[158] Nguyen, D., Luo, W., Venkatesh, S., and Phung, D. Effective Identification of Similar Patients Through Sequential Matching over ICD Code Embedding. *Journal of Medical Systems 42*, 5 (Apr. 2018), 94.

[159] Nguyen, P., Tran, T., Wickramasinghe, N., and Venkatesh, S. Deepr: A Convolutional Net for Medical Records. *IEEE Journal of Biomedical and Health Informatics 21*, 1 (Jan. 2017), 22–30.

[160] Nikbakhtian, S., Reed, A. B., Obika, B. D., Morelli, D., Cunningham, A. C., Aral, M., and Plans, D. Accelerometer-derived sleep onset timing and cardiovascular disease incidence: a uk biobank cohort study. *European Heart Journal-Digital Health 2*, 4 (2021), 658–666.

[161] of the European Society of Cardiology the North American Society of Pacing Electrophysiology, T. F. Heart Rate Variability: Standards of Measurment, Physiological Interpretation, and Clinical Use. *Circulation 93* (1996), 1043–1065.

[162] Oikawa, K., Ishihara, R., Maeda, T., Yamaguchi, K., Koike, A., Kawaguchi, H., Tabata, Y., Murotani, N., and Itoh, H. Prognostic value of heart rate variability in patients with renal failure on hemodialysis. *International Journal of Cardiology 131*, 3 (2009), 370–377.

[163] Olshansky, B., Sabbah, H. N., Hauptman, P. J., and Colucci, W. S. Parasympathetic Nervous System and Heart Failure: Pathophysiology and Potential Implications for Therapy. *Circulation 118*, 8 (2008), 863–871.

[164] Orphanidou, C., Bonnici, T., Charlton, P., Clifton, D., Vallance, D., and Tarassenko, L. Signal Quality Indices for the Electrocardiogram and Photoplethysmogram: Derivation and Applications to Wireless Monitoring. *IEEE Journal of Biomedical and Health Informatics 19*, 3 (2015), 832–838.

[165] Papamakarios, G., Pavlakou, T., and Murray, I. Masked Autoregressive Flow for Density Estimation. *Advances in neural information processing systems 30* (2017).

[166] Pearl, J. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence* (2012), pp. 3–11.

[167] Pecchia, L., Castaldo, R., Montesinos, L., and Melillo, P. Are ultra-short heart rate variability features good surrogates of short-term ones? State-of-the-art review and recommendations. *Healthcare Technology Letters 5(3)* (2018), 94–100.

[168] Peltola, M. Role of editing of R-R intervals in the analysis of heart rate variability. *Frontiers in Physiology 3* (2012).

[169] Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., and Goldberger, A. L. Mosaic organization of DNA nucleotides. *Physical review e 49*, 2 (1994), 1685.

[170] Peng, R.-C., Zhou, X.-L., Lin, W.-H., and Zhang, Y.-T. Extraction of Heart Rate Variability from Smartphone Photoplethysmograms. *Computational and mathematical methods in medicine 2015* (2015).

[171] Perini, R., and Veicsteinas, A. Heart rate variability and autonomic activity at rest and during exercise in various physiological conditions. *European journal of applied physiology 90*, 3 (2003), 317–325.

[172] Pham, T., Tran, T., Phung, D., and Venkatesh, S. DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (2016), Springer, pp. 30–41.

[173] Pincus, S. M. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences 88*, 6 (1991), 2297–2301.

[174] Pincus, S. M., and Goldberger, A. L. Physiological time-series analysis: what does regularity quantify? *American Journal of Physiology-Heart and Circulatory Physiology 266*, 4 (1994), H1643–H1656.

[175] Pitsavos, C. H., Chrysohoou, C., Panagiotakos, D. B., Kokkinos, P., Skoumas, J., Papaioannou, I., Michaelides, A. P., Singh, S., and Stefanadis, C. I. Exercise capacity and heart rate recovery as predictors of coronary heart disease events, in patients with heterozygous familial hypercholesterolemia. *Atherosclerosis 173*, 2 (2004), 345–350.

[176] Plans, D., Morelli, D., Sütterlin, S., Ollis, L., Derbyshire, G., Cropley, M., et al. Use of a biofeedback breathing app to augment poststress physiological recovery: randomized pilot study. *JMIR formative research 3*, 1 (2019), e12227.

[177] Plans, D., Ponzo, S., Morelli, D., Cairo, M., Ring, C., Keating, C. T., Cunningham, A. C., Catmur, C., Murphy, J., and Bird, G. Measuring interoception: The phase adjustment task. *Biological Psychology 165* (Oct. 2021), 108171.

[178] Plews, D., Laursen, P., Kilding, A., and M, B. Heart-Rate Variability and Training-Intensity Distribution in Elite Rowers. *International Journal of Sports Physiology and Performance* (2014).

[179] Pomeranz, B., Macaulay, R. J., Caudill, M. A., Kutz, I., Adam, D., Gordon, D., Kilborn, K. M., Barger, A. C., Shannon, D. C., Cohen, R. J., and Benson, H. Assessment of autonomic function in humans by heart rate spectral analysis. *American Journal of Physiology - Heart and Circulatory Physiology 248*, 1 (1985), H151–H153.

[180] Ponzo, S., Morelli, D., Kawadler, J. M., Hemmings, N. R., Bird, G., and Plans, D. Efficacy of the digital therapeutic mobile app biobase to reduce stress and improve mental well-being among university students: randomized controlled trial. *JMIR mHealth and uHealth 8*, 4 (2020), e17767.

[181] Ponzo, S., Morelli, D., Suksasilp, C., Cairo, M., and Plans, D. Measuring interoception: The cardiac elevation detection task. *Frontiers in Psychology 12* (2021), 712896.

[182] Porta, A., Guzzetti, S., Montano, N., Furlan, R., Pagani, M., Malliani, A., and Cerutti, S. Entropy, entropy rate, and pattern classification as tools to typify complexity in short heart period variability series. *IEEE Transactions on Biomedical Engineering 48*, 11 (2001), 1282–1291.

[183] Ranpuria, R., Hall, M., Chan, C. T., and Unruh, M. Heart rate variability (HRV) in kidney failure: measurement and consequences of reduced HRV. *Nephrology Dialysis Transplantation 23*, 2 (Feb. 2008), 444–449.

[184] Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine 4*, 1 (May 2021), 1–13.

[185] Richman, J. S., and Moorman, J. R. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology 278*, 6 (2000), H2039–H2049.

[186] Rogers, B., Giles, D., Draper, N., Hoos, O., and Gronwald, T. A new detection method defining the aerobic threshold for endurance exercise and training prescription based on fractal correlation properties of heart rate variability. *Frontiers in physiology 11* (2021), 596567.

[187] Rohila, A., and Sharma, A. Correlation Between Heart Rate Variability Features. In *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)* (Feb. 2020), pp. 669–674. ISSN: 2688-769X.

[188] Roque, F. S., Jensen, P. B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., Søeby, K., Bredkjær, S., Juul, A., Werge, T., Jensen, L. J., and Brunak, S. Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. *PLOS Computational Biology 7*, 8 (2011), e1002141.

[189] Rosenberg, A., Weiser-Bitoun, I., Billman, G., and Yaniv, Y. Signatures of the autonomic nervous system and the heart's pacemaker cells in canine electrocardiograms and their applications to humans. *Scientific Reports 10* (2020).

[190] Rosenberg, A. A., Weiser-Bitoun, I., Billman, G. E., and Yaniv, Y. Signatures of the autonomic nervous system and the heart's pacemaker cells in canine electrocardiograms and their applications to humans. *Scientific reports 10*, 1 (2020), 1–15.

[191] Rossi, A., Da Pozzo, E., Menicagli, D., Tremolanti, C., Priami, C., Sirbu, A., Clifton, D., Martini, C., and Morelli, D. A Public Dataset of 24-h Multi-Levels Psycho-Physiological Responses in Young Healthy Adults. *Data* (2020).

[192] Rossi, A., Da Pozzo, E., Menicagli, D., Tremolanti, C., Priami, C., Sirbu, A., Clifton, D., Martini, C., and Morelli, D. Multilevel Monitoring of Activity and Sleep in Healthy People (version 1.0.0). *PhysioNet* (2020).

[193] Rossi, A., Pedreschi, D., Clifton, D., and Morelli, D. Error Estimation of Ultra-Short Heart Rate Variability Parameters: Effect of Missing Data Caused by Motion Artifacts. *Sensors 20* (2020).

[194] Routledge, F. S., Campbell, T. S., McFetridge-Durdle, J. A., and Bacon, S. L. Improvements in heart rate variability with exercise therapy. *The Canadian Journal of Cardiology 26*, 6 (2010), 303–312.

[195] Russoniello, C. V., Pougtachev, V., Zhirnov, E., and Mahar, M. T. A Measurement of Electrocardiography and Photoplethesmography in Obese Children. *Applied psychophysiology and biofeedback 35*, 3 (2010), 257–259.

[196] Salahuddin, L., Cho, J., Jeong, M. G., and Kim, D. Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE* (2007), IEEE, pp. 4656–4659.

[197] Salo, M. A., Huikuri, H. V., and Seppänen, T. Ectopic Beats in Heart Rate Variability Analysis: Effects of Editing on Time and Frequency Domain Measures. *Annals of Noninvasive Electrocardiology: The Official Journal of the International Society for Holter and Noninvasive Electrocardiology, Inc 6*, 1 (Jan. 2001), 5–17.

[198] Sartipy, U., Savarese, G., Dahlström, U., Fu, M., and Lund, L. H. Association of heart rate with mortality in sinus rhythm and atrial fibrillation in heart failure with preserved ejection fraction. *European Journal of Heart Failure 21*, 4 (2019), 471–479.

[199] Scargle. Studies in Astronomical Time Series Analysis. V. Bayesian Blocks, a New Method to Analyze Structure in Photon Counting Data. *The American Astronomical Society 504* (1998).

[200] Schäfer, A., and Vagedes, J. How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram. *International journal of cardiology 166*, 1 (2013), 15–29.

[201] Schroeder, E. B., Liao, D., Chambless, L. E., Prineas, R. J., Evans, G. W., and Heiss, G. Hypertension, blood pressure, and heart rate variability: the atherosclerosis risk in communities (aric) study. *Hypertension 42*, 6 (2003), 1106–1111.

[202] Selvaraj, N., Jaryal, A., Santhosh, J., Deepak, K. K., and Anand, S. Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography. *Journal of medical engineering & technology 32*, 6 (2008), 479–484.

