

Predicting Heart Failure Deterioration



Johanna Elsa-Marie O'Donnell

St Anne's College

Supervised by

Professor Lionel Tarassenko,

Professor Kazem Rahimi

Submitted: Michaelmas Term, 2017

This thesis is submitted to the Department of Engineering Science, University of
Oxford, in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Declaration

I declare that this thesis is entirely my own work, and except where otherwise stated, describes my own research.

J. E.-M. O'Donnell,
St Anne's College

To Mama. To Papa. To Tom.

Predicting Heart Failure Deterioration

Johanna E.-M. O'Donnell

Thesis submitted for the degree of Doctor of Philosophy

St Anne's College, Michaelmas Term 2017

Abstract

Chronic heart failure (HF) is a condition that affects more than 900,000 people in the UK [1]. Mortality rates associated with the condition are high, with nearly 20% of patients dying within one year of diagnosis [2]. Continuous monitoring and risk stratification can help identify patients at risk of deterioration and may consequently improve patients' likelihood of survival [3]. Current repeated-measure risk stratification techniques for HF patients often rely on subjective perception of symptoms, such as breathlessness, and markers of fluid retention in the body (e.g. weight) [4]. Despite the common use of such markers, studies have shown that they offer limited effectiveness in predicting HF-related events [5].

This thesis set out to identify and evaluate new markers for repeated-measure risk stratification of HF patients. It started with an exploration of traditional HF measurements, including weight, blood pressure, heart rate and symptom scores, and aimed to improve the performance of these measurements using a data-driven approach. A multi-variate model was developed from data acquired during a randomised controlled trial of remotely-monitored HF patients. The rare occurrence of HF-related adverse events during the trial required the development of a careful methodology. This methodology helped identify the markers with most predictive ability, which achieved moderate performance at identifying patients at risk of HF-related adverse events, clearly outperforming commonly-used thresholds.

Subsequently, this thesis explored the potential value of additional, accelerometer-derived physical activity (PA) and sleep markers. For this purpose, the ability of accelerometer-derived markers to differentiate between individuals with and without HF was evaluated. It was found that markers that summarise the frequency and duration of different PA intensities performed best at differentiating between the two groups and may therefore be most suitable for future use in repeated-measure applications.

As part of the analysis of accelerometer-derived HF markers, a gap in the methodology of automated accelerometer processing was identified, namely the need for self-reported sleep-onset and wake-up information. As a result, Chapter 5 of this thesis describes the development and evaluation of a data-driven solution for this problem.

In summary, this thesis explored both traditional and new, accelerometer-derived markers for the early detection of HF deterioration. It utilised sound methodology to overcome limitations faced by sparse and unbalanced datasets and filled a methodological gap in the processing of signals from wrist-worn accelerometers.

Acknowledgements

This thesis would not have been possible without the help of so many people. Firstly, I would like to express my sincere gratitude to my supervisors Prof. Lionel Tarassenko and Prof. Kazem Rahimi for their continuous support and guidance. Their experience, perspective and scientific rigour have been invaluable to the completion of this work.

I have been lucky to have been supported in my work by many brilliant people. I would like to acknowledge the contribution of Dr. Aiden Doherty, whose mentorship has significantly shaped my DPhil experience. I would also like to thank Dr. Carmelo Velardo and Dr. Gholamreza Salimi-Khorshidi for their supportive guidance, much-appreciated advice and time. Thanks also go to Dr. Dario Salvio, Dr. Ahmar Shah and Dr. Andreas Triantafyllidis.

I would like to acknowledge my colleagues at the Institute of Biomedical Engineering and at the George Institute for Global Health. In particular, I would like to thank Farah Colchester, Hamish Tomilson, Dr. Rebecca Pullon, Dr. Kate Niehaus, Drew Birrenkott, Dr. Lise Loerup, Sarah-Jane Rodgers, Farah Shamout, Rasheed El-Bouri, Elina Naydenova, Hasmila Omar, Thais Roque, Nathalie Conrad, Moira Allison, Dr. Karl Smith-Byrne, Dr. Elizabeth Millett, Dr. Annelien de Kat, Dr. Jenny Tran, Dr. Shobana Nagraj, and Debbie Hedgecote. I would also like to thank Jo Armitage, Caroline Brown, Debbie McCusker and Naseem Akhtar.

I would like to acknowledge the financial support of the RCUK Digital Economy Programme and the much appreciated contribution of my examiners, Prof. John Cleland and Prof. Maarten de Vos.

Completing a DPhil can be an intense and at times challenging experience. Especially in difficult times, I have always felt the love and support from my family and friends. I would like to especially thank my parents, my grandparents, my sister Fee and my brother Fritz, who have been a source of strength and have (literally) travelled great lengths to support me. I would also like to thank Heidi Dritschel and Sarah Reeves for making Oxford a more homely place to study. Most of all, I have felt the support from my husband Tom, without whom none of this would have been possible; during the last four years, you have made many sacrifices to allow me to thrive. Thank you.

Contents

1	Background	1
1.1	Heart Failure	1
1.1.1	Diagnosis	3
1.1.2	Symptoms	4
1.1.3	Treatments	5
1.1.3.1	Medication	5
1.1.3.2	Cardiac Devices	5
1.1.4	Adverse Events	6
1.2	Heart Failure Management	6
1.2.1	Single-measure Risk Stratification	7
1.2.2	Repeated-measure Risk Stratification	8
1.2.2.1	Weight	8
1.2.2.2	Symptoms	9
1.2.2.3	Blood Pressure	10
1.2.2.4	Other Markers	10
1.2.3	Physical Activity as a Heart Failure Marker	11
1.2.3.1	Physical Activity Patterns of Heart Failure Patients . . .	11
1.2.3.2	Physical Activity as a Marker for Single-measure Risk Stratification	12
1.2.3.3	Physical Activity as a Marker for Repeated-measure Risk Stratification	13
1.2.3.4	Lessons from Other Disease Areas	14
1.2.4	Sleep as a Heart Failure Marker	15
1.2.5	Motivation for Thesis	15
2	Datasets	17
2.1	Introduction	17
2.2	Heart Failure Tele-monitoring Data (SUPPORT-HF)	18
2.2.1	The SUPPORT-HF System	18

2.2.1.1	Devices	19
2.2.1.2	Daily Monitoring	20
2.2.1.3	Data Transfer	21
2.2.2	SUPPORT-HF2 Study	22
2.2.2.1	Study Population	22
2.2.2.2	Data Overview	24
2.3	Labelled Actigraphy Dataset (Cognitive Health in Ageing Exercise Study)	27
2.3.1	Study Population	28
2.3.2	Devices	28
2.3.3	Data Overview	30
2.4	Heart Failure Accelerometer Dataset (UK Biobank)	31
2.4.1	Study Population	31
2.4.2	Data Overview	33
2.5	Summary	34
3	Methods	35
3.1	Introduction	35
3.2	Probability Density Functions	36
3.3	Logistic Classification	37
3.4	Random Forest Classifiers	38
3.5	Change-point Detection	40
3.6	Evaluation Tools	41
3.6.1	Statistical Tests	41
3.6.2	Accuracy, Sensitivity and Specificity	42
3.6.3	Kappa Statistic	43
3.6.4	Receiver Operating Curves	44
3.7	Summary	45
4	Repeated-Measurement Risk Prediction using Traditional Heart Failure Markers	46
4.1	Introduction	46
4.2	Background	47
4.2.1	The Predictive Value of Physiological Measurements and Symptom Scores	47
4.2.1.1	Weight Measurements	47
4.2.1.2	Blood Pressure Measurements	50
4.2.1.3	Heart Rate Measurements	53
4.2.1.4	NYHA Classification Scores	54
4.2.1.5	Early Warning Systems	55

4.3	Methodology	58
4.3.1	Database	58
4.3.2	Data Preparation	59
4.3.2.1	Data Cleaning	59
4.3.2.2	Missing Value Imputation	60
4.3.2.3	Variable Extraction	61
4.3.2.4	Adverse Events	64
4.3.3	Data Analysis	64
4.3.3.1	Data Exploration	64
4.3.3.2	Logistic Classifiers	67
4.3.3.3	Evidence-based Thresholds	68
4.3.4	Evaluation	69
4.4	Results	69
4.4.1	Dataset	69
4.4.2	Alerts	69
4.4.3	Typical Results	70
4.4.4	Adverse Events	73
4.4.5	Data Analysis	75
4.5	Discussion	86
4.6	Conclusion	87
5	Automated Sleep-boundary Detection using Wrist-worn Accelerometers	88
5.1	Introduction	88
5.2	Background	89
5.2.1	Accelerometer-derived Sleep Analysis	89
5.3	Methodology	94
5.3.1	Database	95
5.3.2	Data Preparation	97
5.3.2.1	Calibration	97
5.3.2.2	Wear-time Detection	98
5.3.3	Data Analysis	99
5.3.3.1	Sleep Detection using a Data-driven Threshold	101
5.3.3.2	Sleep Detection using a Random Forest Classifier	103
5.3.3.3	Sleep Detection using Change-point Detection	105
5.3.3.4	Post-processing using a Smoothing Filter	107
5.3.4	Evaluation	108
5.4	Results	110

5.4.1	Data Collection	110
5.4.2	Data Preparation	111
5.4.2.1	Calibration	111
5.4.2.2	Wear-time Detection	111
5.4.3	Data Analysis	113
5.4.3.1	Sleep Detection using Data-driven Thresholds	113
5.4.3.2	Sleep Detection using a Random Forest Classifier	119
5.4.3.3	Sleep Detection using Change-point Detection	125
5.4.4	Evaluation	129
5.5	Discussion	130
5.6	Conclusion	133
6	Heart Failure Classification using Wrist-worn Accelerometer Data	135
6.1	Introduction	135
6.2	Background	136
6.2.1	Accelerometer-derived Markers for Physical Activity and Sleep	136
6.3	Methodology	143
6.3.1	Study Population	143
6.3.2	Data Preparation – Self-reported Physical Activity Data	144
6.3.3	Data Preparation – Accelerometer Data	146
6.3.4	Data Preparation – Hospital Admission Data	150
6.3.5	Statistical Analysis	150
6.4	Results	153
6.4.1	Baseline Analysis	154
6.4.2	Accelerometer Analysis	161
6.5	Discussion	169
6.6	Conclusion	170
7	Conclusion and Future Work	171
7.1	Repeated-Measure Risk Stratification using Traditional Heart Failure Markers	171
7.2	Heart Failure Classification using Wrist-worn Accelerometer Data	172
7.3	Automated Sleep-boundary Detection using Wrist-worn Accelerometers	174
7.4	SUPPORT-HF3: Proposal for the Next Stage of the Study	175
	References	178

List of Figures

1.1	United Kingdom heart failure incidents in 2002 and 2014.	2
1.2	Heart failure and the ejection fraction.	3
2.1	Components of the SUPPORT-HF system.	19
2.2	Functionality of the SUPPORT-HF system.	21
2.3	The SUPPORT-HF data transfer.	22
2.4	The SUPPORT-HF timeline from recruitment to passive follow-up. . . .	24
2.5	Summary statistics for the SUPPORT-HF2 study.	27
2.6	The Cognitive Health in Ageing Exercise study timeline.	28
2.7	Accelerometer monitors relevant to this thesis.	29
2.8	Sleep diary used to collect sleep-onset and wake-up information.	30
2.9	Timeline of UK Biobank data collection.	32
3.1	Visualisation of the maximum likelihood estimation.	41
3.2	Visualisation of the receiver operating curve.	44
4.1	Visual representation of the repeated-filtering method.	57
4.2	Visual representation of missing value imputation.	60
4.3	Visual representation of the three-day change marker.	61
4.4	Visual representation of the ' <i>scaled</i> ' marker extraction.	63
4.5	Visual representation of adverse event prediction.	65
4.6	Flow-chart of the risk prediction methods.	66
4.7	Visualisation of the importance of training and test set sizes.	67
4.8	Heart rate values of patient 159 before and after medication change. . . .	71
4.9	Weight values for patient 179 before and after medication change.	71
4.10	Receiver operating curves across different partitionings.	78
4.11	Beta coefficients for the classifiers using $\delta\mathbf{x}$ and $\sigma\mathbf{x}$ markers as inputs. . .	80
4.12	Performance of the combined classifier.	81
4.13	Performance of guideline-based model.	83
4.14	Performances of models when excluding adverse events classed as ' <i>other</i> '. .	85

5.1	Sleep diary used to collect sleep-onset and wake-up information.	95
5.2	Flow diagram of the sleep-boundary detection methods.	96
5.3	Visualisation of the PDF approach for sleep-boundary detection.	100
5.4	Accelerometer signal for a typical subject from the CHA Exercise study. .	101
5.5	Parameters for the PDF method for sleep-boundary detection.	103
5.6	Flow diagram summarising random forest training and testing steps. . . .	105
5.7	Visual representation of the change-point-detection method.	107
5.8	Visual representation of the Bland-Altman plot.	109
5.9	Distribution of self-reported sleep-onset and wake-up times.	110
5.10	Calibration offset and slope values.	111
5.11	Accelerometer days with self-reported sleep-boundary information.	112
5.12	Sleep and wake periods for a typical subject from the CHA Exercise study.	114
5.13	Parameters for PDF-based method.	115
5.14	Self-reported and algorithm-derived sleep-boundary times (PDF).	117
5.15	Bland-Altman plots (PDF).	118
5.16	Parameters for RF-based sleep-boundary-detection.	120
5.17	RF feature importance.	120
5.18	Evaluation of the effect of large number of input features.	121
5.19	Self-reported and algorithm-derived sleep-boundary times (RF classifier).	123
5.20	Bland-Altman plots (RF classifier).	124
5.21	Parameters for the CPD based sleep-boundary-detection.	125
5.22	Self-reported and algorithm-derived sleep-boundary times (CPD).	127
5.23	Bland-Altman plot (CPD).	128
6.1	Visual representation of the Rowland's method.	141
6.2	Timeline of UK Biobank data collection.	145
6.3	Flow diagram of the accelerometer feature extraction.	147
6.4	Flow diagram showing the training and testing of accelerometer-derived markers.	152
6.5	Age distributions by sex before and after control-matching in the baseline analysis.	156
6.6	Feature selection for the classification using only baseline markers.	157
6.7	Frequency and rank of markers selected for baseline analysis.	158
6.8	Performance of the classifier using only baseline markers.	159
6.9	PDFs showing the performance of the classifier using only baseline markers.	160
6.10	Age distributions by sex before and after control-matching in the ac- celerometer analysis.	162
6.11	Feature selection for the accelerometer classifier.	163

6.12	Frequency and rank of markers selected for accelerometer analysis.	165
6.13	Performance of the accelerometer classifier.	166
6.14	PDFs of the accelerometer classifier.	167

List of Tables

1.1	Definitions of New York Heart Association (NYHA) functional classes. . .	10
2.1	SUPPORT-HF baseline characteristics.	25
2.2	CHA baseline characteristics.	31
2.3	Summary information for the UK Biobank dataset.	33
3.1	Definitions of true and false negatives and positives.	42
4.1	Evidence for the predictive value of weight-based markers.	48
4.2	Evidence for the predictive value of blood pressure markers.	53
4.3	Evidence for the predictive value of heart rate markers.	55
4.4	Evidence for the predictive value of NYHA markers.	55
4.5	Types and severity of adverse events in the intervention group.	72
4.6	Types and severity of adverse events in the control group.	73
4.7	SUPPORT-HF2 baseline characteristics (stratified by adverse event). . .	74
4.8	Types and severity of adverse events in the multi-variate analysis.	75
4.9	Performance across different dataset partitionings.	77
4.10	Performance using three-day and seven-day markers.	79
5.1	Overview of sleep-detection algorithms.	91
5.2	Features for the random forest classifier.	102
5.3	Example of a confusion matrix	109
5.4	Accelerometer summary statistic for subjects in the CHA Exercise study.	113
5.5	Confusion matrix (PDF).	116
5.6	Confusion matrix (RF classifier).	122
5.7	Confusion matrix (CPD).	126
5.8	Summary of performances of sleep-boundary detection methods.	130
6.1	Overview of algorithms for accelerometer data.	138
6.2	Overview of intensity-scoring thresholds for the GENEActiv.	140
6.3	List of markers included in the UK Biobank analysis.	149

6.4	Baseline information for UK Biobank participants.	155
6.5	Performance of the classifier with self-reported physical activity data. . .	159
6.6	Confusion matrix for the classifier using only baseline data.	160
6.7	Confusion matrix for the classifier with self-reported physical activity data.	160
6.8	Summary of the performances of all classifiers.	164
6.9	Confusion matrix for the accelerometer analysis using established markers.	168
6.10	Confusion matrix for the accelerometer analysis using accelerometer-derived markers.	168
6.11	Confusion matrix for the accelerometer analysis using the combined markers.	168

List of Abbreviations

6MWT	Six-minute walk test
ACE	Angiotensin-converting enzyme
ARB	Angiotensin receptor blocker
AUC	Area under the receiver operating curve
BMI	Body mass index
BNP	Brain natriuretic peptide
bpm	Beats per minute
CHA	Cognitive Health in Ageing
CKD	Chronic kidney disease
COPD	Chronic obstructive pulmonary disease
CPD	Change-point detection
CRD	Cardiac-resynchronisation device
DWT	Discrete wavelet transform
EF	Ejection fraction
ESC	European Society of Cardiology
FDA	Food and Drug Administrative
FFT	Fast Fourier transform
HF	Heart failure
HR	Hazard ratio

IQR	Interquartile range
ICD	Implantable cardioverter defibrillator
MET	Metabolic equivalent of task
HRV	Heart rate variability
OR	Odds ratio
PA	Physical activity
PDF	Probability density function
PSG	Polysomnography
REC	Research ethics committee
RF	Random forest
ROC	Receiver operating curve
SHF	SUPPORT-HF
UKB	UK Biobank

Background

1.1 Heart Failure

Heart failure (HF) is a common and debilitating condition with rising prevalence due to an ageing population [1][6]. It is believed to affect more than 900,000 individuals living in the UK [1] and consume up to 2% of the National Health Service's budget [3][7]. Between 2002 and 2014 the number of new HF diagnoses in the UK grew by 12% from an estimated total of 170,727 to 190,798. This rise is believed to be largely driven by an increase in population size and age [1]. Figure 1.1, based on a recent study on HF patients in the UK, illustrates the age distribution of HF incidence in the UK. Most HF cases occur in individuals over the age of 60, with a mean HF population age of 77 years [1]. As life expectancy increases, the incidence of HF continues to rise. Other factors leading to the rise in HF cases include improved medical treatments leading to increased survival rates after adverse events such as myocardial infarctions [8].