[203] Shaffer, F., and Ginsberg, J. P. An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health 5* (2017).

[204] Shaffer, F., McCraty, R., and Zerr, C. A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. *Front Psychol 5* (2014), 1040.

[205] Shah, S., Velardo, C., Farmer, A., and Tarassenko, L. Exacerbations in Chronic Obstructive Pulmonary Disease: Identification and Prediction Using a Digital Health System. *Journal of Medical Internet Research 19*, 3 (2017), e69.

[206] Shahar, E., Whitney, C., Redline, S., Lee, E., Newman, A., Nieto, F., O'connor, G., Boland, L., Schwartz, J., and Samet, J. Sleep-disordered Breathing and Cardiovascular Disease Cross-sectional Results of the Sleep Heart Health Study. *American Journal of Respiratory and Critical Care Medicine 163*, 1 (2001).

[207] Shang, J., Ma, T., Xiao, C., and Sun, J. Pre-training of Graph Augmented Transformers for Medication Recommendation. *arXiv preprint arXiv:1906.00346* (2019).

[208] Shashikant, R., and Chetankumar, P. Predictive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter. *Applied Computing and Informatics* (2020).

[209] Shcherbina, A., Mattsson, C., Waggott, D., Salisbury, H., Christle, J., Hastie, T., Wheeler, M., and Ashley, E. Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *Journal of Persersonalized Medicine 7* (2017).

[210] Shetler, K., Marcus, R., Froelicher, V. F., Vora, S., Kalisetti, D., Prakash, M., Do, D., and Myers, J. Heart rate recovery: validation and methodologic issues. *Journal of the American College of Cardiology 38*, 7 (2001), 1980–1987.

[211] Sibson, R. A brief description of natural neighbour interpolation. *Interpreting multivariate data* (1981).

[212] Singh, J. P., Larson, M. G., Tsuji, H., Evans, J. C., O'Donnell, C. J., and Levy, D. Reduced Heart Rate Variability and New-Onset Hypertension. *Hypertension 32*, 2 (Aug. 1998), 293–297.

[213] Singh, N., Moneghetti, K. J., Christle, J. W., Hadley, D., Froelicher, V., and Plews, D. Heart Rate Variability: An Old Metric with New Meaning in the Era of Using mHealth technologies for Health and Exercise Training Guidance. Part Two: Prognosis and Training. *Arrhythmia & Electrophysiology Review 7*, 4 (Dec. 2018), 247–255.

[214] Stein, P. K., Barzilay, J. I., Chaves, P. H., Mistretta, S. Q., Domitrovich, P. P., Gottdiener, J. S., Rich, M. W., and Kleiger, R. E. Novel Measures of Heart Rate Variability Predict Cardiovascular

Mortality in Older Adults Independent of Traditional Cardiovascular Risk Factors: The Cardiovascular Health Study (CHS). *Journal of Cardiovascular Electrophysiology 19*, 11 (2008), 1169–1174.

[215] Stein, P. K., Kleiger, R. E., and Rottman, J. N. Differing Effects of Age on Heart Rate Variability in Men and Women. *The American journal of cardiology 80*, 3 (1997), 302–305.

[216] Stein, P. K., and Reddy, A. Non-Linear Heart Rate Variability and Risk Stratification in Cardiovascular Disease. *Indian Pacing and Electrophysiology Journal 5*, 3 (July 2005), 210–220.

[217] Stirling, J. R., Zakynthinaki, M., Refoyo, I., and Sampedro, J. A Model of Heart Rate Kinetics in Response to Exercise. *Journal of Nonlinear Mathematical Physics 15*, sup3 (2008), 426–436.

[218] Strimbu, K., and Tavel, J. A. What are Biomarkers? *Current opinion in HIV and AIDS 5*, 6 (Nov. 2010), 463–466.

[219] Taelman, J., Vandeput, S., Spaepen, A., and Van Huffel, S. Influence of Mental Stress on Heart Rate and Heart Rate Variability. In *4th European conference of the international federation for medical and biological engineering* (2009), Springer, pp. 1366–1369.

[220] Tamayo, T., Rosenbauer, J., Wild, S. H., Spijkerman, A. M. W., Baan, C., Forouhi, N. G., Herder, C., and Rathmann, W. Diabetes in Europe: An update. *Diabetes Research and Clinical Practice 103*, 2 (Feb. 2014), 206–217.

[221] Tanaka, H., Monahan, K. D., and Seals, D. R. Age-predicted maximal heart rate revisited. *Journal of the american college of cardiology 37*, 1 (2001), 153–156.

[222] Tang, Y., Shah, H., Junior, C., Sun, X., Mitri, J., Sambataro, M., Sambado, L., Gerstein, H., Fonseca, V., Doria, A., and Pop-Busui, R. Intensive Risk Factor Management and Cardiovascular Autonomic Neuropathy in Type 2 Diabetes: The ACCORD Trial. *Diabetes Care 44* (2021).

[223] Tarassenko, L., and Clifton, D. Semiconductor Wireless Technology for Chronic Disease Management. *Electronics Letters S30* (2011), 30–32.

[224] Taylor, J., and Hamilton, S. Some tests of the Vaníček method of spectral analysis. *Astrophysics and Space Science 17*, 2 (1972), 357–367.

[225] Teng, X. F., and Zhang, Y. T. Study on the peak interval variability of photoplethysmogtaphic signals. In *Biomedical Engineering, 2003. IEEE EMBS Asian-Pacific Conference on* (2003), IEEE, pp. 140–141.

[226] Thayer, J. F., Yamamoto, S. S., and Brosschot, J. F. The relationship of autonomic imbalance, heart rate variability and cardiovascular disease risk factors. *International journal of cardiology 141*, 2 (2010), 122–131.

[227] Thoits, P. A. Stress and Health: Major Findings and Policy Implications. *Journal of Health and Social Behavior 51*, 1_suppl (Mar. 2010), S41–S53.

[228] Tsuji, H., Larson, M. G., Venditti, F. J., Manders, E. S., Evans, J. C., Feldman, C. L., and Levy, D. Impact of Reduced Heart Rate Variability on Risk for Cardiac Events. *Circulation 94*, 11 (Dec. 1996), 2850–2855.

[229] Tsutsui, K., Monfredi, O., Sirenko-Tagirova, S., Maltseva, L., and EA, L. A coupled-clock system drives the automaticity of human sinoatrial nodal pacemaker cells. *Science Signaling 11* (2018).

[230] Tulppo, M. P., Makikallio, T. H., Takala, T. E., Seppanen, T., and Huikuri, H. V. Quantitative beat-to-beat analysis of heart rate dynamics during exercise. *American journal of physiology-heart and circulatory physiology 271*, 1 (1996), H244–H252.

[231] Tzoulaki, I., Liberopoulos, G., and Ioannidis, J. P. Assessment of claims of improved prediction beyond the framingham risk score. *Jama 302*, 21 (2009), 2345–2352.

[232] Ucak, S., Dissanayake, H., Sutherland, K., de Chazal, P., and Cistulli, P. Heart rate variability and obstructive sleep apnea: Current perspectives and novel technologies. *Journal of Sleep Research 0* (2021).

[233] Uusitalo, A., Mets, T., Martinmäki, K., Mauno, S., Kinnunen, U., and Rusko, H. Heart rate variability related to effort at work. *Applied Ergonomics 42*, 6 (Nov. 2011), 830–838.

[234] Vander Plas, J. T. Understanding the Lomb–Scargle Periodogram. *The Astrophysical Journal Supplement Series 236*, 1 (2018), 16.

[235] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \., and Polosukhin, I. Attention is All you Need. *Advances in neural information processing systems 30* (2017).

[236] Violán, C., Roso-Llorach, A., Foguet-Boreu, Q., Guisado-Clavero, M., Pons-Vigués, M., Pujol-Ribera, E., and Valderas, J. M. Multimorbidity patterns with K-means nonhierarchical cluster analysis. *BMC Family Practice 19*, 1 (July 2018), 108.

[237] Wagner, G. S., Freye, C. J., Palmeri, S. T., Roark, S. F., Stack, N. C., Ideker, R. E., Harrell Jr, F. E., and Selvester, R. H. Evaluation of a QRS scoring system for estimating myocardial infarct size. I. Specificity and observer agreement. *Circulation 65*, 2 (1982), 342–347.

[238] Waller, A. D. A Demonstration on Man of Electromotive Changes accompanying the Heart's Beat. *The Journal of physiology 8*, 5 (1887), 229.

[239] Wartelle, A., Mourad-Chehade, F., Yalaoui, F., Chrusciel, J., Laplanche, D., and Sanchez, S. Clustering of a Health Dataset Using Diagnosis Co-Occurrences. *Applied Sciences 11*, 5 (Jan. 2021), 2373.

[240] Weiler, D., Villajuan, S., Edkins, L., Cleary, S., and JJ, S. Wearable Heart Rate Monitor Technology Accuracy in Research: A Comparative Study Between PPG and ECG Technology. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 61* (2017).

[241] Wilkins, E., Wilson, L., Wickramasinghe, K., Bhatnagar, P., Leal, J., Luengo-Fernandez, R., Burns, R., Rayner, M., and Townsend, N. *European Cardiovascular Disease Statistics 2017*. European Heart Network, Brussels, Feb. 2017.

[242] Willson, K., Francis, D. P., Wensel, R., Coats, A. J. S., and Parker, K. H. Relationship between detrended fluctuation analysis and spectral analysis of heart-rate variability. *Physiological Measurement 23*, 2 (Mar. 2002), 385.

[243] Xhyheri, B., Manfrini, O., Mazzolini, M., Pizzi, C., and Bugiardini, R. Heart Rate Variability Today. *Progress in Cardiovascular Diseases 55*, 3 (Nov. 2012), 321–331.

[244] Xiong, J., Liang, X., Zhao, L., Lo, B., Li, J., and Liu, C. Improving accuracy of heart failure detection using data refinement. *Entropy 22*, 5 (2020), 520.

[245] Yadav, R. L., Yadav, P. K., Yadav, L. K., Agrawal, K., Sah, S. K., and Islam, M. N. Association between obesity and heart rate variability indices: an intuition toward cardiac autonomic alteration – a risk of CVD. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy 10* (Feb. 2017), 57–64.

[246] Yaniv, Y., Lyashkov, A. E., and Lakatta, E. G. The fractal-like complexity of heart rate variability beyond neurotransmitters and autonomic receptors: signaling intrinsic to sinoatrial node pacemaker cells. *Cardiovascular pharmacology: open access 2* (2013).