Clinically, HF is defined as a syndrome characterised by symptoms (such as breathlessness, swelling of the ankles) that results in reduced cardiac output and/or increased intracardiac pressure [9]. As a result, the heart is no longer able to pump sufficient blood around the body to supply its cells with oxygen [10]. HF can have a number of functional and structural causes, including amongst others damaged heart tissue due to

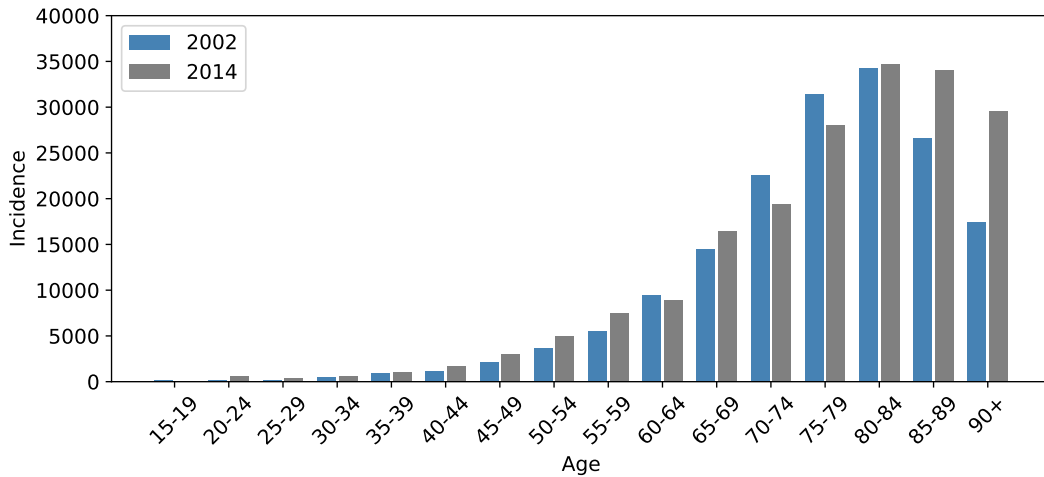
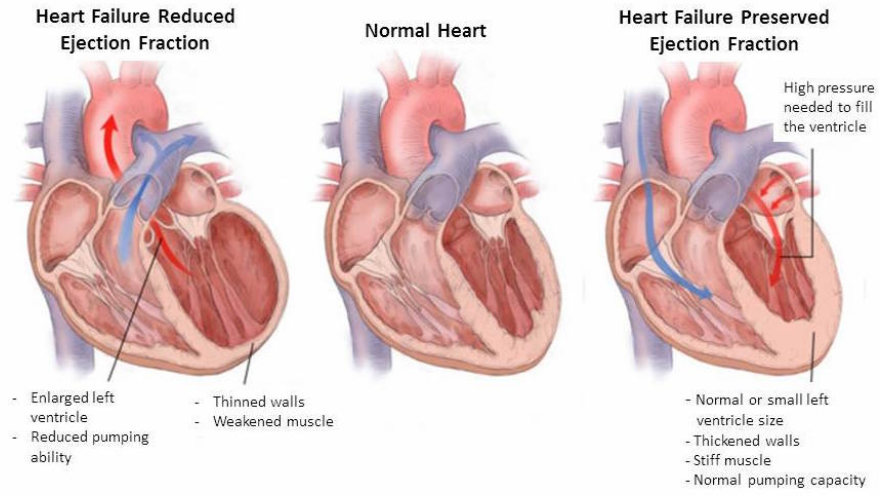


Figure 1.1: *United Kingdom heart failure incidence in 2002 and 2014 stratified by age [1].*

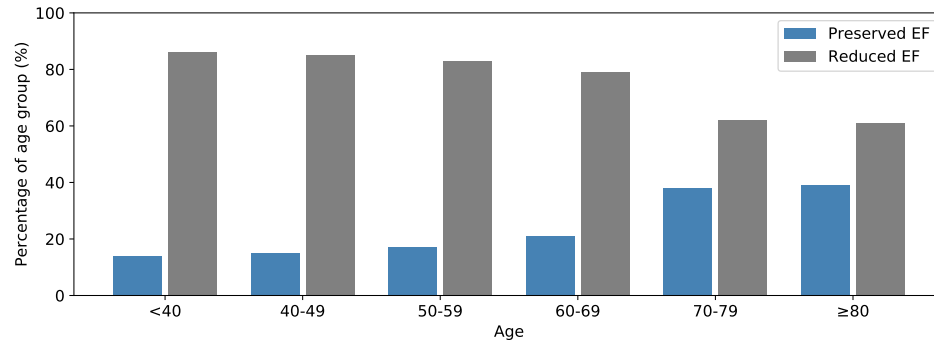
myocardial infarction (heart attack), high blood pressure or coronary heart disease [10]. In younger age groups, viruses causing an inflammation of the heart muscle or dilated cardiomyopathy (an enlarged and weakened left ventricle) are amongst the prevalent causes of HF [11].

HF is commonly divided into HF with reduced ejection fraction (EF), also known as left ventricular HF, and HF with preserved EF, where the EF is a measure of the amount of blood that is pumped out of the left ventricle as a percentage of the amount of blood entering the ventricle (see Figure 1.2a). In a healthy adult, the EF typically ranges between 50% and 70%. HF patients with reduced EF have EF values of less than 40%. In patients with preserved EF the EF is greater than 40%; however stiffness or damage of the heart means that the heart of these patients does not relax properly, leading to a reduced filling of the ventricles during the relaxation phase of the heartbeat [9]. Figure 1.2b shows the percentage of patients who participated in a cross-European HF study with reduced and preserved EF, stratified by age [12]. Especially in younger HF patients, HF with reduced EF is more common.

In the forthcoming sections an overview of HF diagnosis, symptoms, treatments and adverse events will be provided. This overview will set the context in which the analyses



(a) Visual representation of the heart's structure in patients with heart failure with reduced (left) and preserved (right) ejection fraction. Adapted from [13]



(b) Percentage of heart failure patients with reduced and preserved ejection fraction (EF), stratified by age. Based on 41,926 patients studied as part of the MAGGIC study [12].

Figure 1.2: Heart failure and the ejection fraction.

described in this thesis were carried out. It will also demonstrate the need for accurate and unobtrusive repeated-measure monitoring systems.

1.1.1 Diagnosis

An initial HF diagnosis is based on a detailed assessment of a patient's symptoms and their history of cardiovascular disease [9]. In non-emergency settings, the European

Society of Cardiology (ESC) recommends that patients suspected of suffering from HF undergo a screening test, whereby the plasma concentration of natriuretic peptides in their blood is measured. Natriuretic peptides are short chains of amino acids that trigger the kidney to extract sodium. A certain type of peptide, brain natriuretic peptide (BNP), is produced as a result of pressure changes in the heart and its concentration is raised in patients with HF. If the screening test finds that a patient's BNP concentration is above a given threshold, an echocardiographic examination (ultrasound examination of the heart) is recommended. The echocardiogram is the gold standard for the diagnosis of HF [9]. Using advanced ultrasound techniques, the ejection fraction of the heart can be determined, as well as abnormalities in the structure and function of the heart.

1.1.2 Symptoms

HF patients may experience a range of symptoms, including amongst others breathlessness, fatigue and swelling [14]. As the kidneys are no longer supplied with sufficient blood, they start producing hormones that trigger fluid and salt retention in the body. This in turn leads to swelling (oedema) particularly in the legs, ankles and feet of patients. The increased levels of sodium and other substances in the blood can also lead to an impaired mental state. As blood is no longer pumped through the body efficiently, fluid may be pushed into the lungs, which may cause breathlessness and coughing [14].

As a result of the reduced blood supply in the body, important parts of the body, such as the heart and the brain, may be prioritised leading to limited blood flow in areas such as the legs and arms. This may cause fatigue in patients with HF. Patients may also experience increased blood pressure and heart rate as the body tries to compensate for the reduced blood flow [14].

1.1.3 Treatments

1.1.3.1 Medication

For most people, HF is considered a chronic condition and as such there is no cure for it. Instead, a combination of medication and lifestyle interventions are recommended to mitigate symptoms such as swelling or breathlessness and reduce the long-term likelihood of disease deterioration [9]. Commonly prescribed HF medications include diuretics, angiotensin-converting enzyme inhibitors, angiotensin receptor blockers and beta-blockers.

Diuretics target the problem of fluid retention in the body and can offer short-term relief from HF-related water retention in the body [14]. They instigate the kidneys to produce more urine resulting in a reduction of swelling, breathlessness and fatigue.

Angiotensin-converting enzyme inhibitors (ACE inhibitors) and angiotensin receptor blockers (ARBs) on the other hand allow blood vessels to relax reducing blood pressure in the body and therefore the strain on the heart, in turn increasing the EF of HF patients [14]. Similarly, beta-blockers allow the heart to beat more slowly, resulting in a reduction in the strain on the heart. Both ACE inhibitors [15] and beta-blockers [16] have been shown to reduce the long-term risk of mortality in HF patients with reduced EF. For patients with preserved EF, angiotensin-targeting medications and beta-blockers have shown limited effectiveness [17].

1.1.3.2 Cardiac Devices

Some HF patients are affected by unevenly dispersing electrical signals across the heart, which can reduce the effectiveness of the heart and also be a cause of sudden death [9]. Cardiac-resynchronisation devices (CRDs) can be implanted to control the beating of the heart. In some cases these are also fitted with automatic defibrillators.

1.1.4 Adverse Events

Despite existing HF treatments that can control symptoms and prevent long-term HF deterioration, HF is a complex disease to manage and HF-related hospitalisation and mortality rates remain high [2][18]. Every year, nearly one in three patients diagnosed with HF experiences at least one hospitalisation [18]. According to a recent study of 54,000 UK HF patients, 19% of HF patients die within one year of diagnosis and almost 50% of HF patients die within five years of diagnosis [2].

Decompensated HF is defined as an acute worsening of HF-related symptoms and is a major cause of HF-related hospitalisations and deaths [19]. Fluid accumulation in the lungs or other tissues of the body can result in extreme breathlessness and/or swelling [14]. Acute decompensation can for example be caused by non-compliance with treatment or by arrhythmias and ischemias. Other causes of HF-related hospitalisations may include renal dysfunction (kidney failure) or dizziness and falls.

In addition to decompensated HF, sudden cardiac death is another common cause of mortality for HF patients [14][19]. Acute irregularities in the heart beat (e.g. tachycardia), which lead to a break-down of the signalling pattern of the heart, are believed to be a primary cause for sudden cardiac death in HF patients [14].

1.2 Heart Failure Management

Appropriate patient management can improve quality of life and reduce a patient's likelihood of experiencing adverse events [3]. The National Institute for Health and Care Excellence (NICE) therefore advises regular consultations with HF specialists. In order to determine the frequency and intensity of visits/monitoring required, single-contact risk stratification methods can be employed. To obtain a more granular picture of patient well-being over time, particularly in patients with severe HF, repeated risk assessment

(e.g. through daily measurements) can be used. The following section provides a brief overview of some of the commonly used HF risk-assessment methods. It starts with an introduction to single-contact risk assessment methods and progresses to describe ways to monitor HF progression more closely over time.

1.2.1 Single-measure Risk Stratification

Risk stratification is the process of assigning patients to risk groups based on factors such as their clinical characteristics and demographics. This may help inform treatment decisions and provide clarity to patients and their families [20]. A number of risk stratification algorithms have been developed to date and these vary with respect to the outcome they are trying to predict (e.g. hospitalisation, death) and the variables that they include [20].

Single-contact scores include algorithms such as the Get With the Guidelines-Heart Failure (GWTG-HF) score and the Meta-analysis Global Group in Chronic HF (MAGGIC) score [21][22]. The GWTG-HF score uses seven demographic, medical history and laboratory variables to predict in-hospital mortality in patients hospitalised with HF [21]. The MAGGIC score uses 13 variables as a predictor of one- or three-year mortality in patients with HF [22]. According to the authors, patients assigned to the highest decile of the three-year mortality score are at a 70% risk of death, whereas patients assigned to the lowest quintile are at a 10% risk of death. Validation on an external dataset showed the model achieved a C index of 0.741 [23]. Rahimi *et al.* conducted a systematic review of 64 risk prediction models in patients with HF and identified variables that were strong predictors of death [20]. These included, amongst others, age, renal function, blood pressure and ejection fraction.

1.2.2 Repeated-measure Risk Stratification

The purpose of repeated-measure risk stratification is to identify patients at risk of deterioration early and to optimise patient care accordingly [4]. In HF patients, this can be achieved by

- Adjusting a patient’s fluid balance through diuretic medication management,
- Detecting and acting on potential causes for HF adverse events, such as heart rhythm abnormalities, hypertension or infections,
- Optimising medical therapy to improve quality of life and reduce likelihood of adverse events.

For this purpose, patients are asked to take regular readings of a variety of physiological or symptom markers. These readings can be stored in a number of different ways, ranging from hand-written diaries to purpose-built tele-monitoring systems. Once recorded, readings are analysed to identify any abnormalities that may suggest patient deterioration (e.g. weight increases suggesting fluid retention in the body) or opportunities to improve a patient’s long-term likelihood outcomes (e.g. ACE inhibitor titration) [4]. Traditionally, this is done through review of patient data by a healthcare professional. More recently, automated stratification algorithms (or early-warning scores) have been developed to identify and flag patients at risk of deterioration [5]. Algorithms may be based on different types of physiological measurements, such as weight, blood pressure or other measurements. The following paragraphs provide an initial overview of measurements used within repeated-measure systems for HF patients as well as the evidence behind them. A more detailed review of some of the markers will be provided at the beginning of Chapter 4.

1.2.2.1 Weight

Many measurements used within repeated-measure risk-stratification systems for HF patients aim to capture fluid retention in the body [24]. Increases in weight may indicate

fluid retention and require adjustment of medication. The ESC has previously advised to carefully monitor weight changes and to respond to sudden weight gain [25]. In particular, they recommend that weight changes of more than 2kg within a period of three days should be reported to a healthcare professional. Remote monitoring trials have implemented absolute weight change thresholds as recommended by the ESC and studied their ability to predict patient deterioration [26]. In addition to this, other studies have considered the percentage change in weight and the standard deviation of weight over a given period as potential markers for HF deterioration [26][27][28].

1.2.2.2 Symptoms

Symptoms such as swelling in the legs and breathlessness can be further markers for fluid accumulation in the body. Other symptoms, such as chest pain or fatigue, may not be caused by fluid retention but still provide an important insight into patient well-being [29]. Worsening symptoms are indicators of disease deterioration and have been used as predictors in a number of HF home monitoring studies [5][30].

In particular, the New York Heart Association (NYHA) functional classification score [31] has been used as a marker for deterioration. The NYHA classification scores patients on a scale of 1-4 based on the amount of breathlessness they experience. Patients who experience no unusual amount of breathlessness during rest or activity are given a score of 1, patients who experience shortness of breath even at rest are given a NYHA score of 4 (see Table 1.1).

Despite the common use of symptoms as markers for disease worsening, concerns have been raised regarding the accuracy of symptom reports, which may be affected by cognitive ability and comorbidities [24].

NYHA Class	Description
I	No limitation of physical activity. Ordinary physical activity does not cause undue breathlessness, fatigue or palpitations.
II	Slight limitation of physical activity. Comfortable at rest, but ordinary physical activity does cause undue breathlessness, fatigue or palpitations.
III	Marked limitation of physical activity. Comfortable at rest, but less than ordinary physical activity does cause undue breathlessness, fatigue or palpitations.
IV	Unable to carry on any physical activity without discomfort. Symptoms at rest can be present. If any physical activity is undertaken, discomfort increases.

Table 1.1: *Definitions of New York Heart Association (NYHA) functional classes.*

1.2.2.3 Blood Pressure

Both high and low levels of blood pressure can be potential risk factors in HF patients. HF patients with high blood pressure are at increased risk of experiencing myocardial infarctions. Certain HF medications aim to dilate blood vessels and reduce blood pressure (ACEs, ARBs). If the blood pressure is reduced by too great an amount however, this can lead to dizziness and fainting. In order to ensure patients are not affected by extreme levels of blood pressure, the ESC advises to carefully monitor blood pressure [25].

1.2.2.4 Other Markers

In addition to the previously described markers, a number of other measurements have been proposed for repeated-measure risk stratification in HF patients. Intrathoracic electrical impedance measurements can detect fluid retention in the lungs (thorax) and have been found to perform moderately well at identifying HF deterioration up to 30 days early [32]. Heart rate variability (HRV) has also been considered as a marker for repeated-measure monitoring of HF patients [33]. Adamson *et al.* [33] studied the ability of HRV to predict HF-related hospitalisations. In their study of 397 patients with implantable devices they found that HRV was 70% sensitive in predicting hospitalisations with 2.4 false-positives per patient year of follow-up [33]. Other studies

have proposed the use of haemodynamic monitors to identify patient deterioration [24]. Haemodynamic monitors invasively measure arterial and venous blood pressures. In a multi-centre study by Abraham *et al.* [34], a tele-monitoring system based on implanted haemodynamic monitors proved to significantly reduce hospitalisations in HF patients with NYHA functional status of III. Despite the fact that such monitors have shown promising performances in predicting HF events, Desay and Konstam [5] suggest that it is unlikely that invasive haemodynamic devices will offer a scalable solution.

1.2.3 Physical Activity as a Heart Failure Marker

Current non-invasive repeated-measure risk-stratification techniques have been shown to offer limited accuracy in predicting HF-related adverse events [24]. The work discussed in this thesis will try to improve the accuracy of traditional markers as well as explore the predictive performance of additional non-invasive markers, namely physical activity (PA) and sleep patterns of HF patients.

The relationship between PA and HF has been the focus of a number of studies [35][36]. Whilst some analyses have been concerned with the relationship between PA and risk of developing HF later in life, there is also interest in understanding activity patterns of patients with diagnosed HF. Analysis of these activity patterns can offer insights into the impact of HF on a patient's daily living and may provide a marker for risk stratification [37]. The following paragraphs provide background information on PA patterns in HF patients and their potential role in repeated-measure risk stratification.

1.2.3.1 Physical Activity Patterns of Heart Failure Patients

Patients with chronic HF have been shown to be less physically active than healthy individuals of their age. Chessum *et al.* [36] conducted a study on 24 patients with HF and 24 healthy controls, in which patients were asked to self-report their levels of activity. Their results suggested that patients with HF spent on average 20-40% less

time in moderate to high intensity activity than healthy controls [36].

Even within the HF patient population, different levels of PA can be observed. Dontje *et al.* [38] measured PA in 68 HF patients using a wrist-worn accelerometer and found that higher NYHA classification and lower self-efficacy (dependent on confidence) led to reduced levels of PA. Their findings were supported by the outcomes of a study run by Braunschweig *et al.* [39], according to which, walking distance measured post device implantation with an accelerometer-fitted cardiac-resynchronisation therapy device was correlated with the NYHA classification of a patient. Other factors such as comorbidities, weather and encouragement may also affect the activity levels of HF patients [40].

1.2.3.2 Physical Activity as a Marker for Single-measure Risk Stratification

Device-captured PA offers an objective measure of patient well-being and may therefore be a suitable tool for single-measure risk stratification [41]. Activity data collected from implanted devices has repeatedly been proposed for this purpose: Conraads *et al.* [37] conducted a study on 781 HF patients with implantable cardioverter defibrillators (devices used to control irregular heart rhythms) and found that PA in the 30 days after device implantation was correlated with time to death in patients with HF as well as risk of re-hospitalisation. Their study used the average daily minutes above a set activity threshold as their risk marker [37]. Tyagi *et al.* [42] studied 62 patients with an implanted pacemaker and found that pacemaker-derived PA predicted all-cause mortality in their patient population. Patients who had on average ≥ 2 hours of >70 steps/minute activity per day were three times less likely to die within an eight-year follow-up [42].

Non-invasive accelerometers and pedometers have also been proposed as single-measure risk stratification markers. Walsh *et al.* [43] followed 84 HF patients for a mean period of 710 days. Patients were asked to wear two hip-worn pedometers for a period of one week. The study found that the mean weekly pedometer score was predictive of HF-

related deaths. Izawa *et al.* [44] found that average daily step counts over a period of eight days were a predictor of mortality in Japanese HF patients. Izawa *et al.* followed 170 HF outpatients over a mean period of four years. Patients with average daily step counts of $<4,889$ at the beginning of the study were twice as likely to die than patients with higher step counts. Rullman *et al.* [45] studied the predictive value of PA in patients with HF and found that peak activity variance was related to five-year mortality and re-hospitalisations. Activity was measured in 60 HF patients using a waist-worn accelerometer over a period of seven days. The predictive value of the accelerometer output was increased when including peak oxygen uptake and HF survival score in the prediction. The HF survival score rates cardiovascular performance based on a number of variables, including amongst others ejection fraction and blood pressure, and was originally developed to identify candidates suitable for cardio-transplantation [46].

1.2.3.3 Physical Activity as a Marker for Repeated-measure Risk Stratification

Limited work has been conducted on repeated-measure PA monitoring for HF patients [47]. Jehn *et al.* [48] integrated an at-home six-minute walk test (6MWT) into a tele-monitoring system. The 6MWT is a test to evaluate patient fitness and endurance and has been shown to be predictive of long-term outcomes of patients [49]. Patients were provided with a waist-worn accelerometer and asked to perform a monthly 6MWT. The accelerometer was equipped with a button that needed to be pressed at the beginning of the recording period and automatically stopped recording after six minutes. The device also captured walking intensity and step counts during the 6MWT. The authors found that the at-home 6MWT was more correlated with questionnaire-based quality of life than the NYHA classification.

Aranki *et al.* [47] described a repeated-measure real-time activity monitoring system for HF patients. This was a smartphone based system that used the in-device accelerometer to calculate daily energy expenditure and forwarded this information to clinicians.

Aranki *et al.* ran a three-month pilot study to test their system on 15 HF patients and found that smartphone position and compliance were challenges to optimal system performance [47].

Suh *et al.* [50] developed two versions of a tele-monitoring system that recorded PA as well as weight and blood pressure in HF patients. Their initial version of the system included a pocket-worn accelerometer to collect daily energy expenditure readings. The second version of the system collected activity readings through a smartphone-embedded accelerometer. Suh *et al.* tested their tele-monitoring system on 26 HF patients and found that it reduced abnormal (according to the guidelines) weight and blood pressure readings as a result of feedback from clinical staff to patients. Within this study, activity data were used to ensure PA compliance rather than as a risk marker for HF deterioration [50].

1.2.3.4 Lessons from Other Disease Areas

Activity monitors have been used in various areas of medicine, including chronic disease management, recovery and mental health [51] [52]. In a large proportion of studies, activity monitors are worn for relatively short periods of time, e.g. one or two weeks at the beginning of a study [53]. Studies that record data for a longer time period seem to be based on intuitive and convenient recording methods. Alahmari *et al.* [54] conducted a study on 73 chronic obstructive pulmonary disease (COPD, a chronic lung condition) patients. Patients were asked to keep a manual record of their symptoms and step count, on a daily basis. Step counts were collected using a pedometer. After a 12-month period, data were analysed to study activity characteristics during stable and exacerbation (acute worsening) periods [54]. The study found that step counts of COPD patients fell by more than 10% during exacerbation periods. Patients with comparably large reductions in step counts took longer to recover from their exacerbation.

Other systems used smartphone devices to repeatedly monitor PA behaviours. Bardram

et al. [52] developed a smartphone-based system to record activity in patients with bipolar disorder. Patients manually entered information on their daily activity level in addition to objectively-recorded activity by the smartphone. The evaluation of the system on 17 patients showed that the physical and social activity of patients correlated with clinically-rated depressive symptoms [55].

1.2.4 Sleep as a Heart Failure Marker

In addition to PA, the sleeping patterns of HF patients may also be of potential interest. HF has been shown to affect both the quality and duration of sleep [56][57]. More than 50% of HF patients suffer from sleep apnoea [56], a condition in which breathing intermittently stops during sleep. Sleep apnoea has been shown to negatively impact the quality of life of patients and potentially affects their likelihood of survival [58]. In addition to this, reversal of sleep apnea has shown to have a positive impact on HF markers, such as blood pressure and systolic contraction, however studies on the its effect on mortality are lacking [59].

Despite the potential benefits of sleep monitoring, research on sleep monitoring in HF patients has been limited [60][61]. Studies include an evaluation of an under-mattress sensor to detect changes in breathing rates of patients [60] and the analysis of a posture sensor to detect up-right sleeping position as a derivative marker of worsening in HF symptoms [61].

1.2.5 Motivation for Thesis

In summary, HF is a common and complex condition that is associated with high rates of hospitalisations and deaths [18]. A large number of HF-related adverse events are related to HF decompensation, a situation in which HF symptoms, such as breathlessness and swelling, gradually increase [19]. Using appropriate techniques this gradual deterioration may be detected early allowing for timely intervention and the prevention of HF-related

hospitalisations and deaths [5].

Home monitoring technologies can be used to monitor physiological and symptom markers over time and may provide a suitable tool to identify HF deterioration early [24]. In addition to this, such technologies may improve patient management post hospital discharge and therefore improve HF outcomes [4]. Factors through which this may happen include the optimisation of diuretic therapy, identification of clinical abnormalities (e.g. heart rhythm abnormalities) or abnormal readings and the implementation of evidence-based therapy.

Despite the potential benefits of HF home monitoring systems, they have been shown to have varying effectiveness in reducing HF-related adverse events [5] [30]. A review of HF tele-monitoring studies by Desai *et al.* [5] found that such systems provide little incremental value to HF patients' quality of life and clinical outcomes, especially once patients are well-managed and on evidence-based clinical therapy. The authors conclude that effective HF home monitoring systems will need to detect HF deterioration early and accurately with sufficient time to allow for intervention. According to Desai *et al.*, current weight-change thresholds and symptom markers do not fulfil these criteria.

As a result, this thesis set out to look beyond traditional thresholds and markers to predict HF deterioration. It set out to i) improve the predictive value of traditional HF markers using a data-driven approach (Chapter 4) and then to go a step further by ii) analysing the potential value of new PA and sleep markers for future home monitoring systems (Chapter 6). As part of the latter analysis, a methodological gap in the processing of accelerometer signals became apparent: current sleep detection methods require self-reported sleep-onset and wake-up times. Chapter 5 of this thesis therefore addresses this gap and describes the development and evaluation of an automated sleep-boundary detection method.

Datasets

2.1 Introduction

This chapter provides an overview of the datasets employed throughout this thesis. A total of three datasets were used, including both HF-specific and more generic accelerometer datasets. Section 2.2 describes the study layout and data collection of the SUPPORT-HF2 (SHF2) datasets. This dataset was collected over a two-year period and contains daily weight, blood pressure and heart rate measurements as well as symptom scores for more than 200 HF patients.

Section 2.3 introduces the dataset used for the development and evaluation of the data-driven sleep-boundary detection methods described in Chapter 5 of this thesis. The accelerometer dataset contains accelerometer signals for 50 individuals as well as their self-reported sleep-onset and wake-up times.

Finally, Section 2.4 provides information on the UK Biobank resource, a data resource which contains both accelerometer recordings and HF diagnostic information. This resource was used to identify potential PA and sleep markers for future HF home monitoring systems (Chapter 6).

The study set-up for the SHF2 study was part of the work undertaken for this thesis. All other datasets were made available through research collaborations, with all data analyses described in the following chapters undertaken as part of the research for this thesis.

2.2 Heart Failure Tele-monitoring Data (SUPPORT-HF)

The initial research on traditional markers for repeated-measure risk-prediction described in this thesis was based on the analysis of data from the SUPPORT-HF (Seamless User-centred Proactive Provision Of Risk-stratified Treatment for HF, SHF) study. The purpose of this study was to develop a user-centered home monitoring system for patients with chronic HF and to evaluate its ability to reduce HF-related hospitalisations and deaths. The study comprised two stages to date; the SHF1 study, during which the home-monitoring system was tested and iteratively improved [62][63] and the SHF2 study (Scotland South East Research Ethics Committee (REC) approval date 28/08/2014, reference 14/SS/1025), a randomised controlled trial evaluating the effect of the developed home monitoring system on evidence-based patient management (including optimal medication therapy) and as a secondary outcome HF-related hospitalisations and deaths.

In the following paragraphs, the SHF system and the SHF2 trial will be introduced. Subsequently, the population studied as part of the SHF2 study will be described.

2.2.1 The SUPPORT-HF System

The SHF system is a mobile health platform for self- and remote-management of HF patients. It was developed using an iterative approach taking into account feedback from both patients and medical staff [63]. The core components of the SHF system

are a computer tablet (running on the Android operating system), a Bluetooth-enabled weighing scale and a Bluetooth-enabled blood pressure monitor (see Figure 2.1), which automatically uploads weight and blood pressure measurements. The computer tablet runs the SHF application, which allows patients to

1. Answer a questionnaire to obtain their NYHA classification score
2. View their self-monitoring measurements and assessment history
3. Present information on HF, its symptoms and commonly used medications
4. Establish contact with the SHF clinical team.



Figure 2.1: *Components of the SUPPORT-HF system.*

2.2.1.1 Devices

Weighing scales

Three types of weighing scales were used during the SHF studies, all of them available from A&D (Tokyo, Japan). The scales were chosen based on their ability to transfer weight measurements via Bluetooth. During the initial study phase (SHF1), A&D UC-321PBT scales were used. However, the manufacturer discontinued this model and a subset of SHF2 patients were therefore provided with A&D UC-355PBT-Ci (maximum weight of 250 kg) and UC-351PBT-Ci (maximum weight of 150 kg) scales. Despite differences across patients, each individual patient was using the same weighing scale throughout the study period.

Blood pressure monitors

The blood pressure monitors used for the SHF study were the Food and Drug Administrative (FDA) and CE approved A&D UA-767PBT and UA-767PBT-Ci monitors, which used oscillometric technology [64] to derive systolic and diastolic blood pressure values as well as heart rate information. The blood pressure monitors were chosen based on their ability to transmit measurements via Bluetooth to the SHF application.

2.2.1.2 Daily Monitoring

When entering the SHF application, patients were directed to the 'My health diary' section (see Figure 2.2a). Here they were guided through a three-step assessment process: at first they were asked to answer the NYHA questionnaire to assess their ability to breathe normally during non-exercise and exercise activities. Next, patients were asked to take their blood pressure measurements by putting the blood pressure cuff on and waiting for the measurement to appear on the tablet screen (see Figure 2.2b). Finally, patients were asked to weigh themselves (see Figure 2.2c).



Figure 2.2: *Functionality of the SUPPORT-HF system.*

2.2.1.3 Data Transfer

Participants' data were transferred regularly from the tablet computer via the Internet to a secure NHS server acting also as the SHF backend. Data were transferred over an encrypted channel to guarantee data security. Data on the SHF backend were stored encrypted for the same reason. Clinical staff could review patient information by logging into the system using username and password. The SHF backend offered the option to review patient measurements and store additional patient information, such as demographics, blood test results, prescribed medications or general notes. The components and functionalities of the SHF system are summarised in Figure 2.3

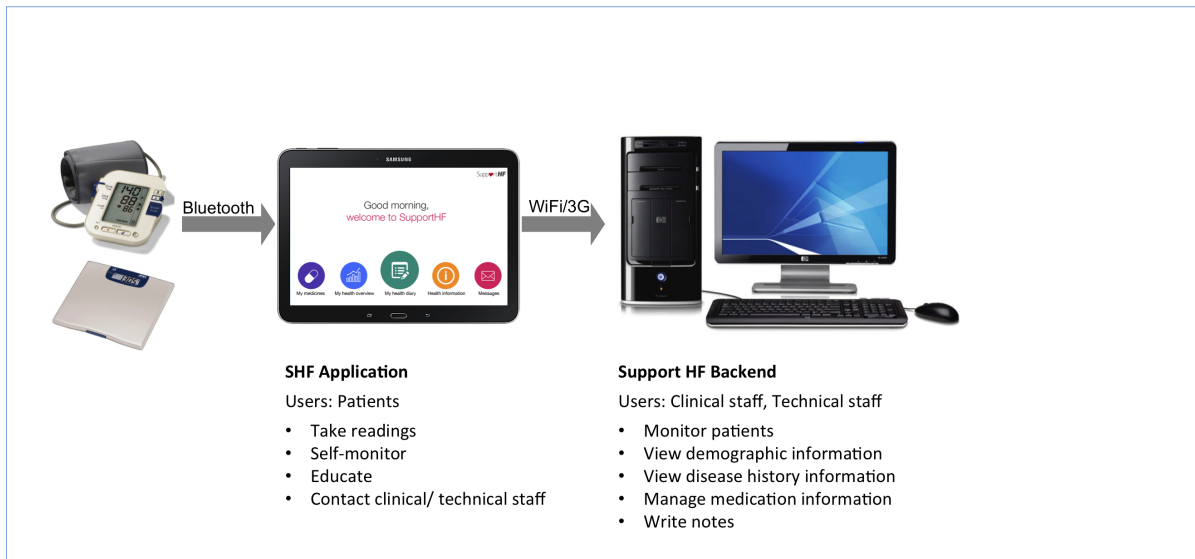


Figure 2.3: *Components of the SUPPORT-HF system. Data were recorded on the computer tablet and transferred from there to the system server on a daily basis.*

2.2.2 SUPPORT-HF2 Study

2.2.2.1 Study Population

The SHF2 study was a randomised controlled trial with the aim of evaluating the ability of a home-monitoring system to optimise medical therapy and reduce adverse events in HF patients. Ethical approval for the trial was obtained from the South-East Scotland multi-centre REC (approval date 28/08/2014, reference 14/SS/1025).

Participants were recruited from seven different sites across the UK and Northern Ireland (Belfast, Exeter, Frimley, Manchester, Milton Keynes, Leicester and Oxford). Participants were eligible to take part in the study if they had a confirmed diagnosis of HF and were likely to benefit from home monitoring. The latter was defined as HF patients with NYHA functional class scores of II to IV, BNP > 100 pg/L (see 1.1.1) or NT-pro-BNP (released during the production of BNP) > 360 pg/L in the last 30 days or patients not on optimal treatment. In addition to this, patients had to either have a >10% probability of one-year mortality (calculated using the MAGGIC score) or had to have experienced at least one HF-related hospital admission within the previous 12

months.

Participants interested in taking part in the SHF study were visited by a HF nurse and introduced to the SHF system. They were asked to provide demographic, medication and disease history information. A blood sample was taken in order to measure their BNP and NT-pro-BNP values. Finally, patients were asked to provide informed consent, if they were interested in taking part in the study. They were then provided with the SHF system.

After evaluation of the data collected during the first visit, participants eligible to take part in the SHF2 study were randomly assigned to the intervention or control group. This was done whilst adjusting for type of HF (reduced versus preserved ejection fraction), baseline risk of death (based on MAGGIC score) and study site. All patients were subsequently asked to use the system for a period of six months. Patients in the control arm were able to use the SHF system to record and review their measurements, contact the clinical team and read the HF educational advice; however only participants in the intervention group were provided with feedback and clinical advice on medication use. After the completion of the six-month monitoring period, patients could stop monitoring, in which case they were monitored for health outcomes without any intervention '*passive monitoring*', or continue monitoring until trial end, in which case the control/intervention monitoring was extended (see Figure 2.4).

Feedback was based on a set of technical and clinical scores that enabled clinical staff to monitor patients and identify those in need of clinical attention. The technical score provided information on system-usage and any participants with missing measurements for three or more days were flagged. This included participants from both the intervention and control arm of the study.

The clinical scores included a patient well-being (medical) score and a treatment

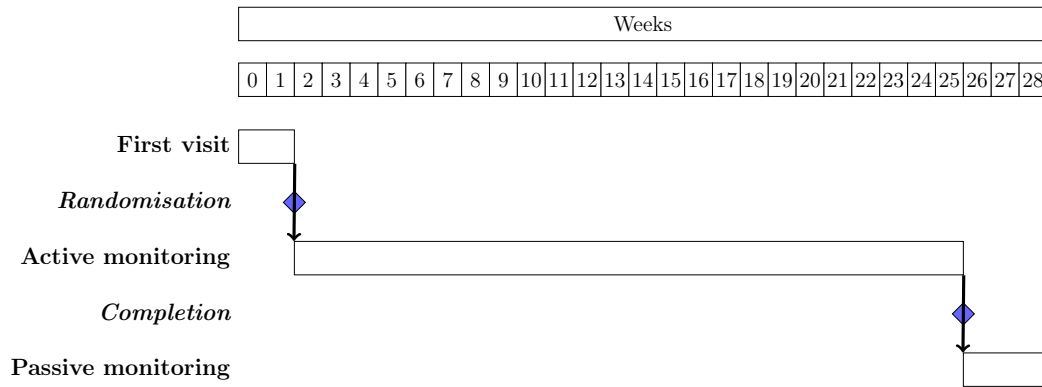


Figure 2.4: *The SUPPORT-HF timeline from recruitment to passive follow-up.*

score. The medical score was based on clinical guidelines and aimed to make clinical staff aware of any potentially deteriorating patients in the intervention group. The following thresholds were used:

- Weight change of ≥ 2 kg in three days
- Systolic blood pressure of > 150 mmHg or < 100 mmHg
- Diastolic blood pressure of > 100 mmHg or < 60 mmHg
- NYHA ≥ 1 higher than median across previous six days

The treatment score was implemented to visualise how closely the treatment regimen of a patient followed the guideline-recommended HF treatment in order to encourage evidence-based adjustments. The SHF clinical staff received daily emails with a list of participants who were flagged according to the technical and clinical scores. They then used their clinical judgement and knowledge of the patients to decide which patients to contact.

2.2.2.2 Data Overview

A total of 202 patients were randomised during the SHF2 study, including 57 (28%) female and 145 (72%) male participants. Patients had a median (interquartile range)

age of 73 (64-81) years and a median MAGGIC score of 23 (18-28)%. 60% of participants had HF with reduced ejection fraction and 92% of patients rated their technical ability of using a computer tablet as poor or intermediate. Table 6.1 provides an overview of the baseline characteristics of the SHF2 patients stratified by NYHA score.

NYHA	I (n = 64)	II (n = 54)	III (n = 67)	IV (n = 17)	All (n = 202)
Age, years (median (IQR))	73 (65-80)	73 (65-78)	76 (64-81)	68 (64-78)	73 (64-81)
Sex, female	23%	17%	39%	41%	28%
Reduced EF	64%	54%	64%	53%	60%
MAGGIC score	20 (14-25)%	20 (17-24)%	26 (22-31)%	28 (25-31)%	23 (18-28)%
Comorbidities					
Hypertension	48%	33%	58%	35%	47%
DM	31%	28%	33%	41%	32%
COPD	12%	17%	19%	29%	17%
CKD	9%	6%	22%	35%	15%
Technical skills					
Poor	36%	44%	46%	59%	44%
Intermediate	53%	48%	46%	35%	48%
Expert	11%	7%	7%	6%	8%
Monitoring duration, days (median (IQR)) (months)	198 (154-234)	203 (165-230)	201 (180-227)	197 (127-216)	200 (158-232)
Missing values (median (IQR))	11 (4-19)%	10 (5-24)%	9 (3-21)%	7 (1-38)%	10 (3-22)
Adverse events (n)	7	10	15	6	38

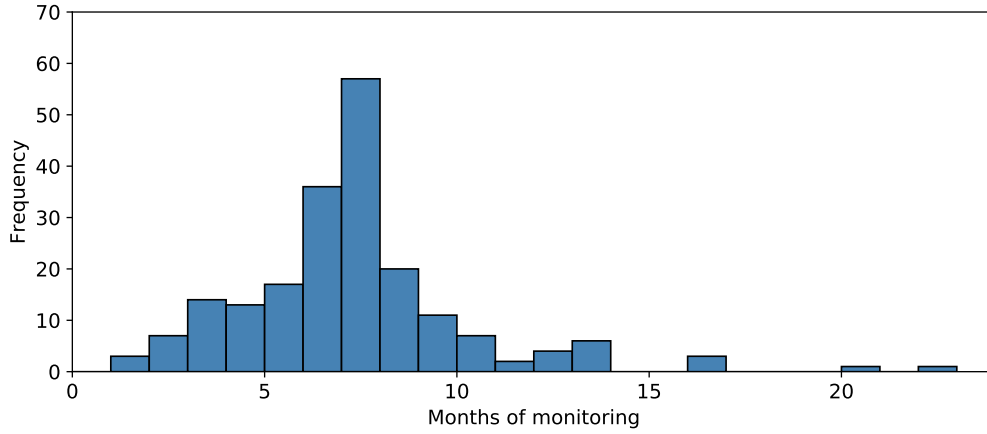
Table 2.1: *Baseline characteristics and compliance and adverse event information for participants randomised into the SUPPORT-HF2 study. Data are stratified by New York Heart Association (NYHA) class. IQR = interquartile range. EF = Ejection fraction. DM = Diabetes mellitus. COPD = Chronic obstructive pulmonary disease. CKD = Chronic kidney disease.*

Patients recorded daily measurements for a median period of 200 (158-232) days or 7 (5-8) months. Figure 2.5a shows the distribution of monitoring durations in months. Monitoring duration was defined as the first to last day with any weight, blood pressure

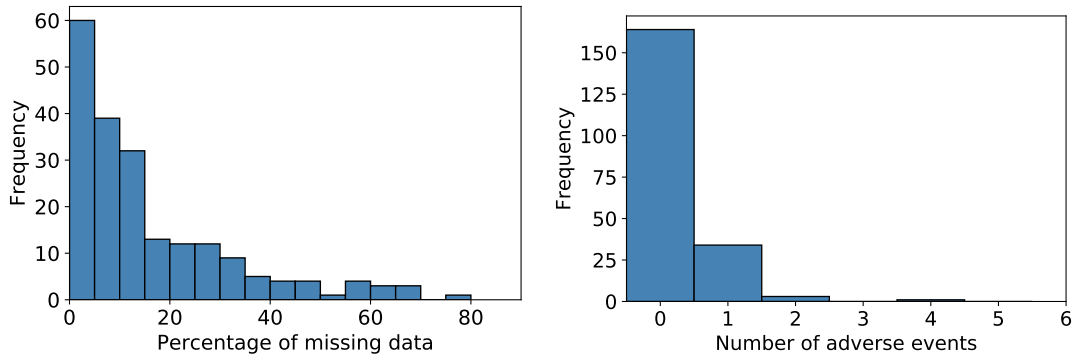
or NYHA data. 53 participants had less than six months worth of monitoring data. These included a small number of participants who discontinued the study ($n=14$), as well as patients who stopped recording their data because of death or unknown reasons ($n = 6$). In addition, 33 participants had not reached their six-month data collection point by the end of the study.