[247] Yuan, H., Plekhanova, T., Walmsley, R., Reynolds, A. C., Maddison, K. J., Bucan, M., Gehrman, P., Rowlands, A., Ray, D. W., Bennett, D., et al. Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality. *medRxiv* (2023).

[248] Yugar, L. B. T., Yugar-Toledo, J. C., Dinamarco, N., Sedenho-Prado, L. G., Moreno, B. V. D., Rubio, T. d. A., Fattori, A., Rodrigues, B., Vilela-Martin, J. F., and Moreno, H. The role of heart rate variability (hrv) in different hypertensive syndromes. *Diagnostics 13*, 4 (2023), 785.

[249] Zhang, L., Wu, H., Zhang, X., Wei, X., Hou, F., and Ma, Y. Sleep heart rate variability assists the automatic prediction of long-term cardiovascular outcomes. *Sleep medicine 67* (Mar. 2020), 217–224.

[250] Zhang, X., and Xu, H. Prediction of angina pectoris events in middle-aged and elderly people using rr interval time series in the resting state: A cohort study based on shhs. *International Journal of Computational Intelligence Systems 16*, 1 (2023), 10.

[251] Zhang, Z., Liu, J., and Razavian, N. BERT-XML: Large scale automated ICD coding using BERT pretraining. *arXiv preprint arXiv:2006.03685* (2020).

[252] Zygmunt, A., and Stanczyk, J. Methods of evaluation of autonomic nervous system function. *Archives of Medical Science 6* (2010), 11–18.

Appendix A

Hyper-parameter search

In this section I report details of the hyper parameter search for each of the models, with a figure that shows the AUC throughout the trials, and a figure with a plot for each parameter showing the AUC reached in the Y axis and the values attempted in the X axis.

A.1 BASIC

This section reports the best hyperparameters for the BASIC model, that uses no HRV and not EHR data, in table A.1.

param_name	param_value
BATCH_SIZE_exponent	4
layers	1
learning_rate	0.000103
units_hidden_exponent	7

Table A.1: Selected hyper parameters for the BASIC model

Figure A.1 shows the AUC achieved by the trials during the hyperparameter search. The search did not seem to converge to the maximum AUC, indicating that either more trials would have been necessary, or that the optimal performance was immediately achieved. Figure A.2 shows the values attempted for the hyperparameters with the AUC achieved. It can be noted that none of the parameters seems to have a clear optimal value, also indicating that either more trials would be needed, or that the hyperparameters made little difference to the models performance. The fact that the range of achieved AUC is very narrow (between 0.618 and 0.632) indicates that it is more likely that the model achieves the best possible results quickly, regardless of the choice of the parameters.

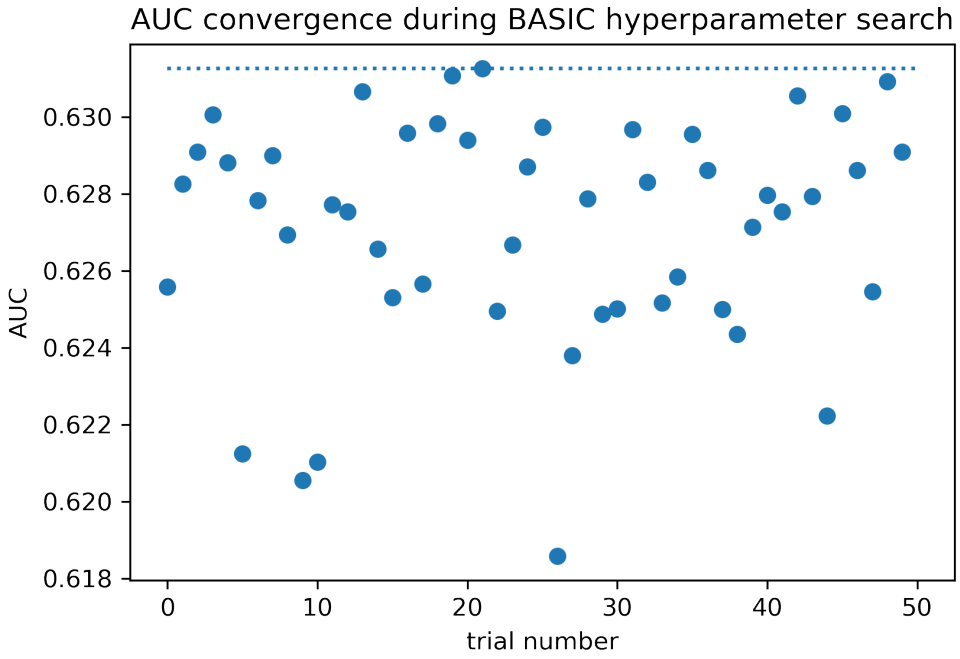


Figure A.1: AUC of the trials during the hyper-parameter search using BASIC as inputs

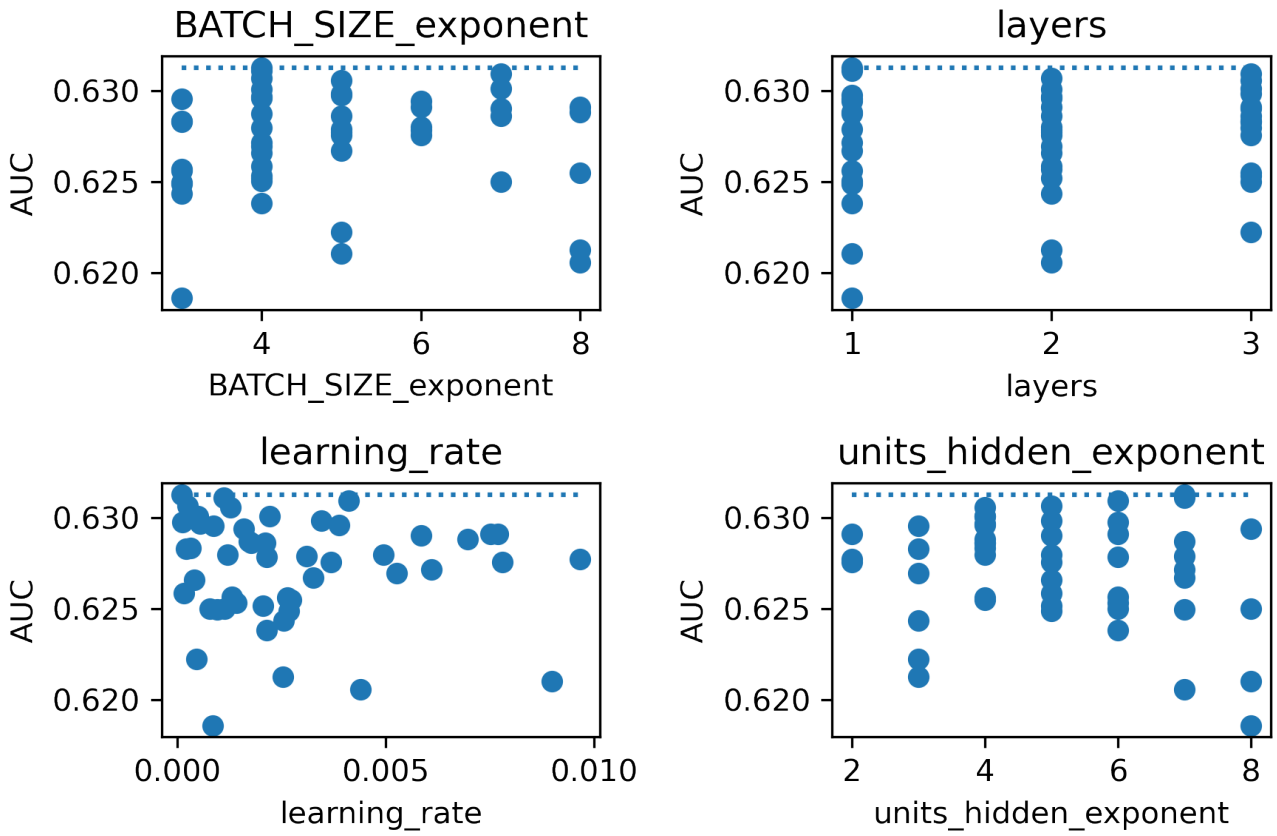


Figure A.2: Parameters attempted during the hyper-parameter search using BASIC as inputs

A.2 EHR

Table A.2 shows the hyper parameters selected by the bayesian optimisation for the model that uses EHR data.

This model shows better convergence to the best AUC, and a larger spread in AUC achieved by trials, with early trials close to random AUC (0.5) and late trials more stably closer to the best AUC (0.66), as shown in figure A.3. The graphs that show the AUC achieved by the different attempted values (figure A.4) do not seem to indicate a single winning combination that achieves the highest AUC, indicating that different combinations of batch size, number of layers, and number of hidden units can achieve similar performance.

param_name	param_value
BATCH_SIZE_exponent	4
layers	1
learning_rate	0.00059
units_hidden_exponent	7

Table A.2: Selected hyper parameters for the EHR model

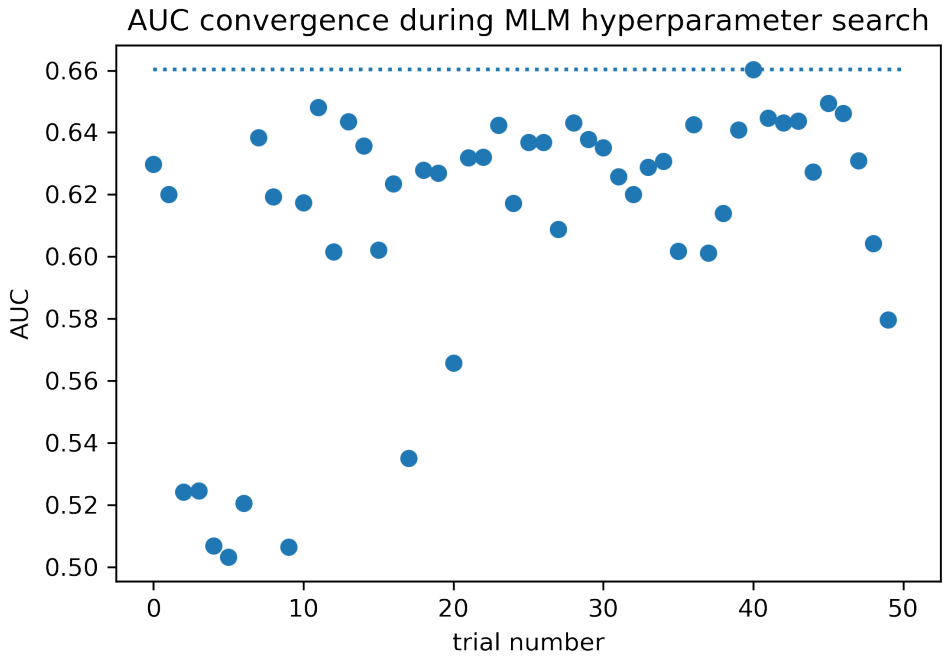


Figure A.3: AUC of the trials during the hyper-parameter search using EHR as inputs