The median percentage of missing data was 10 (3-22)%. For this purpose, missing days were defined as days with no measurements. Any day with a single measurement (e.g. only weight) was not classed as having missing data. Figure 2.5b shows the distribution of the percentage of missing data days. It can be seen that a majority of patients had complete data for more than 85% of days, however a few participants have up to 80% missing values. The percentage of missing data days appears to be similar across NYHA-based disease severity (see Table 6.1).

Figure 2.5c shows the distribution of HF-related adverse events recorded throughout the SHF2 study. 44 adverse events were recorded from 38 participants. More than 160 study participants did not experience HF-related adverse events during the SHF2 study.



(a) *Histogram of monitoring durations across patients.*



(b) *Histogram of % missing data* (c) *Histogram of number of adverse events across patients.*

Figure 2.5: *Summary statistics for the SUPPORT-HF2 study.*

2.3 Labelled Actigraphy Dataset (Cognitive Health in Ageing Exercise Study)

Accelerometer recordings and self-reported sleep and wake data were collected as part of the Cognitive Health in Ageing Exercise (CHA) study. Ethical Approval for this study was obtained from the local REC (Oxford REC B Ref 10/H0605/48).

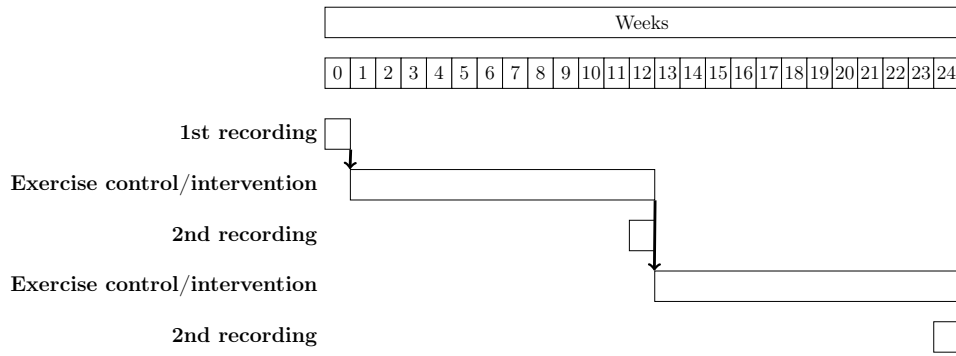


Figure 2.6: *The Cognitive Health in Ageing Exercise study timeline.*

2.3.1 Study Population

The purpose of the CHA study was to analyse the effect of an anaerobic exercise intervention on brain MRI (Magnetic Resonance Image) measures. Adults between the ages of 60 and 80 years that self-reported less than 60 minutes of heart-rate increasing PA per week and no contraindications to MRI scanning or fitness testing were eligible to take part in the research study. After providing informed consent, eligible participants were randomly assigned to a PA intervention or ‘wait-list’ control group. Participants in the intervention group were asked to take part in three monitored training sessions a week; participants in the control group were told to continue with their normal activity programme. After a period of 12 weeks the intervention and control groups were interchanged (see Figure 2.6).

2.3.2 Devices

Participants wore the GENEActiv (Activinsights, Cambridge, UK) accelerometer on their non-dominant wrist for nine consecutive days in weeks 0, 12 and 24 of the study and filled in daily sleep diaries with lights-off, asleep (sleep-onset) and wake-up information during these periods. Information on sleep interruptions or day-time naps was also recorded (see Figure 2.8). The GENEActiv sensor (see below) contains a tri-axial accelerometer as well as a temperature sensor and has previously been validated for

sleep research [65][66]: Van Hees *et al.* compared GENEActiv sleep scores to the gold-standard for sleep detection, polysomnography [67], in 28 healthy adults and found that accelerometer scores achieved sensitivities (correctly identified sleep epochs of five-minute length) and specificities (correctly identified wake epochs of five-minute length) of 91% and 45% respectively. Accelerometer data were collected at a sampling frequency of 87.5Hz.

The self-reported sleep-onset and wake-up times were recorded by means of a so-called sleep-diary (see Figure 2.8). Patients were asked to fill in the lights-off, asleep (sleep-onset) and wake-up times. In addition to this, information on day-time naps and sleep interruptions was recorded.



(a) *GENEActiv* wrist-worn monitor used as part of the Cognitive Health in Ageing (CHA) Exercise study.



(b) *Arivity* wrist-worn monitor used as part of the UK Biobank resource.

Figure 2.7: Accelerometer monitors relevant to this thesis.

What time did you: *Please use the 24-hour clock*

- Turn off the light last night (ready to go to sleep)? __ : __
- First fall asleep last night? __ : __
- Wake up today (eyes open, ready to get up)? __ : __

Did you wake up during the night? ☐ Yes ☐ No

If yes, for how long were you awake in total? __ hours __ minutes

Did you have a nap today? ☐ Yes ☐ No

If yes, please detail any naps:

	1 st Time	2 nd Time	3 rd Time	4 th Time
Time you fell asleep	__ : __	__ : __	__ : __	__ : __
Time you woke up	__ : __	__ : __	__ : __	__ : __

Figure 2.8: *Sleep diary used to collect sleep-onset and wake-up information.*

2.3.3 Data Overview

51 participants took part in the CHA Exercise study, 46 of whom had at least one day's worth of accelerometer recordings with corresponding self-reported sleep information. A total of 972 days' worth of accelerometer recordings were included in the analysis. The mean age of all included participants was 66.7 (standard deviation: 5.4) years. 29 (63%) participants were female, 17 (37%) male. Table 2.2 provides an overview of participant characteristics and self-reported data grouped by tertiles of total sleep duration (tertile boundaries: 0h-7.3h, 7.4h-8.0h, 8.1h-9.2h).

	Tertile 1 (0-7.3h)	Tertile 2 (7.4-7.9h)	Tertile 3 (8.0-9.2h)	All
n	15	14	17	46
Age, years (median (IQR))	66 (62-70)	66 (64-70)	64 (62-67)	65 (62-74)
sex, female	60%	64%	59%	61%
Self-reported days (median (IQR))	25 (18-27)	21 (11-27)	18 (16-26)	18 (18-27)
Total sleep time (h) (median (IQR))	6.7 (6.5-7.0)	7.7 (7.5-7.8)	8.3 (8.1-8.6)	7.6 (6.1-8.1)

Table 2.2: *Baseline characteristics and self-reported sleep information of participants taking part in the Cognitive Health in Ageing (CHA) Exercise study. Characteristics are provided by tertiles of self-reported total sleep time. IQR = interquartile range.*

2.4 Heart Failure Accelerometer Dataset (UK Biobank)

The UK Biobank resource [68] was used to study objectively-recorded PA and sleep behaviours of HF patients and to identify potential markers for repeated-measure HF risk-stratification. The UK Biobank is a registered charity with the aim of improving prevention, diagnosis and treatment of life-threatening diseases [68]. Between 2006 and 2010, the charity recruited 500,000 subjects aged 40-69 years from across the United Kingdom and collected their self-reported information, physiological measurements and biological samples. These data are available to researchers across the world. The UK Biobank has ethical approval from the North West multi-centre REC, which covers the UK.

2.4.1 Study Population

Participants whose data are stored in the UK Biobank resource were recruited via their NHS contact details and invited to take part in an assessment centre, where they received further information and, if willing to participate in the study, were asked to sign an informed-consent form. Participants then took part in a baseline assessment, which

involved the completion of a questionnaire on topics such as diet, PA, smoking and medical history, and the collection of physiological measurements (e.g. blood pressure, height, weight, etc.) and biological samples. Repeat assessments collecting the baseline information again have been run on subgroups of the UK Biobank population (see Figure 2.9).

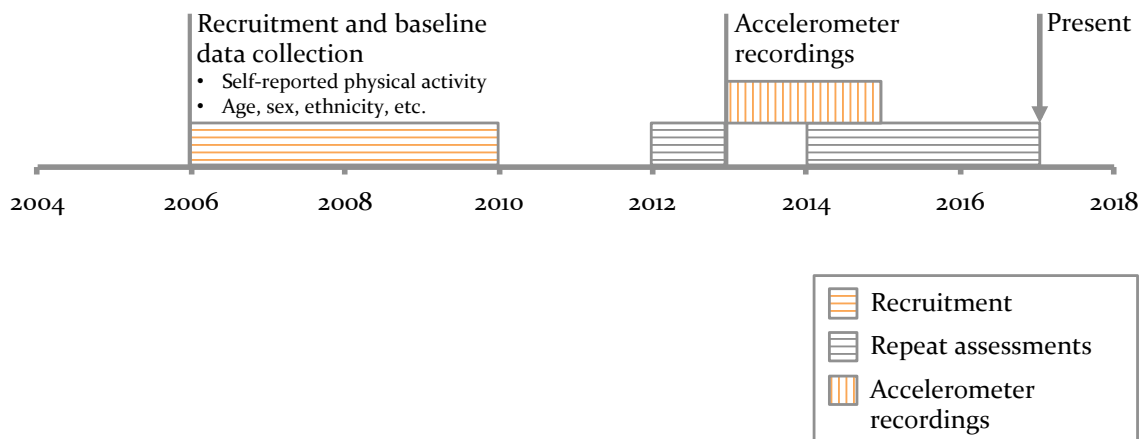


Figure 2.9: *Timeline of UK Biobank data collection showing time of baseline data collection, collection of repeat assessments (run on sub-groups of the original participant pool) and time of accelerometer data collection.*

Accelerometer data from 100,000 UK Biobank participants were collected between May 2013 and 2015 [69]. UK Biobank participants were contacted via email and provided with an Axivity AX3 (Open Lab, Newcastle, UK) accelerometer (see Figure 2.7b) if they agreed to take part in the activity data collection. Participants were told to wear the tri-axial accelerometer continuously on their dominant wrist for a period of seven days. This is different to the wearing location of the CHA Exercise study. Both the dominant and the non-dominant wrist have been used in previous accelerometer studies depending on the types of activities that are to be detected [70]. Generally, activities that involve the use of the dominant arm and wrist, such as holding objects, writing, etc. will be easier to detect using monitors worn on the dominant wrist. The accelerometer recorded data at a frequency of 100 Hz and a range of ± 8 g (gravity units).

	No HF	HF
Self-reported PA	388,469	1,605
Accelerometer-based PA	96,109	596

Table 2.3: *Sample sizes of the subpopulations of the UK Biobank collected at baseline (self-reported Physical Activity (PA)) and at the time of the accelerometer recordings grouped by Heart Failure (HF) status.*

Long-term health-related information is linked to the UK Biobank on an annual basis, from sources including the Office for National Statistics (Death and emigration information) and the Department of Health’s Hospital Episode Statistics (HES). The latter contains information on English residents that are either NHS-funded or NHS-treated, including

1. Admitted in-patient data (1997 onwards)
2. Outpatient data (2003 onwards)
3. Accident and emergency data (2008 onwards).

At the point of data extraction, health-related outcome data were based on hospital-admission data up until March 2015. Figure 2.9 shows a timeline summarising the different stages of the UK Biobank data collection and the data used as part of the analysis described in this thesis.

2.4.2 Data Overview

The UK Biobank resource collected data from a total of 502,602 participants. 390,074 participants provided complete self-reported PA information at baseline, 1,605 of whom had previously been diagnosed with HF (see Table 2.3). Accelerometer-based PA data with good calibration and wear-performance were collected for 96,705 UK Biobank participants, 596 of whom had been diagnosed with HF at the time of recording. Detailed characteristics of the participants in both the self-reported and accelerometer-based PA groups are provided in the results section of Chapter 6.

2.5 Summary

In summary, three datasets were used as part of this thesis: firstly, the SHF2 dataset, which contains approximately six months' worth of tele-monitoring data for 202 HF patients. This dataset was used to develop and test repeated-measure risk-prediction methods based on traditional HF markers.

The second dataset, the CHA Exercise dataset, contains accelerometer recordings and self-reported sleep-onset and wake-up information for 46 participants. These data were used to develop data-driven sleep-boundary detection methods to fill the methodological gap in the automated processing of accelerometer signals.

Finally, the self-reported and accelerometer-derived PA information from UK Biobank participants with and without HF was used to identify potential new markers for HF deterioration, which may prove to be suitable additions to future home monitoring systems.

Methods

3.1 Introduction

The previous chapter introduced the datasets used in this thesis. This chapter will provide an overview of the methods that were employed both for the estimation and evaluation of HF risk-markers, as well as for the development of the sleep-boundary detection methods.

Section 3.2 starts with an introduction on probability density functions (PDFs), which can be used to explore and visualise the distribution of a variable. In Chapter 5, PDFs were used to provide a simple method for sleep-boundary detection. In the remaining analysis chapters, PDFs were used to visualise the performance of HF risk-stratification methods.

Sections 3.3 and 3.4 cover two types of classification methods used in this thesis, i) logistic regression and ii) random forest classifiers. Logistic regression classifiers were used to evaluate the performance of traditional HF markers and PA and sleep-based markers. The random forest classifier was another method used to identify sleep-boundary times in Chapter 5.

In addition to the PDF and random forest methods, Chapter 5 also explored the use of a change-point detection approach. This approach identifies changes in the time-series signal that may occur in the transition from wakefulness to sleep and vice versa. The change-point detection method is explained in section 3.5 of this chapter.

Finally, section 3.6 provides an overview of the concepts used to evaluate the performance of various algorithms throughout this thesis. These include receiver operating curves (ROC), the area under the receiver operating curve (AUC), the kappa statistic, accuracy, sensitivity and specificity.

3.2 Probability Density Functions

PDFs represent the likelihood of a variable taking on any specific value in the variable-space. They can be extracted by applying a smoothing kernel over individual values of the variable and aggregating the kernels into one function. The function is then scaled to have an area under the curve of 1, therefore presenting the probability of the variable to take on any of the possible values.

Throughout this thesis, a Gaussian kernel with a bandwidth (bw) (Equation 3.1) set by Scott's rule of thumb ([71]) was used as the smoothing kernel.

$$bw = N^{\frac{1}{d+4}}, \quad (3.1)$$

where N is the length of the data array and d is the dimension of the data.

PDFs provide a tool to visualise how well a method differentiates between two classes. A mathematical way to quantify the difference between two PDFs is the Kullback-Leibner divergence (D_{KL}).

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}, \quad (3.2)$$

where $P(i)$ is a point on the distribution P and $Q(i)$ is a point on the distribution Q . Small values of D_{KL} suggest a large similarity between distributions; large values correspond to large differences.

3.3 Logistic Classification

Logistic classification [72] is a statistical method that enables the estimation of the relationship between one or more independent variables ($x_1, x_2, \dots, x_m$, where m corresponds to the number of input variables) and a dependent categorical outcome variable (Y) [73]. It can be used to study the association between the independent and dependent variables as well as to evaluate their predictive performance. As part of this thesis, it was used to explore the predictive value of traditional HF markers and analyse the association between PA and sleep markers and HF status.

With linear regression, the dependent variable is continuous, and the relationship between the independent and dependent variables can be described as shown in Equation 3.3,

$$Y = b_0 + b_1x_1 + \epsilon, \quad (3.3)$$

where b_0 is the offset, b_1 is the weight given to the independent variable x_1 and ϵ is the error or noise, assumed to be normally distributed.

Logistic regression deals with categorical dependent variables, for example binary variables ($Y = [0, 1]$). A linear fit is no longer suitable and the relationship between the independent and dependent variable can now be described in form of an alternative function, the sigmoid function (Equation 3.4).

$$P(Y = 1) = \frac{e^{b_0 + b_1x_1 + \epsilon}}{1 + e^{b_0 + b_1x_1 + \epsilon}}, \quad (3.4)$$

The intercept (b_0) and the weights ($b_0, b_1, \dots, b_m$) of the logistic regression function are optimised to reduce the sum (S) of the errors between the estimates (Y_{est}) and the true values (Y), using the *least squares* method (see Equation 3.5).

$$S = \frac{1}{N} \sum_{n=0}^N \sqrt{(Y_{est} - Y)^2} \quad (3.5)$$

The process from adjusting the b values ($b_0, b_1, \dots, b_m$) to evaluating the mean absolute error is repeated until i) the improvement in fit is below a specified threshold or ii) a specified maximum number of iterations has been reached.

The logistic regression method assigns each input vector a label between '0' and '1'. If the dataset is balanced, a threshold of 0.5, above which the outcome is classed as '1' is typically used to assign the final classes.

3.4 Random Forest Classifiers

Random forest classifiers are ensemble classifiers that combine multiple individually weak decision trees into one stronger classifier [74]. The advantage of random forest classifiers over other classifiers is that they create multiple randomly selected subsets of the training data and therefore reduce the impact of hierarchical structures within the dataset. Hierarchical structures are likely to be found in time-series data such as accelerometer data, where within-subject readings may be correlated over time.

At the centre of the random forest classifier lies the decision tree. Decision trees assign classes or labels by repeatedly partitioning datasets using optimised thresholds. Setting the branch-depth is important as decision trees will otherwise overfit on the training data, meaning that they learn the individual variations in the data points rather than the underlying structure of the data.

There are different methods to choose the best-performing variable and threshold

combination. The scikit [75] package in Python was used in the investigations described in this thesis and uses the so-called Gini impurity. Gini impurity is a measure of uncertainty and quantifies the probability of assigning each of the items in a branch to the wrong class. It is calculated as shown in Equation 3.6:

$$G(p) = \sum_{i=1}^I p_i(1 - p_i), \quad (3.6)$$

where I is the number of different class labels and p_i is the fraction of labels of class i . Computing the Gini impurity for each variable and threshold combination allows the decision tree to identify the combination that results in the smallest uncertainty and hence the combination which performs best at differentiating between the two groups.

Even when the depth of a decision tree is restricted, the fact that it uses all input features and all input data makes it prone to fitting the training set too closely. Random forest classifiers try to overcome this challenge by combining multiple decision trees with a randomly selected sub-set of input variables (default value is $\sqrt{n_{feat}}$, where n_{feat} is the total number of features) and input samples (default value is all samples). Each individual tree then 'votes' on the class of an item and the combined classification score (z_{all} , between '0' and '1') is calculated by averaging across the individual decision tree scores (z_i) as shown in Equation 3.7:

$$z_{all} = \frac{1}{N} \sum_{i=1}^N z_i, \quad (3.7)$$

where N is the number of decision trees in the random forest classifier.

The performance of the random forest classifier is dependent on the number of trees included in the classifier as well as the depth of each of the trees. These parameters need to be tuned to the requirements of the data to be fitted. Increasing the number of trees to a certain extent allows for a greater variety of randomly selected input variable combinations to be tested. This in turn reduces the likelihood of individual trees worsening the performance of the combined classifier. As in the case of individual decision trees,

the depth of the trees is another parameter by which the performance of the classifier can be optimised. This can be controlled by i) limiting the number of splits in each tree and therefore prohibiting large growth and ii) by setting a requirement of the minimum number of cases allowed for a branch to split.

The importance of individual variables in the random forest classifier can be estimated by multiplying the reduction in Gini impurity achieved by a node by the probability of reaching that node. This is then averaged across all trees in the classifier.

3.5 Change-point Detection

In Chapter 5 of this thesis, one of the methods employed for sleep-boundary detection was the change-point detection (CPD) method developed by *Killick et al* [76], implemented in the R package '*changepoint*' [77].

This method is based on maximum log-likelihood estimation (MLE), which provides a measure of the integrity/agreement within a time-varying signal. MLE quantifies the likelihood of a given set of values ($x_{1:N}$ of length N) occurring given the best-fitting (or most likely) underlying distribution for these values, when this distribution can be described using parameters θ (e.g. mean, standard deviation of a normal distribution):

$$MLE(N) = \log p(x_{1:N}|\theta) \quad (3.8)$$

If the sum of the MLEs of two segments of a signal prior to and after a time τ ($MLE(\tau)$, Equation 3.9, Figure 3.1) is larger than the MLE of the combined signal ($MLE(N)$), τ is marked as a change point within the signal.

$$\begin{aligned} MLE(\tau) &= MLE_1 + MLE_2 \\ &= \log p(x_{1:\tau}|\theta_1) + \log p(x_{(\tau+1):N}|\theta_2) \end{aligned} \quad (3.9)$$

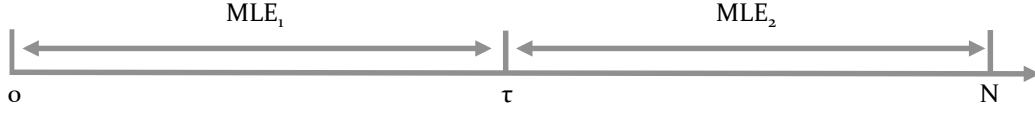


Figure 3.1: *Visualisation of the maximum likelihood estimation.*

In order to identify multiple change points within the signal, a binary-segmentation approach was employed, whereby the dominant change point is identified first and samples either side of the change point are consecutively scanned for additional change points. The method stops searching for change points after a pre-set number of points have been identified.

3.6 Evaluation Tools

3.6.1 Statistical Tests

Throughout this thesis, statistical tests were used to study whether the values of two groups (e.g. participants with and without HF) were significantly different. The t-test is a statistical test that compares two means scaled by a factor dependent on the variances of the two distributions (see Equation 3.10) and was used to compare variables that were continuous and assumed to be normally distributed.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad (3.10)$$

where $\bar{x}_1$ is the mean of the first sample, $\bar{x}_2$ is the mean of the second sample, n_1 and n_2 are the number of observations in each sample and s_1 and s_2 are the standard deviations of each sample. The t-value is then compared to the t-distribution corresponding to the sizes and variances of the samples to obtain the p-value. A p-value of less than 0.01 was considered as significant unless stated otherwise.

Non-normally distributed continuous variables were tested using the Wilcoxon rank sum test. The test ranks all participants by the variable of interest and determines whether the ranks assigned to one group are significantly different to the ranks assigned to the other group. Depending on the number of participants in each group and the difference in rankings a p -value is assigned.

Variables of categorical form were tested using the χ^2 test (Equation 3.11).

$$\chi^2 = \sum_n^N \frac{O_n - E_n}{E_n}, \quad (3.11)$$

where O_n is the observed number of times a category n occurred in the one group and E_n the number of times it occurred in the other group. N is the number of different values (categories) a variable can take. Depending on N the resulting χ^2 value is compared to the corresponding χ^2 -distribution. The likelihood of χ^2 taking on its value or greater on the corresponding distribution is then expressed in form of the p -value.

3.6.2 Accuracy, Sensitivity and Specificity

Accuracy, sensitivity and specificity are summary statistics that quantify the correctness of an estimate or a prediction. For the purpose of this thesis, they were used in the context of how well a classification method can differentiate between two groups (e.g. Will a patient experience a HF deterioration? Yes/No). All three statistics are based on the number of correctly identified cases as shown in Table 3.1.

	True value is '1'	True value is '0'
Method classes data as '1'	True positive (t_p)	False positive (f_p)
Method classes data as '0'	False negative (f_n)	True negative (t_n)

Table 3.1: *Definitions of true positives, false positives, true negatives and false negatives.*

Accuracy refers to the overall proportion of correctly identified instances, sensitivity to the proportion of correctly identified true cases (true value = 1) and specificity to the correctly identified false cases (true value = 0). The equations for all three statistics are shown in below:

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \quad (3.12)$$

$$Sensitivity = \frac{t_p}{t_p + f_n} \quad (3.13)$$

$$Specificity = \frac{t_n}{t_n + f_p} \quad (3.14)$$

3.6.3 Kappa Statistic

Whilst the performance of classification methods is often quantified using accuracy, this statistic may be affected by the frequency of occurrence of each state, e.g. if there are significantly more '0' labels and an algorithm classifies all instances as '0', the method may achieve a high accuracy despite having a poor sensitivity. Cohen's kappa coefficient incorporates the relative probability of chance agreement (P_e) and may therefore be a more appropriate statistic in some cases. It is calculated as shown in Equation 3.15, where P_0 is the probability that a label is '0' and the estimate says '0' and P_1 is the probability that a label is '1' and the estimate says '1'.

$$kappa = \frac{Accuracy - P_e}{1 - P_e}, \quad (3.15)$$

where

$$P_e = P_1 + P_0 \quad (3.16)$$

and

$$P_1 = \frac{t_p + f_p}{t_p + t_n + f_p + f_n} * \frac{t_p + f_n}{t_p + t_n + f_p + f_n} \quad (3.17)$$

$$P_0 = \frac{t_n + f_n}{t_p + t_n + f_p + f_n} * \frac{t_n + f_p}{t_p + t_n + f_p + f_n} \quad (3.18)$$

3.6.4 Receiver Operating Curves

Receiver operating curves (ROCs) provide an alternative method to evaluate the performance of a classifier, by plotting the sensitivity against (1-specificity) for a range of different thresholds. An example ROC is shown in Figure 3.2. The orange line corresponds to random prediction. The further a ROC moves away from this line, the better the method characterised by the ROC is at differentiating between classes. This can be quantified by the area under the receiver operating curve (AUC) as shown in Equation 3.19.

$$AUC = \int_{-\infty}^{\infty} Sensitivity(T)(1 - Specificity(T))dT, \quad (3.19)$$

where T corresponds to the threshold value. A classifier for which performance is no better than random chance yields an AUC of 0.5.

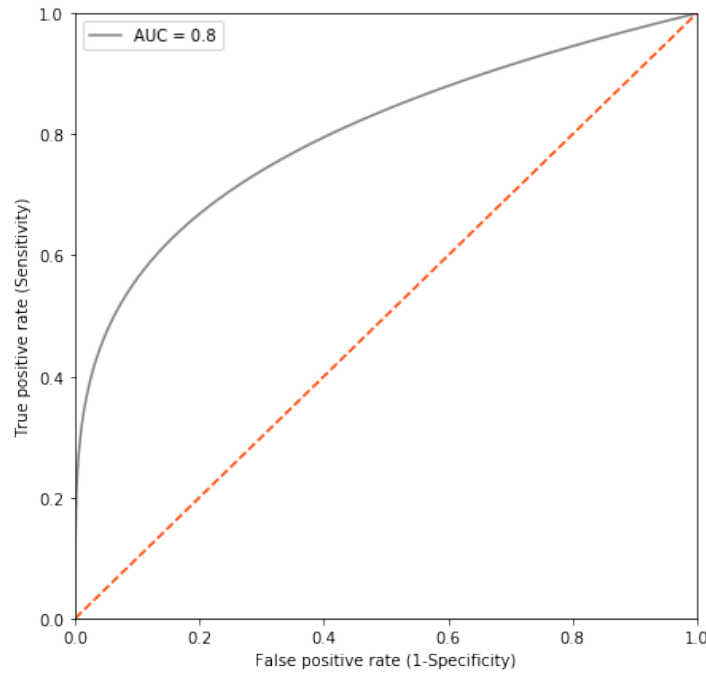


Figure 3.2: *Visualisation of the receiver operating curve.*

3.7 Summary

This chapter provided an overview of the key methods used throughout this thesis. The PDF and logistic regression methods introduced in this chapter were used for the evaluation of both traditional and accelerometer-derived markers for HF deterioration described in Chapter 4 and 6. In addition to this, the change-point detection and random forest methods were used for the development of automated sleep-onset and wake-up algorithms described in Chapter 5 of this thesis.

Repeated-Measurement Risk Prediction using Traditional Heart Failure Markers

4.1 Introduction

This chapter set out to evaluate the combined predictive value of physiological values and symptom scores collected as part of the SHF2 study. The principal contributions of this chapter are:

1. A review of the existing evidence for the predictive value of the physiological and symptom markers collected as part of the SHF2 study.
2. The development and evaluation of a classifier-based early warning system for HF patients using data-driven HF markers.

4.2 Background

4.2.1 The Predictive Value of Physiological Measurements and Symptom Scores

Weight, blood pressure, heart rate and NYHA scores were recorded on a daily basis as part of the SHF2 study. The following sections firstly summarise the evidence for the ability of each parameter to predict HF deterioration and then provide an overview of existing early warning systems developed for remotely-monitored HF patients. Relevant sources from the literature were identified based on the review on HF risk prediction by Rahimi *et al.* [20] and the reviews on HF tele-monitoring systems by Louis *et al.* [4] and Gorthi *et al.* [78]. Where available, information on specificity, sensitivity and accuracy of individual or combined markers is provided in the text. Alternatively, odds ratios or hazard ratios (described in further detail below) were recorded for single measurement risk-prediction markers. It should be noted that the collated performance measures come from a wide range of studies and were evaluated with respect to a variety of outcomes, including amongst others in-hospital mortality and one-year mortality. The performance measures can therefore not be directly compared. Instead, the forthcoming summaries aim to provide an overview of the significance of the different markers with respect to HF deterioration.

4.2.1.1 Weight Measurements

HF-caused fluid retention in the body leads to weight increase, which in turn provides a marker for patient deterioration. As a result, guidelines produced by the European Society of Cardiology recommend that patients monitor their own weight and increase their diuretic dose (or alert their healthcare team) if they experience unexpected weight gains of >2 kg in 3 days [9]. Whilst many HF monitoring studies utilised guideline-recommended thresholds, some have looked deeper into the predictive value of weight-based markers and reported on their performance (see Table 4.1 for more detail).

Measurements	Source type	Threshold	Study/ Author
Weight	Guideline	>2 kg in 3 days	Ponikoswki <i>et al.</i> [9]
Weight	Tele-monitoring	≥ 3 lbs in 5 days; ≥ 5 lbs in 3 days	Abraham <i>et al.</i> [32]
Weight	Tele-monitoring	>2 kg in 2-3 days	Scherr <i>et al.</i> [79]
Weight	Tele-monitoring	>5 lbs in 3 days; >2 lbs from baseline	Chaudhry <i>et al.</i> [80]
Weight	Tele-monitoring	≥ 2 kg in 2-3 days	Lewin <i>et al.</i> [81]
Weight	Tele-monitoring	≥ 1.5 kg from baseline	Roth <i>et al.</i> [82]
Weight	Tele-monitoring	Difference between mean weight over past 5 days and mean over past 80 days >4 lbs	Zhang <i>et al.</i> [26]
BMI	Risk score	HR, 1.18 per decrease in BMI group	Brophy <i>et al.</i> [83]
BMI	Risk score	HR, 0.8 per 2 kg/m ² increase up until 25 kg/m ²	O'Connor <i>et al.</i> [84]
BMI	Risk score	HR, 0.96 per 1 kg/m ² increase	Levy <i>et al.</i> [85]
BMI	Risk score	HR, 0.95 per 1 kg/m ² increase	Wedel <i>et al.</i> [86]
BMI	Risk score	HR, 0.96 per 1 kg/m ² difference from 25 kg/m ²	Manzano <i>et al.</i> [87]
BMI	Risk score	HR, 0.97 per 1 kg/m ² increase up to 30 kg/m ²	Pocock <i>et al.</i> [88]

Table 4.1: Evidence for the predictive value of weight-based markers for HF deterioration. BMI = Body mass index. HR = Hazard ratio. A HR of 0.96 corresponds to a 0.04 times reduction in risk for every 1 kg/m² increase in BMI.

Chaudhry *et al.* [80] compared the weight values of 134 HF patients with, and 134 without, imminent HF hospitalisations [80]. Their analysis showed that changes in baseline-adjusted weight were on average visible 30 days ahead of HF deterioration and increased severely in the seven days leading up to the event. Patients with a seven-day weight increase of 2 lbs (0.9 kg) or more had odds ratios of 2.77 and higher of being hospitalised in comparison with patients with weight increases of less than 2 lbs (0.9 kg), i.e. patients with the given weight increase were 2.77 times as likely to be hospitalised

as those without.

The odds ratio describes the association between weight increases and hospitalisations, and is related to the b-coefficient in logistic classification (see Section 3.3 for more detail). Hazard ratios as shown in Table 4.1 are a similar measure referring to how likely an event is to occur within a certain timeframe. However, neither odds ratio nor hazard ratio describe the predictive performance (typically quantified using sensitivity and specificity) of a marker. Lewin *et al.* [27] evaluated the performance of the ' $\geq 2kg$ weight increase over a period of 2-3 days' marker on its own and in combination with BNP changes in predicting clinical deterioration determined by physicians' review (see Section 1.1.1 of this thesis for more information on BNP). The study with 77 participants showed that the weight marker by itself had a sensitivity of 17% in predicting one day HF deterioration and a specificity of 94%. Combining the weight with the BNP marker increased the sensitivity to 55% and reduced the specificity to 74%. The authors concluded that weight measurements by themselves or in combination with BNP did not achieve sufficient sensitivity to be used as a reliable predictor of HF deterioration.

Another comparison study was conducted by Abraham *et al.* [32], who compared the predictive performance of different weight thresholds to that of intrathoracic impedance measurements. The study of 156 participants showed that impedance measurements outperformed weight markers in predicting 30-day HF hospitalisations, with the latter achieving a sensitivity of 22.5% and an unexplained detection rate of 4.3 cases per patient per year.

Finally, Zhang *et al.* [26] evaluated two rule-of-thumb markers (weight gain of 3 lbs (1.4 kg) in 1 day or 5 lbs (2.3 kg) in 3 days) and one moving average marker (short-term mean over long-term mean) to predict HF worsening within the 14 days prior to hospitalisation. The authors found that the rule-of-thumb algorithms had sensitivities (defined as percentage of correctly-identified 14-day episodes with hospitalisations)

of 58% and 65% and specificities (defined as percentage of correctly-identified 14-day episodes without hospitalisations) of 54% and 58% respectively. The moving average algorithm, optimised based on consideration of acceptable rates of true and false alerts, achieved a sensitivity of 20% and a specificity of 89%. Overall the authors concluded that additional markers might be required to obtain a more suitable prediction of HF worsening.

In addition to changes in weight, absolute weight [89] or BMI [88] may also be indicative of HF deterioration. In particular, lower BMI scores have been associated with an increased risk of mortality [85][88] (see Table 4.1). However, studies evaluating the ability of BMI to predict HF deterioration could not be found.

4.2.1.2 Blood Pressure Measurements

Reduction in the cardiac output of the heart leads to decreased organ perfusion in the early stages of HF. The body reacts by increasing cardiovascular resistance, resulting in a rise in blood pressure. Whilst the increase in blood pressure resolves the immediate need for organ perfusion, it also poses additional stress on the already failing heart leading to long-term remodelling of the heart and worsening of HF [90]. Previous monitoring studies have therefore included warnings for blood pressure values above pre-defined thresholds [82][91].

Cardiovascular medication, such as beta blockers and ACE inhibitors, causes a reduction in blood pressure and has been associated with long-term improvements in HF mortality and morbidity. Over-dosage of such medication can, on the other hand, lead to a clear reduction in systolic blood pressure [92], which can cause dizziness, fainting, falls and has been found to be an independent predictor of HF-related mortality and hospitalisations [93][94] (Table 2).

In an analysis of more than 60,000 HF patients participating in the ADHERE study,

Fonarow *et al.* found that systolic blood pressure values less than 115 mmHg were associated with an increased risk of in-hospital mortality [93]. Hsieh *et al.* developed a clinical prediction rule for short-term in-hospital and out-of-hospital outcomes after adverse events and included a range of systolic blood pressure thresholds at different points in their decision tree [94].

Other analyses have studied the association between low systolic blood pressure and longer-term outcomes. O'Connor *et al.* analysed the data from 4,014 patients hospitalised with HF and found that systolic blood pressure values at discharge of 130 mmHg or below were associated with post-discharge mortality [89]. Every 10 mmHg decrease in systolic blood pressure below 130 mmHg was associated with a 10% increase in mortality risk.

Felker *et al.* studied 949 patients from the OPTIME-CHF dataset and showed that low systolic blood pressure was an independent predictor of 60-day mortality [95]. The authors found that each 10 mmHg decrease in systolic blood pressure was associated with a hazard ratio of 0.78. Results from all the analyses mentioned above are summarised in Table 4.2.

As part of this review, only two repeated-measure monitoring studies with information on the pre-defined blood pressure thresholds used to alert clinical staff were found: In a 118-patient cohort with chronic HF, Roth *et al.* conducted a tele-monitoring study and implemented warning thresholds to allow for early clinical intervention. The authors used systolic blood pressure thresholds of <90 mmHg and >180 mmHg and diastolic blood pressure thresholds of >110 mmHg [82]. Cleland *et al.* used systolic blood pressure thresholds of <90 mmHg or >140 mmHg as triggers for close review of patient management [91]. Neither study reported on the sensitivity or specificity of the blood pressure markers.

Table 4.2 summarises the evidence of systolic and diastolic blood pressures as markers for HF deterioration. Overall, it can be seen that low systolic blood pressure measurements or hypertension have been associated with an increased likelihood of adverse events; however there is limited consistency in the blood pressure thresholds that have been used throughout the literature. The predictive performance of blood pressure measurements, in terms of sensitivity and specificity, has not been reported.

Measurement	Source type	Threshold	Study/ Author
SBP	Guideline	<90 mmHg; hypertension ($\geq 140/90$ mmHg)	Ponikowski <i>et al.</i> [9]
SBP	Tele-monitoring	<90 mmHg or >140 mmHg	Cleland <i>et al.</i> [91]
SBP	Tele-monitoring	<90 mmHg or >180 mmHg	Roth <i>et al.</i> [82]
SBP	Risk score	HR, 0.95 per 10 mmHg increase	Huynh <i>et al.</i> [96]
SBP	Risk score	HR, 0.78 per 10 mmHg increase	Felker <i>et al.</i> [95]
SBP	Risk score	HR, 0.85 per 10 mmHg increase	Lee <i>et al.</i> [97]
SBP	Risk score	Various thresholds	Hsich <i>et al.</i> [94]
SBP	Risk score	<115 mmHg	Fonarow <i>et al.</i> [93]
SBP	Risk score	≤ 130 mmHg; 10% increase per 10 mmHg decrease	O'Connor <i>et al.</i> [89]
SBP	Risk score	HR, 0.77 per 10 mmHg increase up to 160 mmHg	Abraham <i>et al.</i> [98]
SBP	Risk score	Various thresholds	Peterson <i>et al.</i> [21]
SBP	Risk score	<110 mmHg;	Bouvy <i>et al.</i> [99]
SBP	Risk score	HR 0.80 per 10 mmHg increase until 160 mmHg	Levy <i>et al.</i> [85]
SBP	Risk score	HR, 0.93 per 10 mmHg increase	Manzano <i>et al.</i> [87]
SBP	Risk score	HR, 0.88 per 10 mmHg increase	Pocock <i>et al.</i> [88]
DBP	Tele-monitoring	>110 mmHg	Roth <i>et al.</i> [82]
DBP	Risk score	HR, 0.86 per 10 mmHg increase up to 80 mmHg	Manzano <i>et al.</i> [87]
DBP	Risk score	HR, 0.73 per 10 mmHg increase up to 100 mmHg	Abraham <i>et al.</i> [98]
DBP	Risk score	<70 mmHg	Bouvy <i>et al.</i> [99]
DBP	Risk score	HR, 0.94 per 5 mmHg increase	O'Connor <i>et al.</i> [84]

Table 4.2: Evidence for the predictive value of blood pressure markers for HF deterioration. SBP = Systolic blood pressure. DBP = Diastolic blood pressure. HR = Hazard ratio.