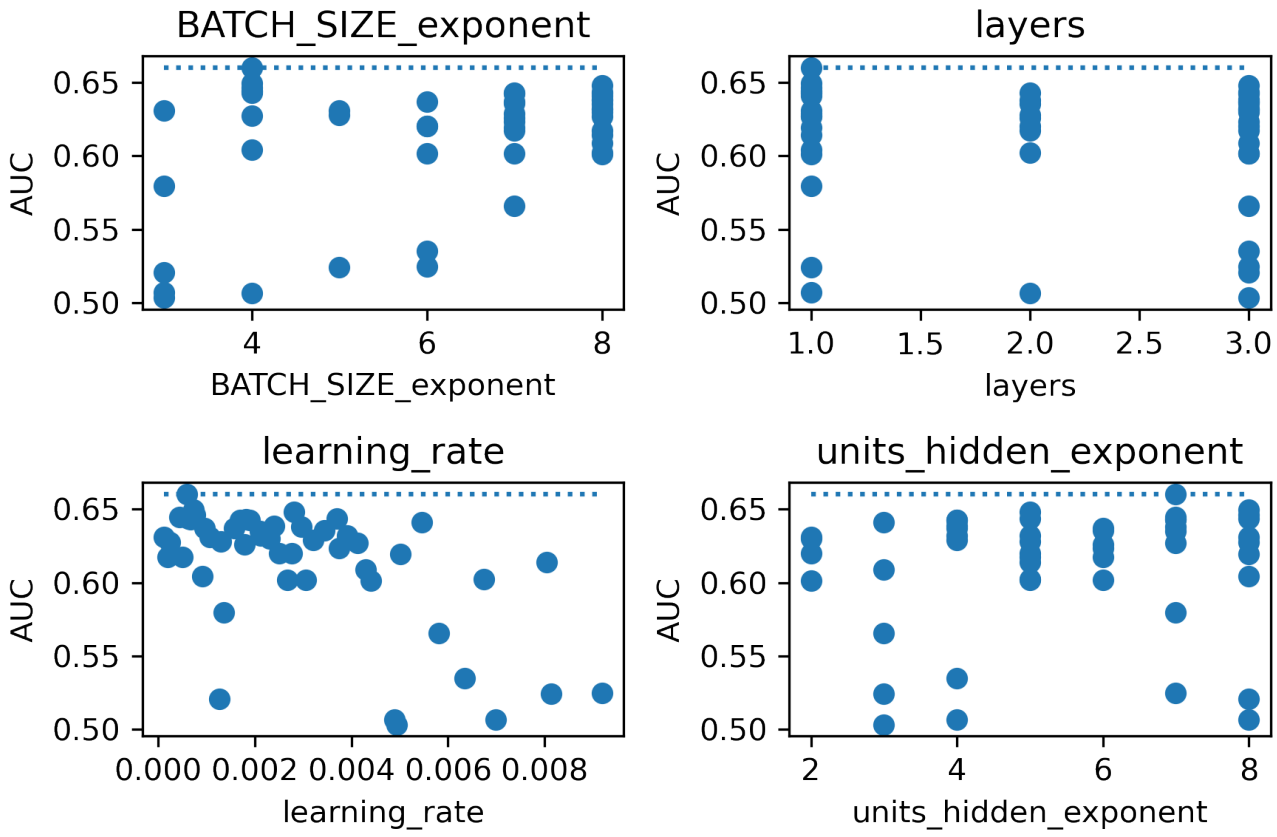


Figure A.4: Parameters attempted during the hyper-parameter search using EHR as inputs

A.3 HRV

Table A.3 shows the hyper parameters selected by the Bayesian optimisation for the model that uses HRV data.

Also for this model the AUC achieved by trials started at levels near random (0.5), but quickly achieved the best AUC (0.66), and showed to achieve optimal results often, indicating good convergence (figure A.5).

Figure A.6 shows that for HRV data large batch sizes, small learning rates, a single layer topology, and large numbers of nodes in the hidden layer, resulted in better AUC.

param_name	param_value
BATCH_SIZE_exponent	8
layers	1
learning_rate	0.000586
units_hidden_exponent	7

Table A.3: Selected hyperparameters for HRV

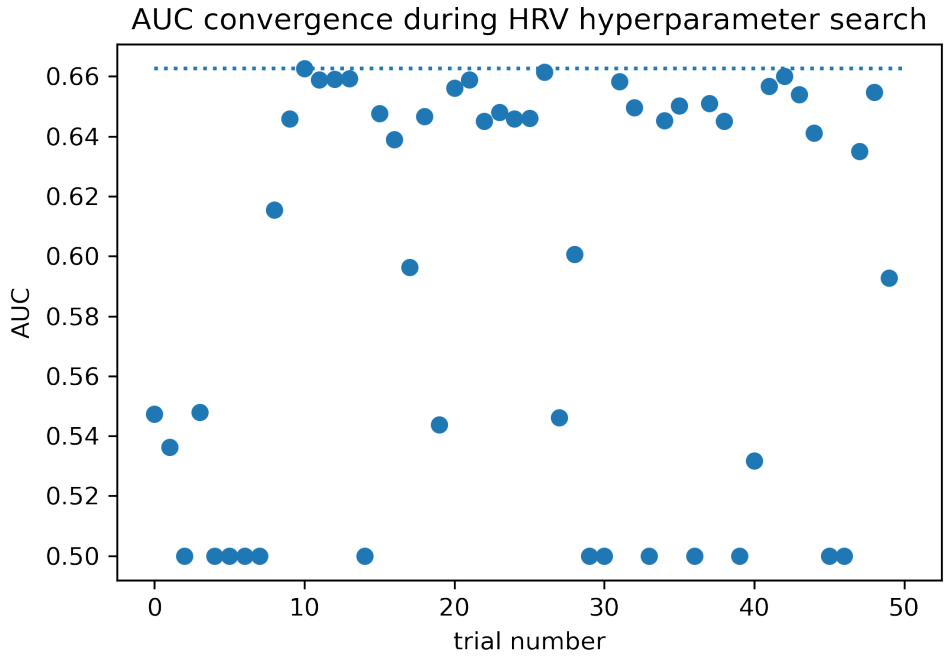


Figure A.5: AUC of the trials during the hyper-parameter search using HRV as inputs

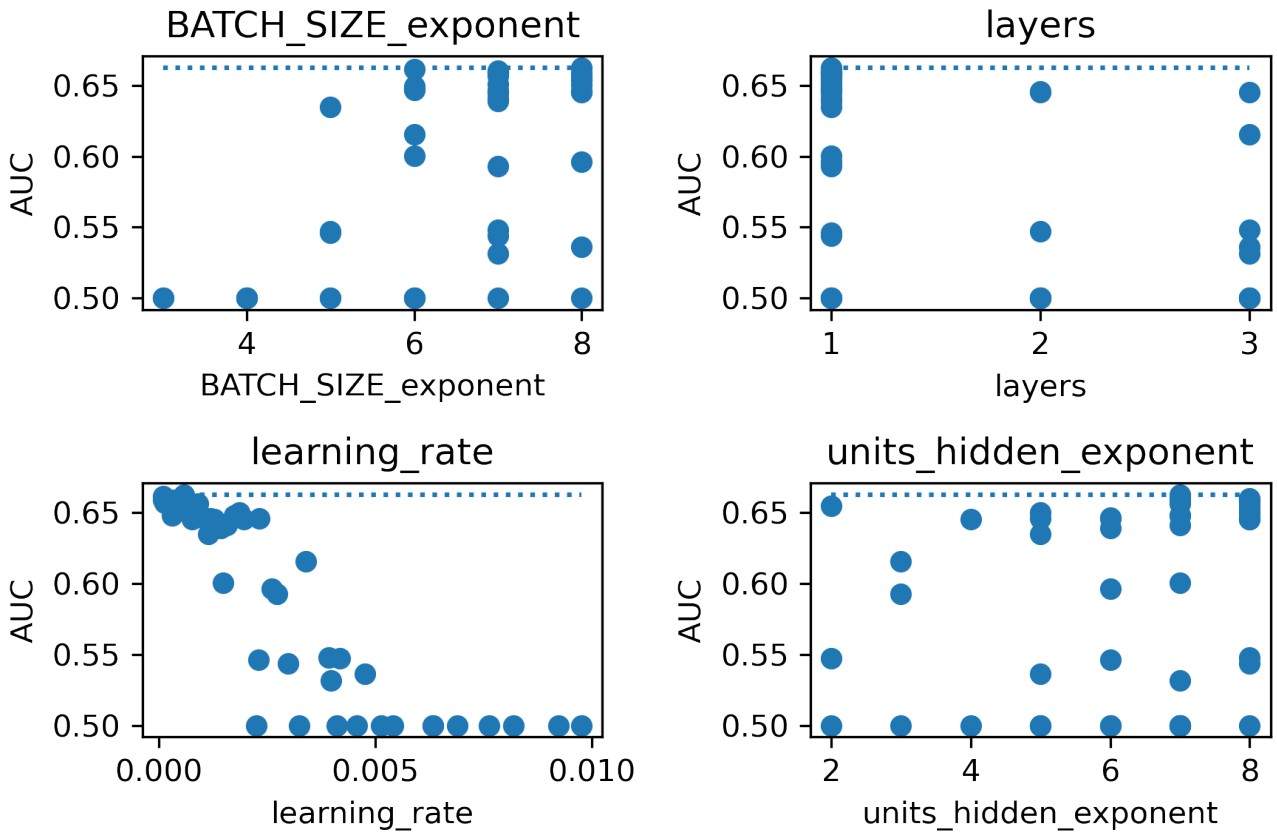


Figure A.6: Parameters attempted during the hyper-parameter search using HRV as inputs

A.4 BASIC+HRV

Table A.4 shows the hyper parameters selected by the Bayesian optimisation for the model that uses both BASIC and HRV data.