4.2.1.3 Heart Rate Measurements

Increasing the resting heart rate is another coping mechanism of the body to reduce oxygen deprivation caused by a weakened heart; however, this poses additional strains on the myocardium. In addition to high resting heart rates, low resting heart rates

(e.g. $<50\text{bpm}$ [91]) can cause low oxygen circulation in the body and may therefore put patients at risk of breathlessness and cardiac death. The evidence on the predictive value of heart rate is ambiguous with some research suggesting that heart rate is an independent predictor of morbidity and mortality in HF patients [86][100] and others suggesting it is not [89].

Böhm *et al.* [100] analysed data from 6,505 participants in a randomised controlled trial, which evaluated the performance of a heart-rate lowering drug. The authors found that in the control group patients with resting heart rates ≥ 87 bpm had more than twice the risk of experiencing cardiovascular death or hospitalisation in comparison to patients in the lowest heart rate quintile (70-72 bpm, $p = 0.352$). Patients' risk of experiencing an event increased 16% with every 5 bpm increase in heart rate [100]. However, the predictive value of heart rate is not unambiguous, and O'Connor *et al.* found that heart rate was not associated with post-discharge mortality in the OPTIMIZE-HF dataset (4,402 subjects) [89]. Kotecha *et al.* [101] conducted a meta-analysis on eleven randomised controlled trials to study the association between heart rate and six-month all-cause mortality in HF with reduced ejection fraction and sinus rhythm or atrial fibrillation. The authors found an increased risk of mortality in patients in sinus rhythm, however not in HF patients with atrial fibrillation. Table 4.3 provides an overview of heart rate thresholds identified to be associated with HF deterioration.

4.2.1.4 NYHA Classification Scores

NYHA is a marker of breathlessness in HF and has been used to determine the severity of the condition [9]. A number of risk models have found NYHA classification to be associated with long-term hospitalisation and death in HF patients [102][87][88]. Models have included either the absolute NYHA score or analysed the risk of patients with higher NYHA scores in comparison to others (see Table 4.4). The predictive performance of NYHA as a marker for HF adverse event has not been evaluated to date.

Measurement	Source type	Threshold	Study/ Author
Heart rate	Guideline	<40 bpm or >70 bpm	Ponikovski <i>et al.</i> [9]
Heart rate	Tele-monitoring	<50 bpm or >80 bpm.	Cleland <i>et al.</i> [91]
Heart rate	Risk score	16% per 5 bpm increase	Böhm <i>et al.</i> [100]
Heart rate	Risk score	OR, 1.1 per 10 bpm between 65 and 110 bpm	Abraham <i>et al.</i> [98]
Heart rate	Risk score	HR, 1.11 per 10 bpm	Kotecha <i>et al.</i> [101]
Heart rate	Risk score	Various thresholds	Peterson <i>et al.</i> [21]
Heart rate	Risk score	HR, 1.08 per 10 bpm increase	Pocock <i>et al.</i> [88]
Heart rate	Risk score	HR, 1.13 per 10 bpm increase	Wedel <i>et al.</i> [86]

Table 4.3: Evidence for the predictive value of blood pressure-based markers for HF deterioration. OR = Odds ratio. HR = Hazard ratio.

Measurement	Source type	Change	Study/ Author
NYHA	Risk score	HR, 1.32 for NYHA class III; HR, 1.54 for NYHA class IV	Pocock <i>et al.</i> [88]
NYHA	Risk score	HR, 2.40 per NYHA increase	Levy <i>et al.</i> [85]
NYHA	Risk score	HR, 1.94 for NYHA class IV	Felker <i>et al.</i> [95]
NYHA	Risk score	HR, 1.40 per NYHA increase	Wedel <i>et al.</i> [86]
NYHA	Risk score	HR, 1.70 per NYHA increase	Bayes-Genis <i>et al.</i> [102]
NYHA	Risk score	OR, 4.09 NYHA class III or IV	Senni <i>et al.</i> [103]
NYHA	Risk score	HR, 1.70 per NYHA increase	Manzano <i>et al.</i> [87]
NYHA	Risk score	HR, 0.79 for NYHA I; HR, 1 for NYHA II; HR, 1.41 for NYHA III; HR, 11.68 for NYHA IV	Pocock <i>et al.</i> [88]

Table 4.4: Evidence for the predictive value of NYHA-based markers for HF deterioration

4.2.1.5 Early Warning Systems

As presented in the previous sections, individual HF measurements, such as weight, blood pressure, heart rate and NYHA, may contain some meaningful information on patients' short- and long-term likelihood of deterioration; however evidence remains ambiguous and there is limited information, if any, on the predictive performance of

each individual measurement. Tele-monitoring systems have been employed to monitor how the different physiological measurements and symptom scores (NYHA) vary over time and to use this information to optimise HF therapy. These systems primarily used guideline-based or proprietary thresholds to identify patients at risk of deterioration.

In recent years, a limited number of published, non-invasive early warning systems (that go beyond guideline-based thresholds) have been developed based on the monitoring of several physiological markers. Anand *et al.* developed a multivariate algorithm for the early detection (24-hour ahead of adverse event) of HF deterioration [104]. HF patients taking part in the study were asked to wear multi-sensor patches collecting bio-impedance, activity and heart rate data for a period of 90 days. Patients were initially shown how to apply the patches and then required to change them on a weekly basis. The developed early-warning system achieved a sensitivity of 63% and a specificity of 92% on a validation group of 200 HF patients [105]. Limited information is available on the algorithm itself.

Javed *et al.* developed an early warning system for HF patients using nocturnal respiratory data. Their system was based on a change-point-detection approach and triggered a warning if changes in the probability distribution of the respiratory signal were identified. The system was tested on 104 patients with NYHA classifications of II-IV and achieved a sensitivity of 55% and a specificity of 73% in predicting worsening of HF up to three weeks ahead of the event [106].

Finally, Koulaouzidis *et al.* [107] developed an early warning system using weight, blood pressure and heart rate signals. The authors used data from 308 tele-monitoring patients with a total of 27 HF-related adverse events to develop a method to predict HF hospitalisations one day ahead of the event and found that a combination of normalised weight and diastolic blood pressure markers collected in the eight days before the hospitalisations achieved the best performance (AUC = 0.82, sensitivity = 52%, specificity

= 98%). The weight and blood pressure markers were vectors of repeatedly filtered and down-sampled measurements across the eight-day period (see Figure 4.1). According to the authors, this approach was chosen to resemble human behaviour when analysing the data, i.e. looking at the broader trends as well as the more detailed changes.

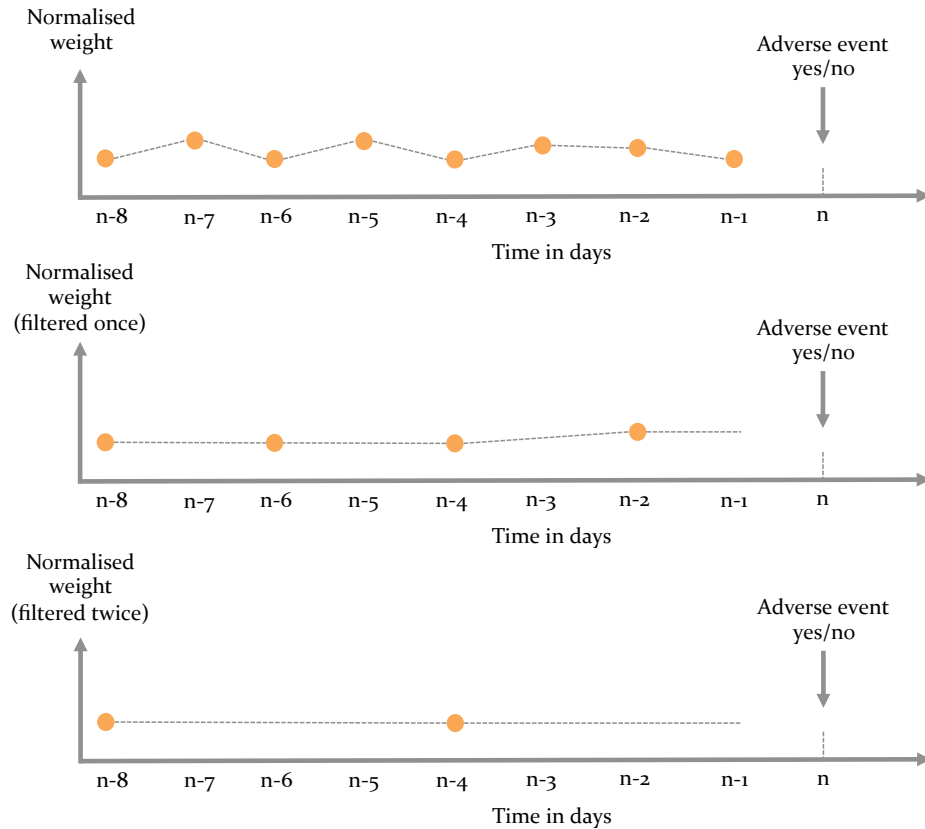


Figure 4.1: Visual representation of the repeated filtering and down-sampling method used by Koulaouzidis et al. [107].

The final model included input vectors of length 28 (14 values each for weight and diastolic blood pressure measurements) and was trained and subsequently tested on a dataset with 9,870 cases without adverse events and 15 cases with adverse events. As the training and test sets were randomly selected using a 90%:10% split (the process was repeated 100 times), only a very small number, if any, of adverse events would have been expected to have been assigned to the test set in any iteration. It is very unusual to train a model for which there are more input variables than examples of one particular

class. Unless specific steps are taken, it is inevitable that overfitting will occur. In addition to this, the clear difference in class sizes was not addressed when training the classifier. Despite the methodological flaws of the study, the authors reported good classification performance using the unprocessed weight and diastolic blood pressure values, and further improved performance when using the repeated filtering method. Whilst it is likely that the overall performance of the model was heavily biased by overfitting, the proposed variables may still be interesting variables to study in the future.

The research described for the remainder of this chapter used similar measurements to those of Koulaouzidis *et al.*; however it also included the NYHA symptom score. Using markers that quantify the three-day changes in measurements the analysis aimed to develop and evaluate early warning methods suitable for HF tele-monitoring systems. Due to the limited size of the dataset studied for the remainder of this chapter, it was decided not to extract the eight-day variables proposed by Koulaouzidis *et al.*

4.3 Methodology

4.3.1 Database

Data from the SHF2 study were used for the development and evaluation of the early warning methods. As described in Section 2.2, the SHF2 study was a randomised controlled trial with the aim of evaluating the effectiveness of a home monitoring system in optimising medical therapy and HF outcomes. In the period between October 2015 and September 2017, HF patients across seven hospital sites in the UK were recruited into the study.

Patients with a verified diagnosis of HF who could benefit from taking part in the home monitoring study were eligible to take part in the study. Such patients were

defined as patients who were not on optimal treatment or had severe symptoms of HF. After providing informed consent, patients were randomly assigned to the intervention or control group. They were provided with a Bluetooth-enabled weighing scale and blood pressure monitor and asked to take daily weight and blood pressure measurements and answer a two-question questionnaire to determine their NYHA functional class.

Patients in both the intervention and control groups were provided with the SHF2 self-monitoring platform. In addition to this, patients in the intervention group received advice on optimal medication use as well as feedback in response to changes in measurements suggesting HF deterioration (see Section 2.2.2.1 for more detail). The intended monitoring period for randomised patients in the study was six months. After this period patients were able to continue in the study or join *passive follow-up*. Adverse event data for all patients were collected every six months and at the end of the trial.

4.3.2 Data Preparation

Unless stated otherwise all data analysis was conducted in Python.

4.3.2.1 Data Cleaning

Weight Data

Weight data were collected in units of *kg* or *lb* and converted to *kg* before further processing. Subsequently potential false values defined as weight values of ≤ 30 kg or day-to-day changes in weight of ≥ 5 kg were removed from the dataset. The requirement of daily weight changes to be smaller than 5 kg was introduced to remove unusual values that may have been caused by individuals other than the study participants (e.g. family members or visitors) using the weighing scale.

Blood Pressure and Heart Rate Data

According to clinical recommendations, physiologically plausible blood pressure and heart rate values were defined as:

- Systolic blood pressure values ≥ 50 mmHg.
- Diastolic blood pressure values ≥ 25 mmHg.
- Heart rate values ≥ 30 bpm.

Values outside of the normal ranges were removed from the analysis as they were likely to be caused by erroneous measurements.

4.3.2.2 Missing Value Imputation

Data entry was reliant on patient adherence. Patients may occasionally forget to take their measurements or miss measurements for other reasons. Based on clinical considerations, a last observation carried forward (LOCF) approach was used to impute values for up to three days after the last taken measurement. Figure 4.2 provides a visual representation of the data imputation process.

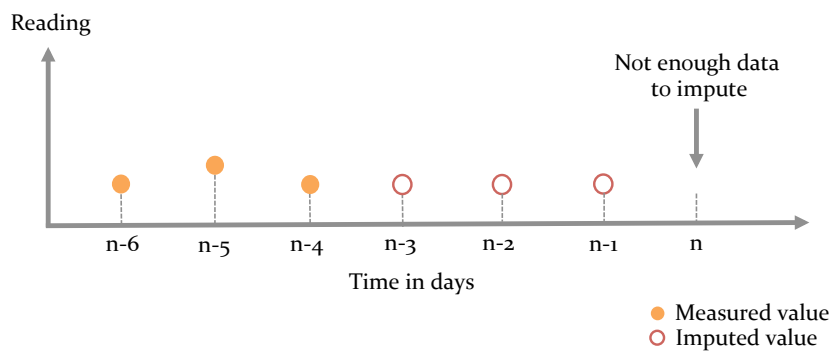


Figure 4.2: Visual representation of the last observation carried forward (LOCF) data imputation process.

4.3.2.3 Variable Extraction

Due to the limited size of the SHF2 dataset only a small number of heuristically-chosen variables were extracted for analysis. In particular, the three-day change of each measurement was selected as the marker of interest. According to European guidelines, three-day changes in weight (Δw) are considered an important marker for HF deterioration [9]. In addition to this, changes in NYHA, blood pressure and heart rate may also be indicative of disease worsening. Three-day changes were calculated as shown in Equations 4.1 - 4.3 below, where w_n corresponds to the weight value on day n of the recording from an individual.

$$\Delta w_n = w_n - w_{n-3} \quad (4.1)$$

$$\Delta w_{n-1} = w_{n-1} - w_{n-4} \quad (4.2)$$

$$\Delta w_{n-2} = w_{n-2} - w_{n-5} \quad (4.3)$$

One-off changes in measurements may be affected by false readings or outliers and the analysis described in this chapter therefore used the three-day average (δw_n) of three-day changes as illustrated in Figure 4.3 and shown below:

$$\delta w_n = \frac{1}{3}(\Delta w_n + \Delta w_{n-1} + \Delta w_{n-2}) \quad (4.4)$$

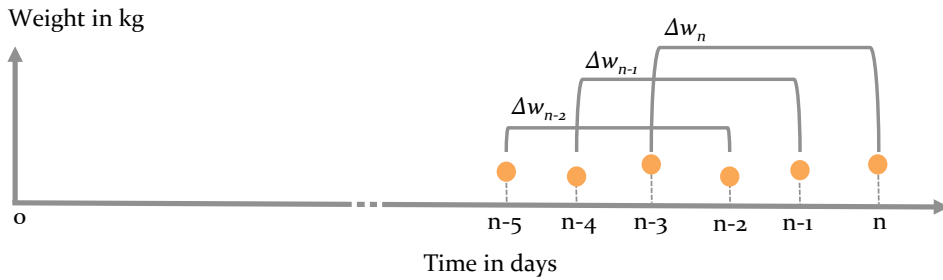


Figure 4.3: Visual representation of the three-day change marker using the example of weight measurements.

The average change markers were extracted for all daily measurements (weight (δw_n), systolic blood pressure (δsbp_n), diastolic blood pressure (δdbp_n), heart rate (δhr_n)) and the NYHA classification ($\delta nyha_n$) to obtain a vector:

$$\delta \mathbf{x} = \{\delta w_n, \delta sbp_n, \delta dbp_n, \delta hr_n, \delta nyha_n\} \quad (4.5)$$

Information may not only be contained in the average three-day change of a measurement but also in the variability of change, e.g. large variations in values may be a sign of instability and may occur ahead of disease deterioration. Such variations can be quantified using the (unbiased) standard deviation of change as shown for the example of weight (σw_n) below:

$$\sigma w_n = \frac{1}{2} \sqrt{(\Delta w_n - \delta w_n)^2 + (\Delta w_{n-1} - \delta w_n)^2 + (\Delta w_{n-2} - \delta w_n)^2} \quad (4.6)$$

The standard deviation of change markers was also extracted for all daily measurements (weight (σw_n), systolic blood pressure (σsbp_n), diastolic blood pressure (σdbp_n), heart rate (σhr_n)) and the NYHA classification ($\sigma nyha_n$) to obtain a vector:

$$\sigma \mathbf{x} = \{\sigma w_n, \sigma sbp_n, \sigma dbp_n, \sigma hr_n, \sigma nyha_n\} \quad (4.7)$$

Both the average and the standard deviation of three-day change values do not take into account a subject's previous characteristics. In a subject with a history of large three-day change variations, a similarly large change may be less meaningful than in a subject with a history of much smaller changes. As part of this analysis, patient-specific markers of the average and standard deviation of three-day changes in measurements were therefore extracted by scaling the current average/ standard deviation of three-day changes by the previous average/ standard deviation of three-day changes, with a three-day guardband between the previous (long-term) and the current (short-term) values.

This extraction process is illustrated in Figure 4.4 and in the Equations below:

$$\phi w_n = \frac{\delta w_n}{\frac{1}{m} \sum_{i=0}^m |\Delta w_i|} \quad (4.8)$$

$$\psi w_n = \frac{\sigma w_n}{\frac{1}{m-1} \sqrt{\sum_{i=0}^m (\Delta w_i - \mu w_m)^2}}, \quad (4.9)$$

where $m = n - 6$ corresponds to the length of the long-term data and

$$\mu w_n = \frac{1}{m} \sum_{i=0}^m \Delta w_i. \quad (4.10)$$

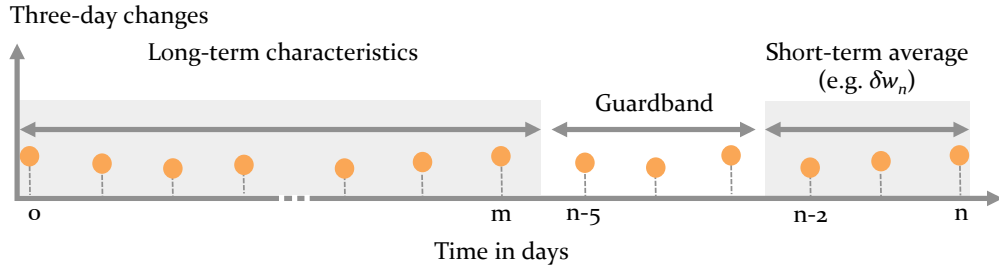


Figure 4.4: Visual representation of the 'scaled' three-day marker extraction with guardband.

The scaled change markers were extracted for all daily measurements to obtain the vectors:

$$\phi \mathbf{x} = \{\phi w_n, \phi sbp_n, \phi dbp_n, \phi hr_n, \phi nyha_n\} \quad (4.11)$$

$$\psi \mathbf{x} = \{\psi w_n, \psi sbp_n, \psi dbp_n, \psi hr_n, \psi nyha_n\} \quad (4.12)$$

In addition to the variables based on three-day averages/ standard deviations of three-day changes, another set of variables using seven-day average/ standard deviation of three-day changes was extracted. These variables were also investigated as they may

capture longer-term trends in three-day changes. They were calculated similarly to the three-day markers with the short-term (current) average now ranging from n to $n - 6$, the guardband from $n - 7$ to $n - 13$ and $m = n - 14$. Both the three-day and seven-day markers were calculated only for m greater than 14 in order to ensure sufficient datapoints were included in the calculation of the long-term characteristics of each measurement.

4.3.2.4 Adverse Events

Adverse events were defined as HF-related hospitalisations and deaths, including events labelled as atrial fibrillation, dizziness, device implantation, myocardial infarction, cardiac death, decompensated HF, renal failure or dehydration. The different adverse event groups were chosen based on clinical recommendations and further information on the number of adverse events in each category is provided in the results section of this chapter.

4.3.3 Data Analysis

4.3.3.1 Data Exploration

The purpose of the analysis was to develop and evaluate methods capable of predicting HF-related adverse events. Three methods for HF deterioration were developed to study the combined predictive power of weight, NYHA, blood pressure measurements and heart rate. These were (a) a data-driven method using a logistic classifier with three-day input variables and (b) the same method, but using seven-day input variables, (c) a method using guideline-based thresholds.

For each participant, input variables were extracted for each day of study enrolment with a complete set of values after data imputation. Each day was given an outcome label 1 or 0 depending on whether or not a HF-related adverse event occurred on the following day ($n+1$ on Figure 4.5). As measurements immediately after an adverse event

may be affected by this adverse event, the 14 days after a HF-related adverse event were excluded from the analysis. Days with missing input variables after imputation were also excluded from the analysis.

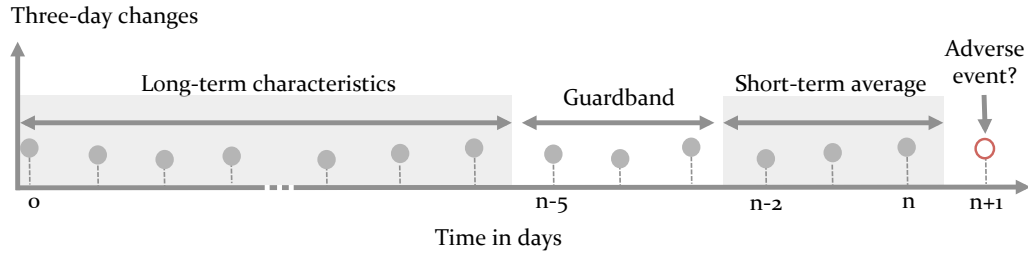


Figure 4.5: *Visual representation of adverse event prediction.*

To evaluate the performance of both the logistic classifiers and the guideline-based method, the dataset was divided into a training and a test set so that a percentage of participants were randomly assigned to the training set and the remaining participants were assigned to the test set (see Figure 4.6). The logistic classifiers were trained on the training set and all methods were tested on the test set. The process of selecting a random subset of participants for the training and test sets to evaluate the performance of the models was repeated 100 times in order to ensure that the performance measures would not be affected by chance similarities or differences between the training and test sets.

The SHF2 dataset only contains a small number of days followed by adverse events. The relative proportions of participants in the training and test sets can therefore significantly influence the performance of the classifier. As illustrated in Figure 4.7, choosing a small number of samples on which to train a classifier can lead to overfitting on the training data. Depending on which subset of event-day data are in the training set, very different decision boundaries between event and non-event days may be found. Choosing a larger proportion of samples to be included in the training set can alleviate this prob-

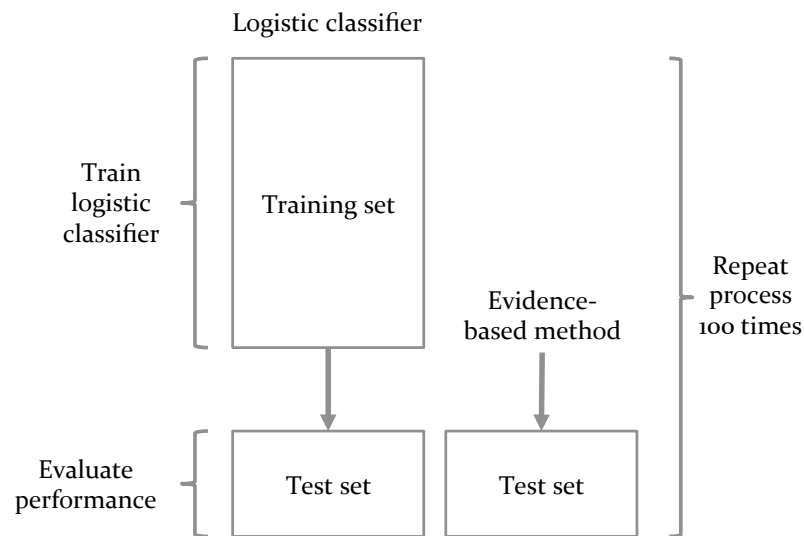
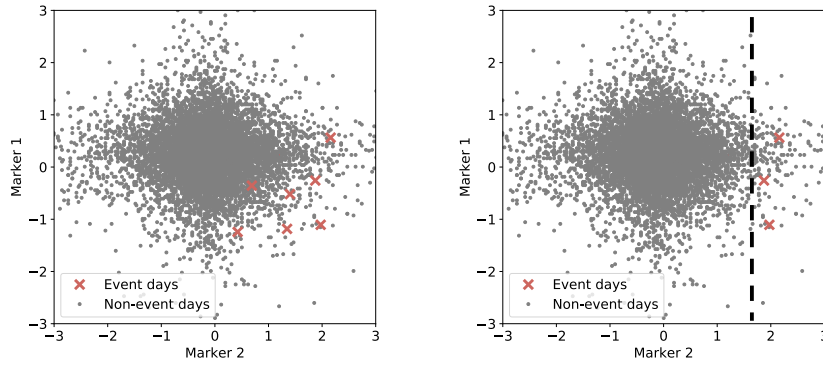
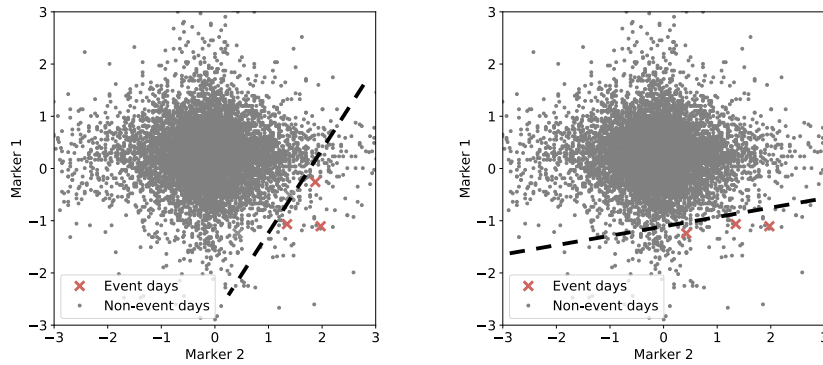


Figure 4.6: *Flow-chart of the early warning methods analysed in this chapter. The guideline-based threshold method was applied to the test set without any prior training whilst the classifiers required training.*

lem, however at the cost of having fewer samples in the test set with which to evaluate the trained classifier. For the purpose of the analysis described in this chapter, a range of dataset partitions were tested and evaluated based on the mean performance of the classifier on the training and the test sets of the 100 randomly selected partitions.



(a) *Example full dataset, with only a small number of event cases.* (b) *Example partition one, with three event days selected for the training set.*



(c) *Example partition two, with three different event days selected for the training set.* (d) *Example partition three, with yet another three different event days selected for the training set.*

Figure 4.7: *Visualisation of the importance of training and test set sizes. The dashed lines in (b), (c) and (d) shows a possible linear decision boundary in each case.*

4.3.3.2 Logistic Classifiers

The 'lme4' package in R [108] was used for the logistic classifiers. Two different sets of input variables were evaluated: 1) the three-day markers and 2) the seven-day markers. All input markers were normalised using a zero-mean and unit variance transformation for each marker in the training set. Standard-deviation markers (e.g. σw_n , σsbp_n , ..., ψw_n , ψsbp_n ...) are by definition always positive. These markers were also normalised

using the zero-mean and unit variance transformation and subsequently shifted so that the minimum value was ≥ 0 . This was done so to maintain the interpretability of the markers.

The SHF2 data is heavily unbalanced with more than 40,000 days of self-monitoring data and only 44 days followed by adverse events. For a subgroup of adverse event days, insufficient self-monitoring data were available further reducing the total number of adverse events included in the analysis (more detail on the adverse events included in the analysis are presented in the results section). As classifiers optimise their decision boundaries to reduce the mean squared error between the predicted and the true outcome scores, an unbalanced dataset may result in the prioritisation of the no-event classification over the event classification. In order to allow for the classifier to learn both classes, the training set and test set were balanced. For the training set, all days with associated next-day adverse events were identified; for each selected event-day a randomly selected non-event day was chosen from the remaining training set data. The trained classifier was then evaluated on the test set data, balanced using the same approach as for the training set data.

4.3.3.3 Evidence-based Thresholds

This method was based on established clinical thresholds used to identify HF patients in need of medication adjustment. Thresholds included:

- Weight increase of > 2 kg in 3 days,
- Blood pressure values of $\geq 140/90$ mmHg
- NYHA classification of 4.

To evaluate the performance of these well-established thresholds, individual variables were created for each threshold and set to 1 (predicted adverse event) whenever the threshold was met or exceeded and to 0 (no predicted adverse event) otherwise.

4.3.4 Evaluation

The primary outcome measure of the analysis was the AUC of the ROC (see Section 3.6.4). In addition to this, PDFs of classifications were created to provide further insight into performance. The mean classification score per type of adverse event was extracted in order to provide an insight into how well the classifiers identified each of the different types of adverse events.

4.4 Results

4.4.1 Dataset

By the end of recruitment, 202 HF patients were randomised into the SHF2 study, accounting for 41,448 days worth of tele-monitoring data. Patients had a median (interquartile range) age of 73 (64-81) years and 28% of the participants were female. A total of 34,494 weight values (median (IQR) of 173 (122-200) per patient), 31,139 NYHA values (155 (106-186) per patient) and 34,331 blood pressure measurements (this includes systolic and diastolic blood pressure values and heart rate measurements) (171 (122-200) per patient) were recorded. At least one of the measurements was available for 35,956 days (84% of days) with a median percentage of missing days of 10 (3-22) across participants. After data cleaning, 39,493 patient days with at least one measurement of weight, NYHA and blood pressure values were captured within the dataset. This included 5,404 days with imputed data (2,876 imputed weight values, 4,548 imputed NYHA scores and 2,897 imputed blood pressure values).

4.4.2 Alerts

During the duration of the SHF2 study, the clinical team received 6,230 patient deterioration alerts for patients in the intervention group: 641 (median of 3 (1-9) per patient) as a result of abnormal weight increases, 468 (0 (0-7) per patient) as a result of NYHA

increases, 3,729 (14 (1-56) per patient) as a result of systolic blood pressure values and 1,392 (2 (0-10) per patient) as a result of diastolic blood pressure values. A large number of blood pressure alerts were generated because the system used hard thresholds that did not take into account patient physiology, e.g. a patient with hypertension would trigger daily alerts. A recent retrospective study by Murphy *et al.* [109] showed similar results, with population-based threshold alerts leading to large numbers of false alerts in a tele-monitoring study of 187 veterans with HF. Using a system that can learn the normal distribution of an individual patient's measurements and adjust alert thresholds accordingly would reduce the number of false alerts [110]. This was however not part of the SHF2 protocol.

Instead of responding to every alert, the SHF2 clinical team used clinical judgement and their knowledge of patients to determine when interventions were required. As a result of this clinical judgement, a total number of 110 messages were sent to a total of 47 patients in the intervention group (a median of 2 (1-2) messages for the 47 patients). Subsequent medication changes (within 30 days of the message) were recorded 64 times for 30 patients (median of 2 (1-2) for the 30 patients).

4.4.3 Typical Results

In some cases, clinical intervention led to an improvement in physiological measurements as shown in the example of patient 159 (see Figure 4.8). Here, the patient experienced high heart rates for a number of days immediately after study enrolment. The patient was recommended an increase in beta blockers and as a result of the medication change their heart rate stabilised for the remainder of the study.

In other cases, the patient's responses were delayed or subsequent adverse events could not be prevented. Figure 4.9 shows the weight data for a patient (patient 179) who experienced a gradual increase in weight. After a short-term rise in weight, the

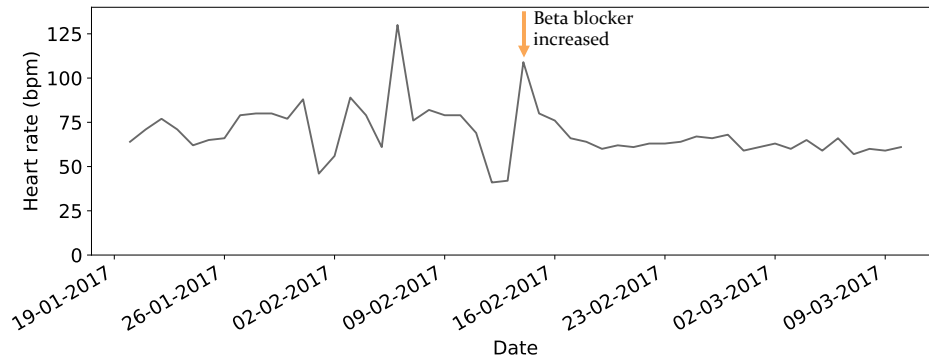


Figure 4.8: Heart rate values of patient 159 before and after medication change.

patient was advised to increase their dose of diuretics. This led to a brief recovery of their weight; however a week after the intervention, the patient continued to gain weight and eventually had to be hospitalised with fluid overload. All other values, including NYHA scores, remained stable over the decompensation period.

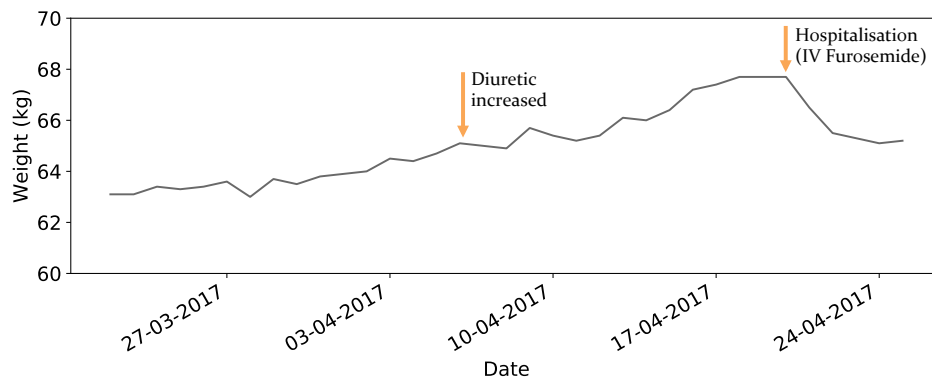


Figure 4.9: Weight values for patient 179 before and after medication change.

During the SHF2 study, 44 HF-related adverse events occurred in 38 participants (events that occurred in the 14 days after a previous adverse event were not included in this number ($n=3$)). 23 of the 44 events, including six HF-decompensations and three renal dysfunctions, in 18 participants, were experienced by participants in the intervention group. Another 21 events, including three HF-decompensations, in 20 participants were experienced by patients in the control group (see Tables 4.5 and 4.6).

The tele-monitoring intervention specifically targeted patients who experienced changes in weight and/ or NYHA symptom score and/or blood pressure. As a result, the greatest effect of the intervention would have been expected to be on the number of adverse events as a result of HF-deterioration, dizziness or renal dysfunction. Whilst some adverse events may have been prevented through early intervention, the total number and types of adverse events occurring in the intervention and control groups suggest that there were limited differences, if any. As a result, for the remainder of this chapter the analysis of adverse events was conducted for the entire study population.

Type	Total (n)	Death (n)	Life threatening (n)	Hospitalisation (n)	Other (e.g. check-up) (n)
Atrial fibrillation	2	-	-	2	-
Dizziness	-	-	-	-	-
Chest infection	-	-	-	-	-
CRT-D implantation	-	-	-	-	-
CRT-P implantation	1	-	-	1	-
Heart failure deterioration	6	2	-	4	-
Hypotension	-	-	-	-	-
ICD implantation	-	-	-	-	-
Myocardial infarction	-	-	-	-	-
Renal dysfunction	3	-	-	3	-
Ventricular tachycardia	1	-	-	1	-
Other	10	2	-	3	5

Table 4.5: *Types and severity of adverse events that occurred in the intervention group, where n indicates the number of cases falling in each category. More information on the category 'other' can be found in Section 4.4.4.*

Type	Total (n)	Death (n)	Life threatening (n)	Hospitalisation (n)	Other (e.g. check-up) (n)
Atrial fibrillation	-	-	-	-	-
Dizziness	1	-	-	-	1
Chest infection	-	-	-	-	-
CRT-D implantation	1	-	-	1	-
CRT-P implantation	-	-	-	-	-
Heart failure deterioration	3	-	-	2	1
Hypotension	-	-	-	-	-
ICD implantation	-	-	-	-	-
Myocardial infarction	2	-	-	2	-
Renal dysfunction	-	-	-	-	-
Ventricular tachycardia	-	-	-	-	-
Other	14	1	1	7	5

Table 4.6: *Types and severity of adverse events that occurred in the control group, where n indicates the number of cases falling in each category. More information on the category 'other' can be found in Section 4.4.4.*

4.4.4 Adverse Events

Table 4.7 provides an overview of the SHF2 participants who experienced HF-related adverse events. The first column ('*All adverse events*') shows the summary characteristics for all the participants who experienced HF-related adverse events in the study. The second column shows data for those who experienced HF-related adverse events and for whom input variables could be extracted for the day before the event ($n=26$ subjects and 30 adverse events). These were the participants whose data were used to evaluate the three-day and seven-day markers and train the multi-variate classifiers.

There was no significant difference in age ($p = 0.17$ (t-test)) across participants with and without HF-related adverse events; however, patients with HF-related adverse events were more likely to be female ($p < 0.01$ (χ^2 test)) and have been diagnosed with chronic kidney disease ($p < 0.01$ (χ^2 test)).