The graphs in figure A.8 show that the best performance are achieved with small learning rates, large batch sizes, a single hidden layer with a small number of nodes. This indicates that to avoid overfitting the model needs to be simple.

Figure A.7 shows that the range in achieved AUC is rather narrow: the initial trials achieved 0.62 and the best trials achieved less than 0.67. The best trial was trial 45 of 50, it is possible that allowing more trials I would have found a different configuration of parameters with better AUC.

param_name	param_value
BATCH_SIZE_exponent	7
layers	1
learning_rate	0.000209
units_hidden_exponent	2

Table A.4: Selected hyperparameters for BASIC+HRV

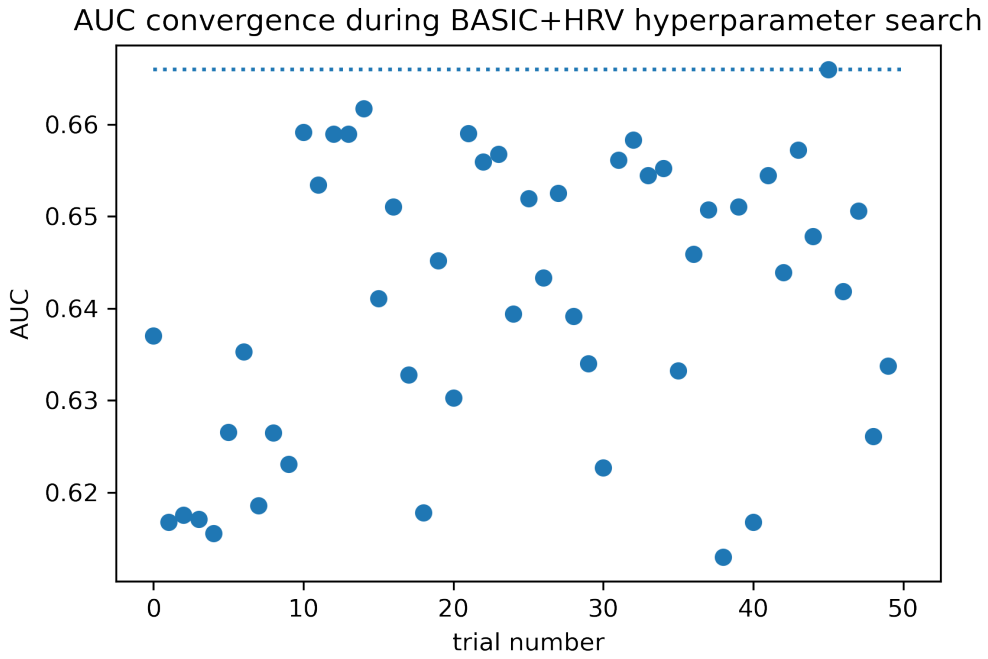


Figure A.7: AUC of the trials during the hyper-parameter search using BASIC+HRV as inputs

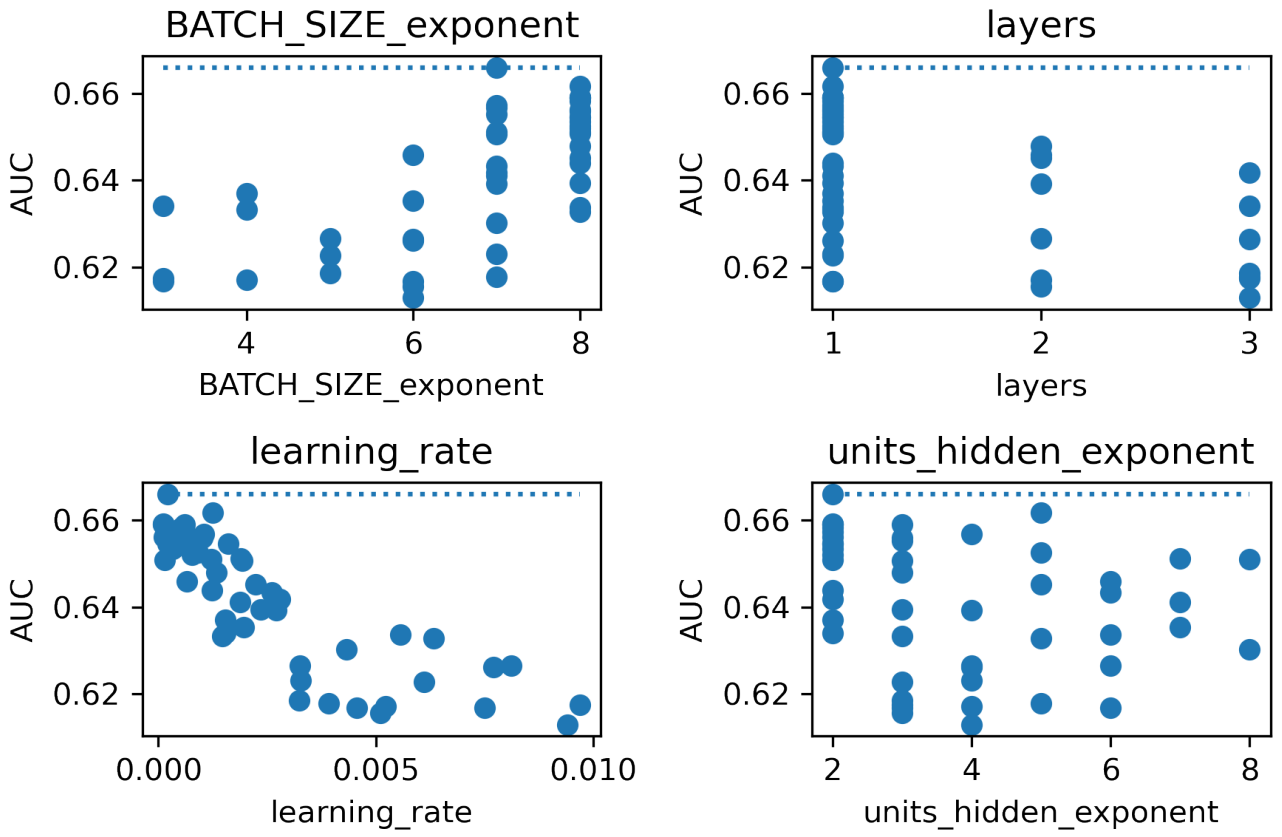


Figure A.8: Parameters attempted during the hyper-parameter search using BASIC+HRV as inputs

A.5 BASIC+EHR

Table A.5 shows the hyper parameters selected by the Bayesian optimisation for the model that uses both BASIC and EHR data.

The history of AUC of the trials during the hyperparameter search, shown in figure A.9, is similar to BASIC+HRV, the AUC range is narrow (0.62 to 0.67), the best AUC is achieved at the end of the search (trial 44 of 50), that indicates that it might be beneficial to allow a larger number of trials in the search.

Figure A.10 show that there is convincing evidence for the chosen hyperparameter values, as they are on very distinct peaks with respect to the other options.

param_name	param_value
BATCH_SIZE_exponent	5
layers	1
learning_rate	0.001126
units_hidden_exponent	5

Table A.5: Selected hyperparameters for BASIC+EHR

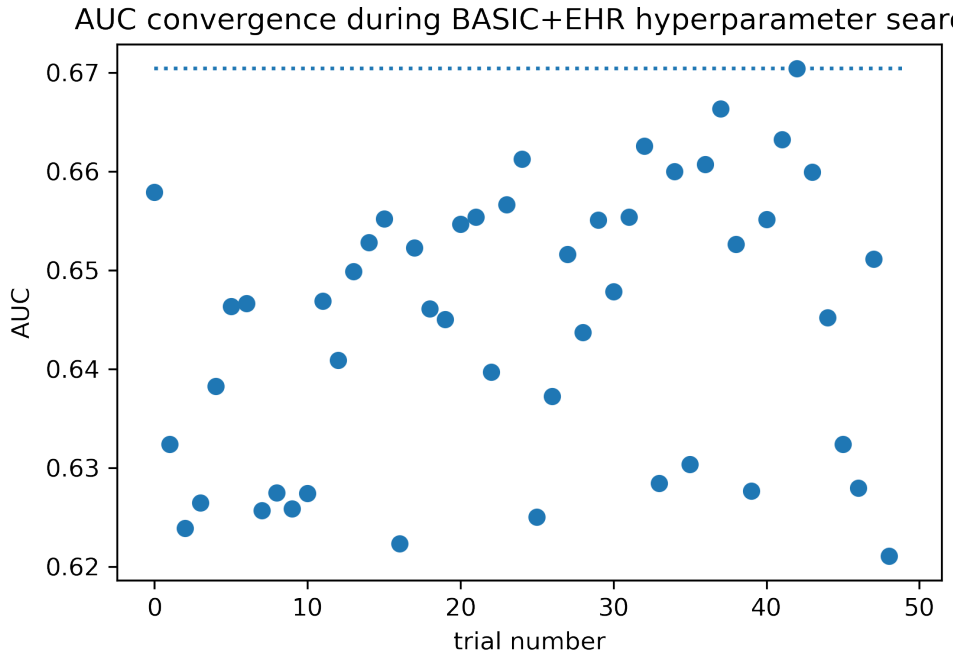


Figure A.9: AUC of the trials during the hyper-parameter search using BASIC+EHR as inputs

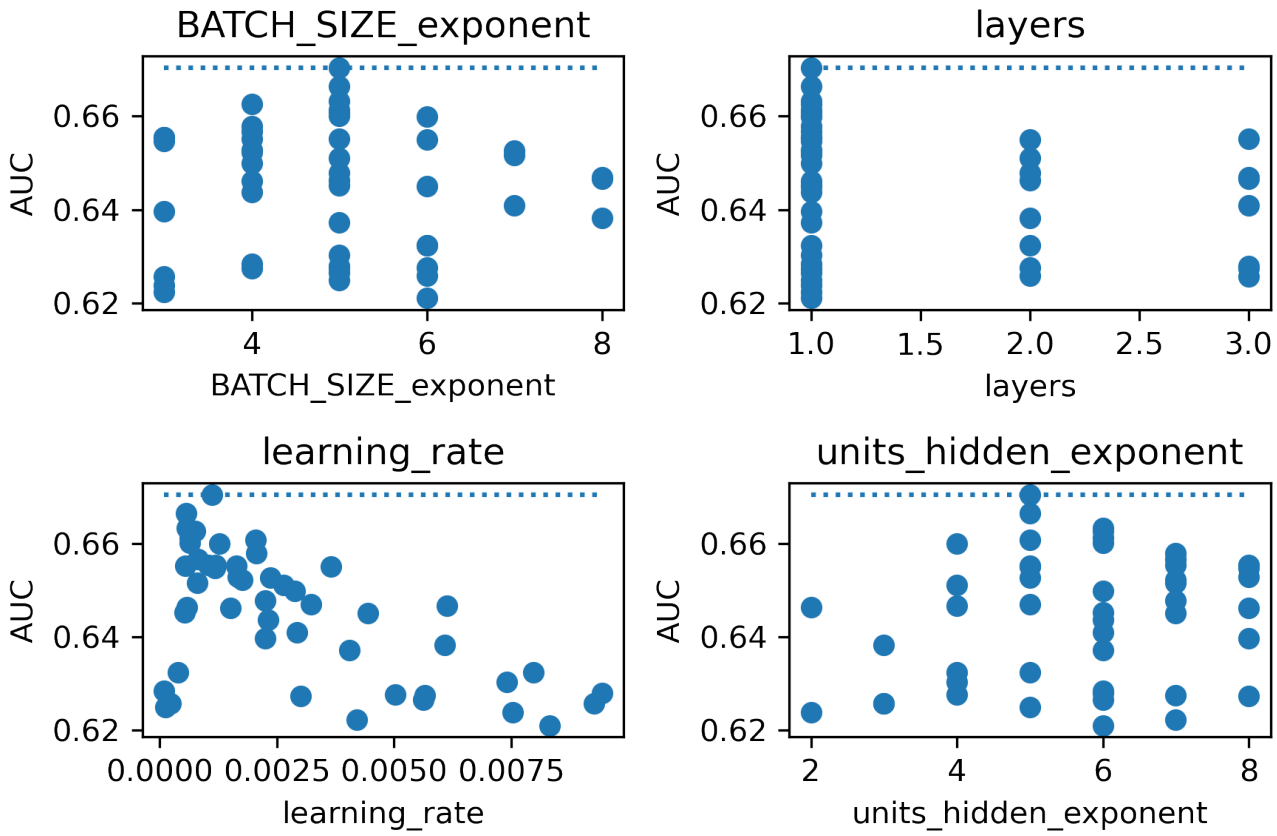


Figure A.10: Parameters attempted during the hyper-parameter search using BASIC+EHR as inputs

A.6 BASIC+HRV+EHR

Table A.6 shows the hyper parameters selected by the Bayesian optimisation for the model that uses BASIC, HRV, and EHR data.

Looking at figure A.11 it can be seen that AUC values of about 0.68 (the best AUC) were obtained multiple times, indicating a satisfactory stability in performance.

Figure A.12 shows clear trends in the correlation between parameter values and obtained AUC. Best performances were obtained with large batch sizes, small learning rates, small number of nodes in the hidden layers, and a topology with two layers.

param_name	param_value
BATCH_SIZE_exponent	8
layers	2
learning_rate	0.001895
units_hidden_exponent	4

Table A.6: Selected hyperparameters for BASIC+HRV+EHR

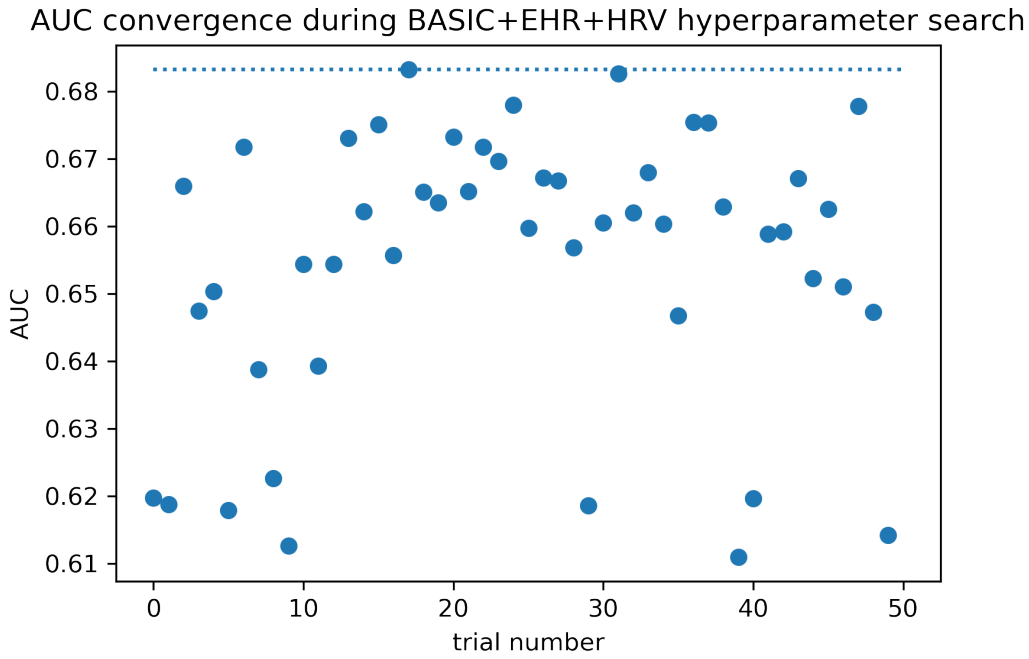


Figure A.11: AUC of the trials during the hyper-parameter search using BASIC+HRV+EHR as inputs

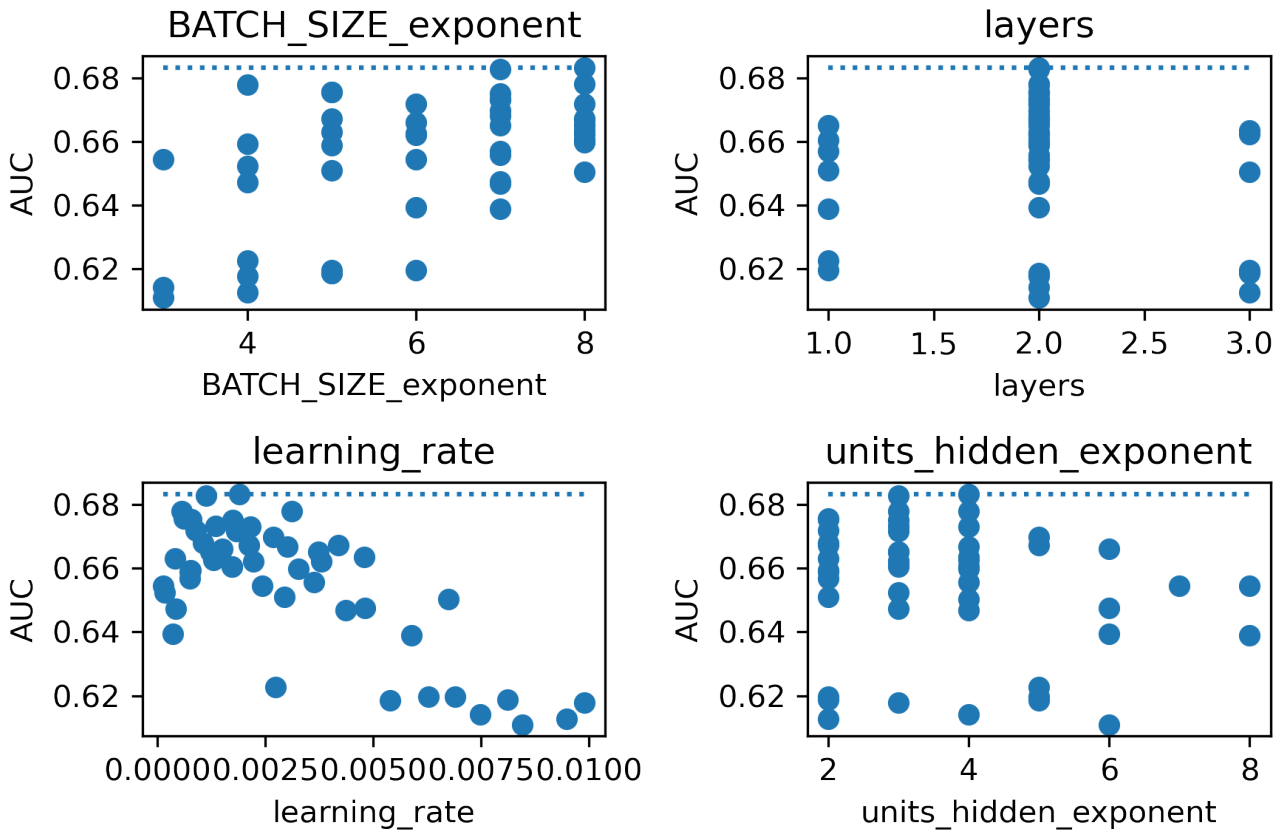


Figure A.12: Parameters attempted during the hyper-parameter search using BASIC+HRV+EHR as inputs

A.7 HRV+EHR

Table A.7 shows the hyper parameters selected by the Bayesian optimisation for the model that uses both HRV and EHR data.

Figure A.13 shows a wide range of AUC (from 0.5 to 0.675), but no increasing trend can be appreciated, and figure A.14 shows little effect of different choices of parameters on resulting AUC: using 1 layer I obtained performance similar to using 2, or 3 layers. Similar considerations can be done for batch size and number of units in the hidden layers. Learning rate is the only hyperparameter that shows inverse correlation: I obtained high AUCs with small learning rates.

param_name	param_value
BATCH_SIZE_exponent	5
layers	3
learning_rate	0.000526
units_hidden_exponent	6

Table A.7: Selected hyperparameters for HRV+EHR

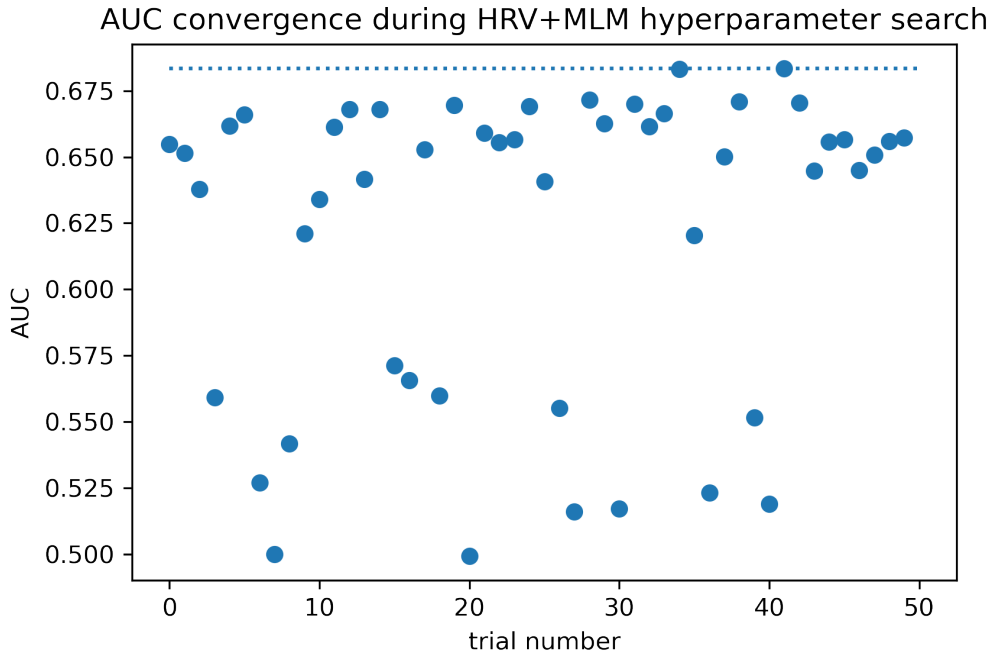


Figure A.13: AUC of the trials during the hyper-parameter search using HRV+EHR as inputs

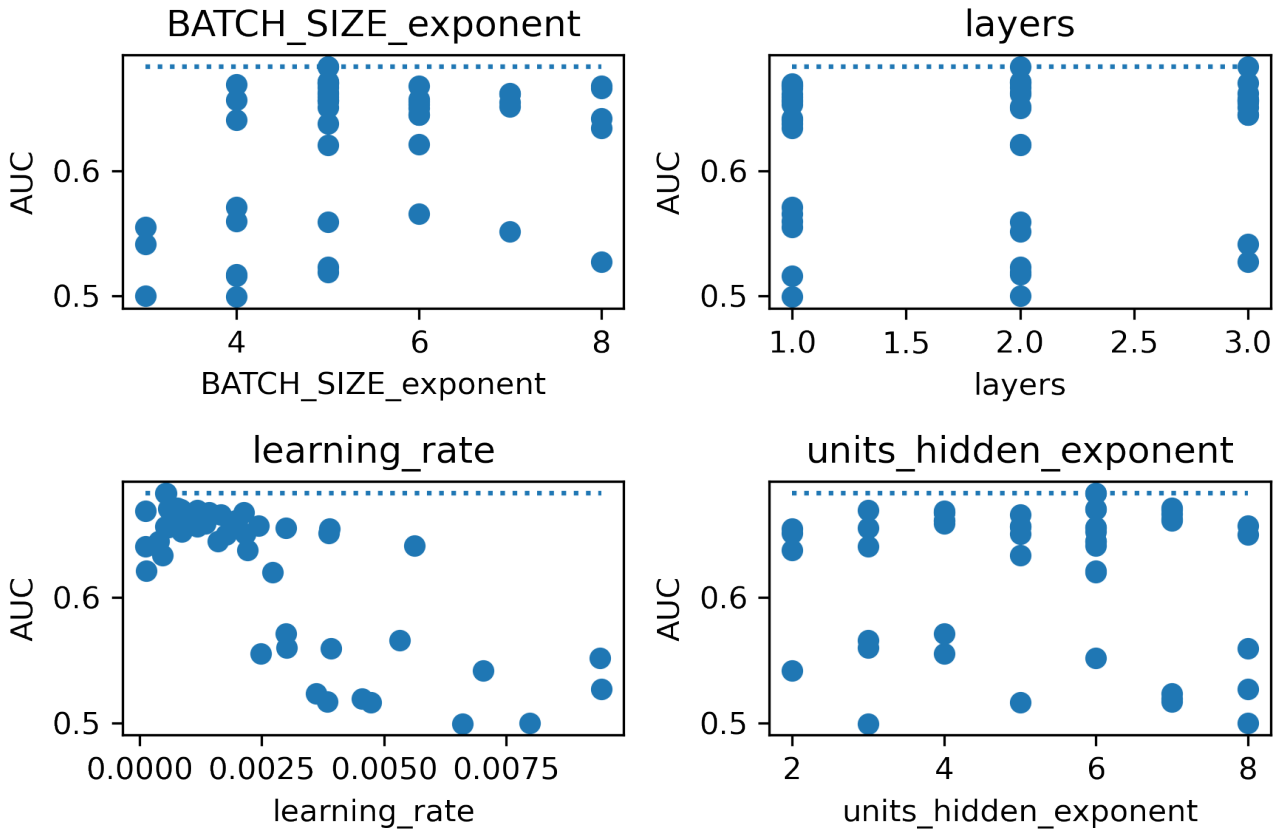
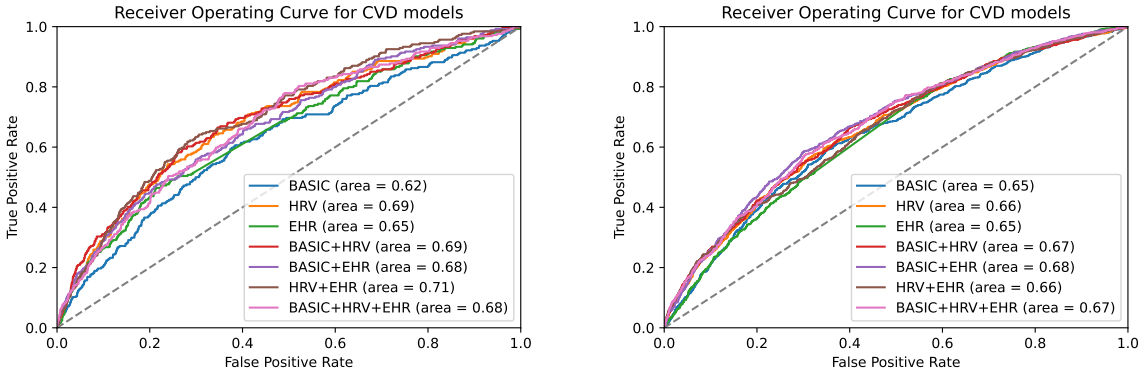


Figure A.14: Parameters attempted during the hyper-parameter search using HRV+EHR as inputs

Appendix B

Receiver Operating Characteristic curves for prediction windows at 5 and 10 years

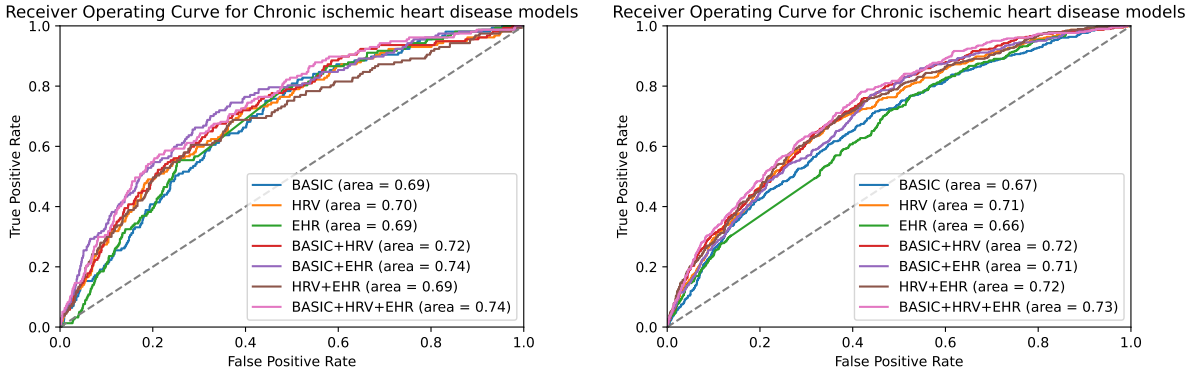
In this section I report the ROC curves of the Health outcome prediction experiment described in Chapter 7. For each target I report the curves of all models (BASIC, EHR, HRV, BASIC+EHR, BASIC+HRV, BASIC+HRV+EHR) at 5 years prediction window and at 10 years prediction window.



(a) 5 years prediction window

(b) 10 years prediction window

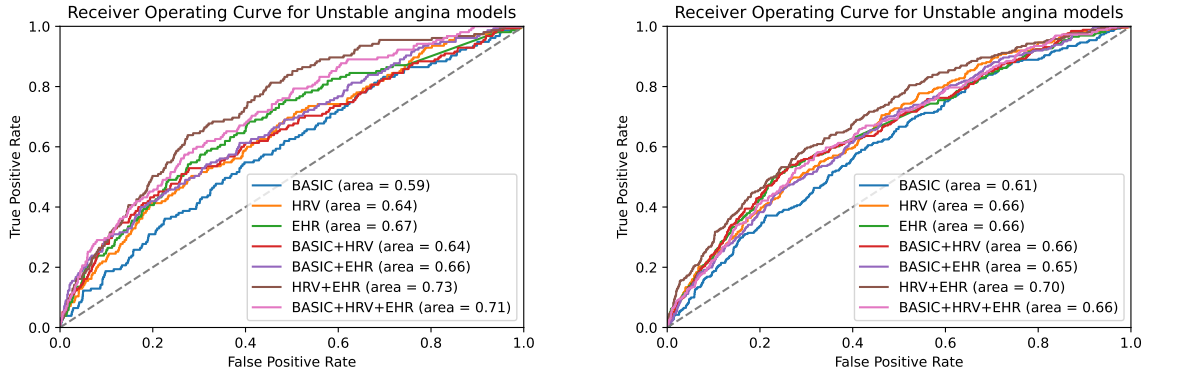
Figure B.1: ROC curves for all models for CVD, on the test set



(a) 5 years prediction window

(b) 10 years prediction window

Figure B.2: ROC curves for all models for CHD (Chronic Ischemic heart disease), on the test set



(a) 5 years prediction window

(b) 10 years prediction window

Figure B.3: ROC curves for all models for Unstable Angina, on the test set

Appendix C

Results for prediction window at 2 years

In this appendix I report the health outcome prediction results for a time windows of 2 years, one table for each target outcome: table C.1 show results for CVD; table C.2 show results for CHD; table C.3 show results for Angina;

	AUC
BASIC	0.59
HRV	0.62
EHR	0.63
BASIC+HRV	0.63
BASIC+EHR	0.63
HRV+EHR	0.63
BASIC+HRV+EHR	0.65

Table C.1: AUC for CVD on the test set with a prediction window of 2 years

	AUC
BASIC	0.67
HRV	0.72
EHR	0.66
BASIC+HRV	0.68
BASIC+EHR	0.68
HRV+EHR	0.71
BASIC+HRV+EHR	0.73

Table C.2: AUC for CHD (Chronic Ischemic heart disease) on the test set

	AUC
BASIC	0.61
HRV	0.69
EHR	0.61
BASIC+HRV	0.71
BASIC+EHR	0.61
HRV+EHR	0.67
BASIC+HRV+EHR	0.67

Table C.3: AUC for Unstable Angina on the test set

Appendix D

Machine used to run the Health Outcome Prediction experiments

The machine used to run the experiments was a VM running in Google Cloud and had the following configuration:

The Operating System was a Linux 4.19.0-22-cloud-amd64 #1 SMP Debian 4.19.260-1 (2022-09-29) x86_64 GNU/Linux

The machine had 32 GB of RAM.

The machine had a GPU: NVIDIA Tesla T4.

Python version was 3.7.12

The main pip packages were:

Package	Version
[...]	
conda	4.11.0
keras	2.8.0rc0
numpy	1.21.5
pandas	1.3.5
tensorflow	2.8.0
[...]	

The architecture used was x86_64, 8 CPUs, 4 cores per socked, 1 socket, CPU was an Intel(R) Xeon(R) CPU @ 2.20GHz, L1 cache 32K, L2 cache 256K, L3 cache 56320K. To train and run the inference I used a GPU NVIDIA Corporation TU104GL [Tesla T4] (rev a1).

Appendix E

Relative Error on HRV features induced by interpolation methods

In Table E.1, a comparison of a selection of outcomes of the experiment delineated in Chapter 3 are presented. The Relative Error is documented for every HRV feature across a selection of interpolation techniques and for each proportion of missing values within the IBI time series.

Interpolation method	pnn50	RMSSD	SD1	SD2	SDNN	VLF	LF	HF	Total Power
30% of NA no-interpolation	24.37	14.65	14.68	37.56	22.23	21.22	48.37	53.18	25.22
30% of NA quadratic on DURATION	41.80	19.34	19.37	62.46	42.68	30.41	33.52	62.26	30.50
30% of NA quadratic on TIME	32.07	15.05	15.08	8.57	9.42	14.50	26.87	32.18	17.16
50% of NA no-interpolation	39.01	23.13	23.18	50.64	31.07	30.23	58.79	69.14	36.90
50% of NA quadratic on DURATION	65.16	31.61	31.50	84.26	75.88	42.38	60.73	77.21	46.58
50% of NA quadratic on TIME	53.20	25.70	25.61	13.49	15.47	23.72	42.42	52.56	27.59
70% of NA no-interpolation	137.37	34.37	34.46	55.23	74.88	65.47	60.59	107.78	171.52
70% of NA quadratic on DURATION	107.01	46.59	45.81	94.45	127.46	68.23	68.58	106.93	104.19
70% of NA quadratic on TIME	89.09	39.91	39.24	19.19	22.76	29.73	56.41	72.98	72.07

Table E.1: Relative Errors of a selection of interpolation methods, for each HRV feature, at each level of missing heartbeats.

Appendix F

Survival Analysis for Health Outcomes

In this appendix, I elaborate on the details of a potential extension of my work from a framework where the predicted health outcome is encoded as a binary variable to a survival analysis setting, in which the model's output is a risk estimation.

The label for the models could be encoded as a tensor y with shape $N \times 2$, where N is the batch size, and the last dimension holds 2 values for each sample: a binary value that indicates if the subject had the event or not (to indicate censoring), and the other value indicates after how many years the subject had the event or dropped out the study (i.e. last follow up date, or died for other causes). I could use y_e for the tensor that indicates if the subject had the event (it only assumes values 0 or 1), and y_t to indicate the predicted time to event.

The output of the model would be a tensor $\tilde{y}$ with shape Nx1, with a value for each element that indicates the time to event.

Loss functions

I could assess the quality of my models using concordance index. The following three loss functions might be used to train the model:

- Event loss $\mathcal{L}_e$
- Time to event loss $\mathcal{L}_t$
- Ranking loss $\mathcal{L}_r$

And the overall loss defined as

$$\mathcal{L} = k_e\mathcal{L}_e + k_t\mathcal{L}_t + k_r\mathcal{L}_r \tag{F.1}$$

where k_e, k_t, and k_r are parameters fitted by the hyper-parameter search.

Event loss $\mathcal{L}_e$ This loss focuses on the binary label that is set to 1 if the subject had the event.

I could use the predicted time to event $\hat{y}$ to estimate the estimated probability for each subject to have the event, as shown in equation F.2 where σ_e is a scaling parameter that is fitted during hyper-parameter search.

$$p_e = 1 - \tanh\left(\sigma_e \tilde{y}\right) \tag{F.2}$$

Figure F.1 shows the effect of σ_e on p_e. it can be seen that p_e is bound between 0 and 1, with values close to 0 (indicating low probability of event) when the model predicts a large time to event $\tilde{y}$, and values close to 1 (indicating high probability of event) when time to event $\tilde{y}$ is close to zero.

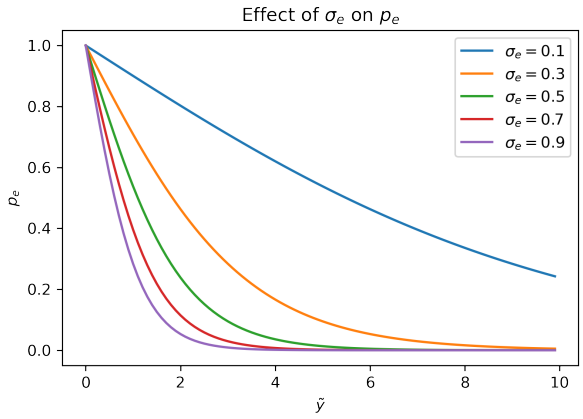


Figure F.1: The effect of σ_e on p_e. The time to event predictions $\tilde{y}$ (years) are mapped to probabilities p_e in different ways: small values of σ_e make p_e to to zero slowly (large $\tilde{y}$ are required for p_e to reach zero; high values of σ_e make p_e go to zero for small values of $\tilde{y}$

The event loss $\mathcal{L}_e$ is then defined as the binary cross entropy between p_e and y_e, where y_e is the label that indicates if the subject had the event, in this loss function the label that indicates the time to event is ignored.

This loss function assumes that most subjects without events have a time to event larger than subjects with events.

Time to event loss $\mathcal{L}_t$ As shown in equation F.3, $\mathcal{L}_t$ is simply a masked version of Root Mean Square Error on the predicted time to event, with a mask that filters out subjects that did not have the event. $\mathcal{L}_t$ therefore only considers subject who had the event and tries to force the model to guess the right amount of years that passed between assessment (time zero) and the actual event (y). The values predicted for subjects who did not have the event are simply ignored.

$$\mathcal{L}_t = \sqrt{\frac{\sum y_e (y_t - \tilde{y})^2}{\sum y_e}} \tag{F.3}$$

If a batch has no subjects with events (i.e. $\sum y_e = 0$) I set $\mathcal{L}_t = 0$.

Ranking loss $\mathcal{L}_r$ This loss function is inspired by the $\mathcal{L}_2$ loss function defined in [115]: I use the same idea to approximate concordance index without requiring the full computation (that would have complexity $\mathcal{O}(N^2)$, therefore impractical for large number of subjects). $\mathcal{L}_r$ is shown in equation F.4.

$$\mathcal{L}_r = \sum_{i \neq j} \mathbb{1}_{ij} \mu(\tilde{y}_i, \tilde{y}_j) \tag{F.4}$$

Where:

- $\tilde{y}_i$ and $\tilde{y}_j$ are the predicted time to event for subject i and j;
- $\mathbb{1}_{ij}$ is the indicator function for acceptable subjects pairs (i, j) (a pair is acceptable is subject i had the event, and subject j either had the event after i or didn't have the event until time y_{ti}, that indicates when subject i had the event);
- μ is defined in equation F.5;
- σ_r is a parameter fitted by hyper-parameter search.

$$\mu(x_1, x_2) = \exp\left(\sigma_r(x_1 - x_2)\right) \tag{F.5}$$

The indicator function $\mathbb{1}_{ij}$ ensures only the acceptable pairs are considered, for which is desirable for the loss function to return a low value if $\tilde{y}_i < \tilde{y}_j$ (the model estimated a lower time to event for subject i) and a large value if $\tilde{y}_i > \tilde{y}_j$ (the model estimated a lower time to event for subject j).

Figure F.2 shows the behaviour of $\mathcal{L}_r$ as a function of the difference in the estimated time to event values for subject pairs (i, j). The loss function $\mathcal{L}_r$ ignores non-acceptable pairs where subject i did not have an event, or where the actual time to event for i was larger than the time to event of j. Is desirable for $\mathcal{L}_r$ to penalise predictions where $\tilde{y}_i > \tilde{y}_j$ (our model predicted a larger time to event for i). It can be seen that μ accomplishes this task by assigning zero loss when $\tilde{y}_i < \tilde{y}_j$ and increasingly large values as $\tilde{y}_i$ grows larger compared to $\tilde{y}_j$. σ_r has the effect of setting how quickly wrong predictions of time to event produce large values of $\mathcal{L}_r$. Minimising $\mathcal{L}_r$ has the effect of maximising the number of times the ranking between acceptable pairs is estimated correctly (i.e concordance index).

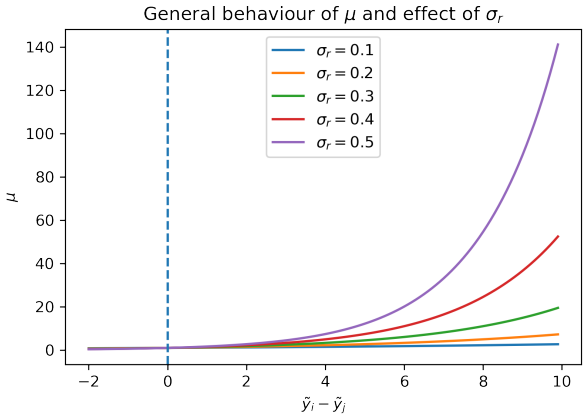


Figure F.2: The effect of σ_r on $\mathcal{L}_r$ and the general behaviour of $\mathcal{L}_r$ as a function of the difference between the estimated time to events for subject pairs (i, j)

.