	All participants (n = 44)	All adverse events (n = 44)	Risk-prediction group (n = 30)
Subjects	202	38	26
Age, years (median (IQR))	73 (64-81)	70 (62-78)	72 (66-78)
Sex, female	28%	33%	35%
Reduced EF	60%	64%	69%
MAGGIC score	23 (18-28)	24 (18-27)	25 (20-28)
Comorbidities			
Hypertension	47%	58%	58%
DM	32%	39%	38%
COPD	17%	13%	15%
CKD	15%	32%	31%
Technical skills			
Poor	44%	34%	38%
Intermediate	48%	63%	62%
Expert	8%	3%	0%
Monitoring duration, days (median (IQR))	200 (158-232)	189 (129-212)	191 (129-209)
Missing values (median (IQR))	10 (3-22)	18 (6-32)%	18 (5-27)%
Monitoring days after imputation (median (IQR))	195 (153-222)	169 (105-201)	179 (110-204)

Table 4.7: *Baseline characteristics and compliance and adverse event information for participants randomised into the SUPPORT-HF2 study. Data are stratified by adverse event occurrence. n refers to the number of adverse events in each group. IQR = interquartile range. DM = Diabetes mellitus. CKD = Chronic kidney disease.*

Table 4.8 lists the 30 adverse events included in the risk-prediction analysis. It can be seen that a majority of events ($n = 15$) did not fit into the clinician-defined categories and were therefore classed as 'other'. The category 'other' ranged from elective check-up events to sudden death. The latter can be caused by a range of factors, including myocardial infarction or arrhythmias, and was therefore assigned to the group 'other'. Five adverse events were due to HF deterioration, a majority of which resulted in hospitalisations. The listed adverse events also include potentially elective procedures such

as implantations of cardio resynchronisation devices with fitted defibrillators (CRT-D) or pacemakers (CRT-P).

Type	Total (n)	Death (n)	Life threatening (n)	Hospitalisation (n)	Other (e.g. check-up) (n)
Atrial fibrillation	1	-	-	1	-
Dizziness	1	-	-	-	1
Chest infection	-	-	-	-	-
CRT-D implantation	1	-	-	1	-
CRT-P implantation	1	-	-	1	-
Heart failure deterioration	5	-	-	4	1
Hypotension	-	-	-	-	-
ICD implantation	-	-	-	-	-
Myocardial infarction	2	-	-	2	-
Renal dysfunction	3	-	-	3	-
Ventricular tachycardia	1	-	-	1	-
Other	15	2	-	8	5

Table 4.8: *Types and severity of adverse events investigated for the prediction analysis, where n indicates the number of cases falling in each category.*

4.4.5 Data Analysis

Two logistic classifiers and one guideline-based model were investigated with respect to their ability to detect HF-related adverse events one day ahead of their occurrence. The logistic classifier was run a) using the three-day input vectors and b) the seven-day input vectors introduced in Section 4.3.2.3. The logistic classifier models trained on the three-day and seven-day variables went beyond the hard thresholds used by the guideline-based model by learning the multi-variate decision boundaries that best fit the training sets. The boundaries learnt by the classifier may differ depending on the selection of event and non-event samples used to train the classifier (see Section 4.3.3.2). When there are only a small number of event-days used to train the classifier (e.g. on average nine event days for a 30:70 split of the dataset into training and test sets), the limited variety in characteristics of these samples may amplify this effect. Depending on which small subsample of event-days is randomly selected into the training set, very

different decision boundaries may be chosen resulting in differences in performance as measured using the AUCs on training and test sets.

Table 4.9 lists the AUCs achieved on the training and test sets. Generally, the classifier using $\delta\mathbf{x}$ and $\sigma\mathbf{x}$ (together) would be expected to fit the training data more closely due to its larger number of input markers. For both the $\delta\mathbf{x}$ and $\delta\mathbf{x}$ and $\sigma\mathbf{x}$ markers, it can be seen that using a small proportion (i.e. 30:70 partition of the dataset) of samples in the training set resulted in the best mean AUC on the training sets; however the corresponding performances on the test sets were poor, showing overfitting to the training data. Figure 4.10 shows the training set performance of the logistic classifier using three-day input markers (i) $\delta\mathbf{x}$, ii) $\delta\mathbf{x}$ and $\sigma\mathbf{x}$) for different partitions of training and test sets (50:50, 70:30, 90:10). The shaded bands referring to the mean $\pm$ the standard deviation of the ROCs were also comparably large suggesting again that the performance of the classifiers was very dependent on the subset of samples used during training (overfitting phenomenon).

As more samples were used in the training set, the amount of overfitting and the variation in performance across the 100 random partition of training and test sets reduced, however the variation in the test set performance (as quantified using the standard deviation) increased as shown in Table 4.9. This is likely to have been caused by the increasingly smaller sample set on which the model was evaluated and hence the greater dependency on the choice of samples for the test set. Overall, the split of training and test set did not appear to have a large impact on the performance of the classifier on the test set and the ability to predict HF-related adverse events using the proposed markers was clearly limited. For the remainder of this analysis, it was decided to use the 70:30 split as this was seen to limit the amount of overfitting whilst maintaining a relatively large number of test set samples. In addition to this, the NYHA markers were removed from the classifiers as it was found that there was very little variation in these markers resulting in large values for the corresponding coefficients in the logistic classifier and

overfitting.

Split	Adverse events per group	Markers	AUC on training set Mean (std)	AUC on test set Mean (std)
30:70	8:21	$\delta\mathbf{x}$	0.78 (0.09)	0.52 (0.08)
40:60	12:18	$\delta\mathbf{x}$	0.74 (0.09)	0.55 (0.08)
50:50	15:15	$\delta\mathbf{x}$	0.71 (0.08)	0.54 (0.09)
60:40	18:12	$\delta\mathbf{x}$	0.69 (0.06)	0.53 (0.12)
70:30	21:8	$\delta\mathbf{x}$	0.67 (0.06)	0.52 (0.11)
80:20	24:6	$\delta\mathbf{x}$	0.67 (0.05)	0.56 (0.16)
90:10	27:3	$\delta\mathbf{x}$	0.65 (0.04)	-
30:70	8:21	$\delta\mathbf{x}$ and $\sigma\mathbf{x}$	0.92 (0.09)	0.51 (0.08)
40:60	12:18	$\delta\mathbf{x}$ and $\sigma\mathbf{x}$	0.89 (0.08)	0.52 (0.09)
50:50	15:15	$\delta\mathbf{x}$ and $\sigma\mathbf{x}$	0.84 (0.07)	0.52 (0.09)
60:40	18:12	$\delta\mathbf{x}$ and $\sigma\mathbf{x}$	0.81 (0.07)	0.54 (0.10)
70:30	21:8	$\delta\mathbf{x}$ and $\sigma\mathbf{x}$	0.78 (0.06)	0.55 (0.12)
80:20	24:6	$\delta\mathbf{x}$ and $\sigma\mathbf{x}$	0.77 (0.06)	0.55 (0.17)
90:10	27:3	$\delta\mathbf{x}$ and $\sigma\mathbf{x}$	0.76 (0.06)	-

Table 4.9: Performance of the logistic regression classifier when using 100 different random partitions of training and test sets for different proportions (from 30:70 to 90:10). Std = standard deviation. AUC = area under the receiver operating curve.

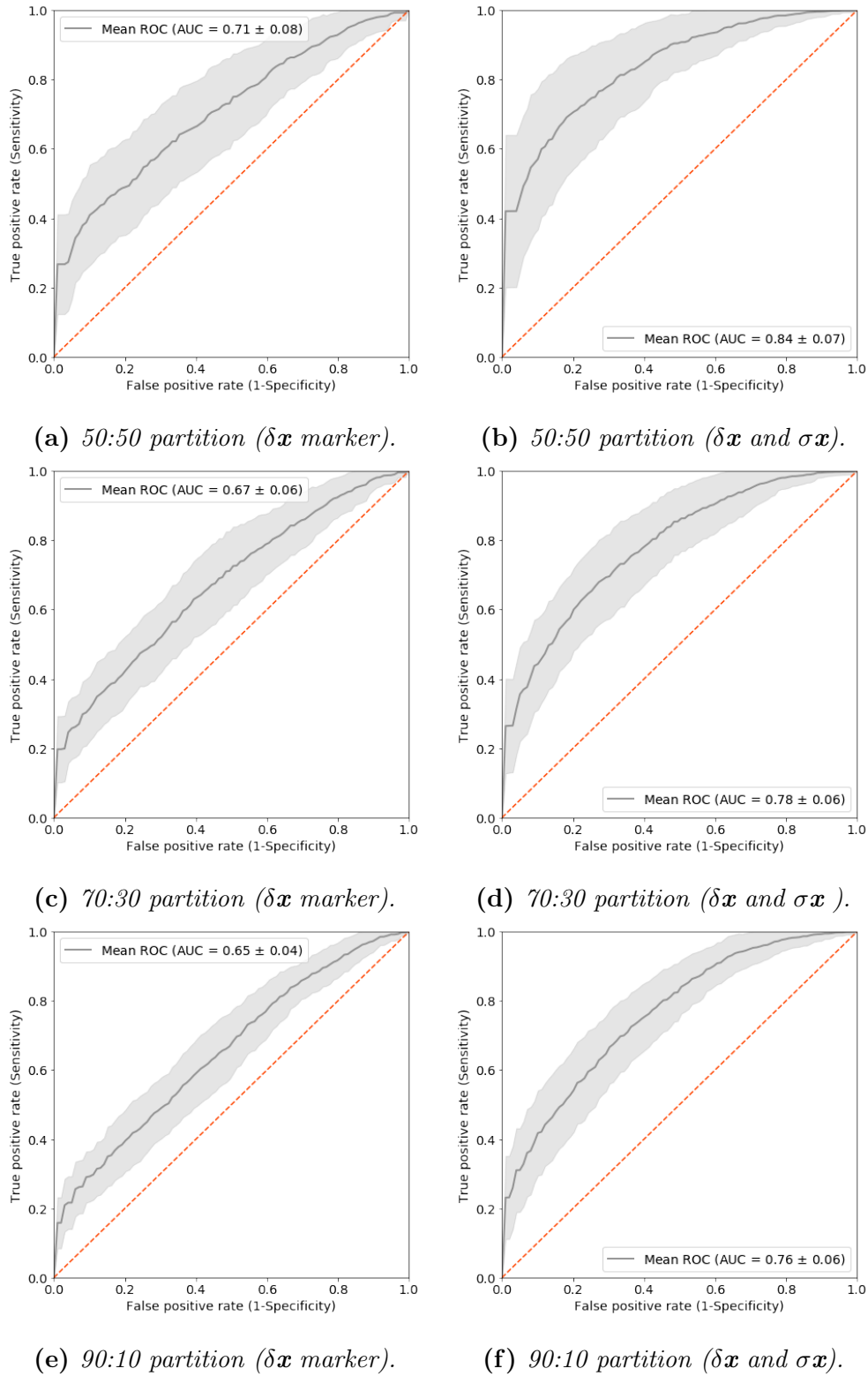


Figure 4.10: Receiver operating curves (ROCs) for different partitionings of training and test data. AUC = area under the receiver operating curve. The figures show the mean ROC curve as a solid grey line and the mean $\pm$ the standard deviation of the ROC curves as the grey shaded area.

Table 4.10 lists the average performance of the classifiers using the different input markers on the training and test sets using the 30:70 partitioning. Test set performances did not exceed AUCs of 0.55 implying moderate predictive performance of the classifiers. The three-day markers appeared to achieve on average higher AUCs than the seven-day markers. In addition to this, the classifiers using markers of standard deviation ($\sigma\mathbf{x}$ and $\psi\mathbf{x}$) outperformed the classifiers using average markers or the combined vectors as their inputs.

Markers	Three-day markers		Seven-day markers	
	Train AUC mean (std)	Test AUC mean (std)	Train AUC mean (std)	Test AUC mean (std)
$\delta\mathbf{x}$	0.64 (0.06)	0.48 (0.12)	0.61 (0.05)	0.45 (0.12)
$\sigma\mathbf{x}$	0.67 (0.05)	0.53 (0.15)	0.63 (0.06)	0.46 (0.12)
$\delta\mathbf{x}$ and $\sigma\mathbf{x}$	0.74 (0.06)	0.52 (0.14)	0.69 (0.06)	0.42 (0.10)
$\phi\mathbf{x}$	0.66 (0.06)	0.50 (0.13)	0.66 (0.06)	0.51 (0.12)
$\psi\mathbf{x}$	0.68 (0.07)	0.55 (0.14)	0.65 (0.06)	0.49 (0.11)
$\phi\mathbf{x}$ and $\psi\mathbf{x}$	0.75 (0.07)	0.53 (0.14)	0.72 (0.05)	0.47 (0.11)

Table 4.10: Performance of the logistic classifiers when using different three-day and seven-day input markers with a 70:30 partitioning into training and test sets. AUC = area under the receiver operating curve. Std = standard deviation.

In an attempt to develop the optimal logistic classifier based on the available SHF2 data, the classifier coefficients for the $\delta\mathbf{x}$ classifier and for the $\sigma\mathbf{x}$ classifier were investigated. Figure 4.11 shows the average (standard deviation) beta coefficients given to the different markers for the 100 random partitions using a 70:30 ratio between training and test sets. It can be seen that the weight and heart rate markers were assigned coefficients either side of zero, indicating that for some partitions the markers had a positive weight in the classifier and for other partitions the markers had a negative weight. The σsbp , σdbp and δdbp markers on the other hand were consistently given coefficients of similar size and sign. For example, an increase in the three-day mean of diastolic blood pressure changes was consistently associated with an increased risk of HF-related adverse events. Similarly, an increase in the variability of three-day change markers of systolic blood

pressure was associated with an increased risk of deterioration.

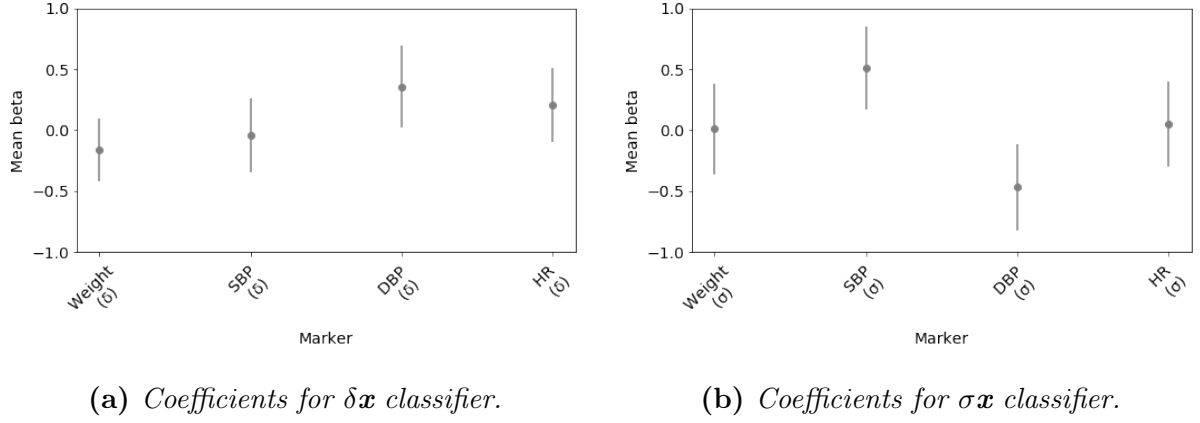
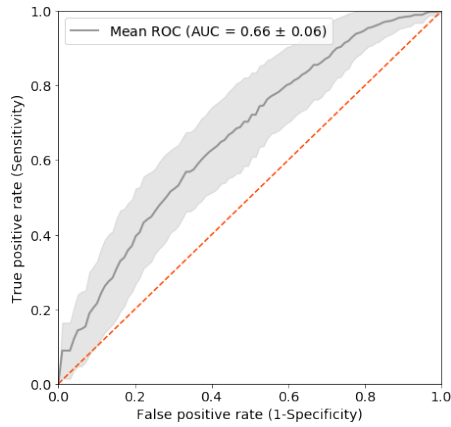
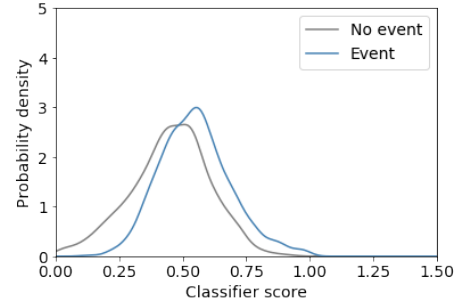


Figure 4.11: Mean and standard deviation of beta coefficients for the classifiers using $\delta \mathbf{x}$ and $\sigma \mathbf{x}$ markers as inputs. SPB = Systolic blood pressure, DBP = Diastolic blood pressure. HR = Heart rate.

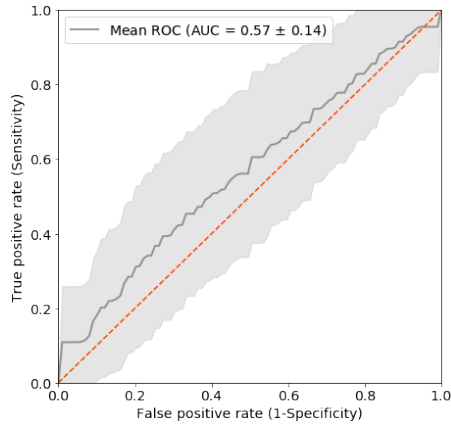
Based on these findings, a model using the three markers (σsbp , σdbp and δdbp) as its input was trained. The average performance of the model on the training and the test set is summarised in Figure 4.12. Figure 4.12a shows the performance of the classifier on the training set. The average AUC achieved on the training set by the classifier across the 100 partitions of the dataset was 0.66. The PDFs in Figure 4.12b show that some differentiation was possible. With respect to the performance of the trained classifiers on the test set, the average AUC was 0.57. The PDFs show that a small degree of differentiation was still possible, although there was a large proportion of HF event days that were given a similar classifier score as non-event days. This may have happened because some adverse events were easier to identify than others. Figure 4.12f suggests that 'myocardial infarction' is an example of an adverse event that was easier to identify (it achieved high average classifier scores), whereas the category 'other' was harder to detect. It is important to note that there were a limited number of adverse events in the dataset, with some adverse event groups (e.g. 'dizziness', 'CRT-P') only occurring once, and therefore no general statements can be made. The weights of the HF markers are shown in Figure 4.12e and were similar to those in the individual $\delta \mathbf{x}$ and $\sigma \mathbf{x}$ models.



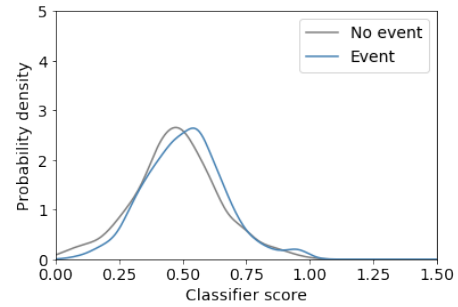
(a) Receiver operating curve (ROC) on the training set.



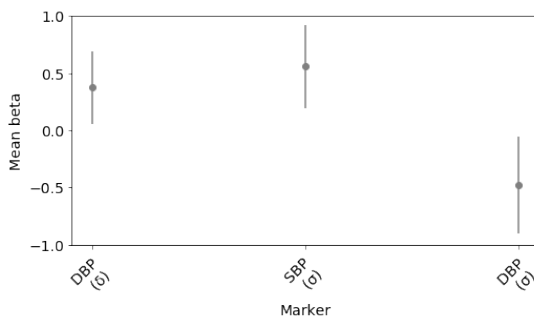
(b) Probability density functions (PDFs) of the training set.



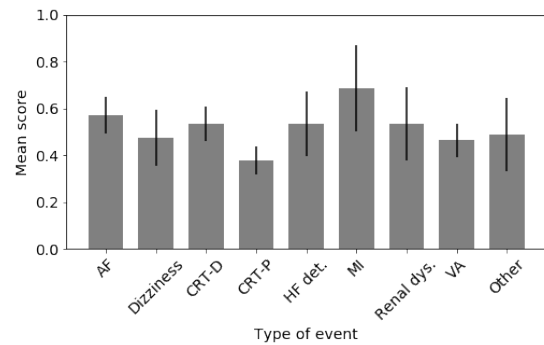
(c) Receiver operating curve (ROC) on the test set.



(d) Probability density functions (PDFs) of the test set.



(e) Mean and standard deviation of coefficients learnt by the classifier.



(f) Mean and standard deviation of classifier scores assigned to adverse events.

Figure 4.12: Performance of the combined classifier. AUC = Area under the receiver operating curve. DBP = Diastolic blood pressure. SBP = Systolic blood pressure. AF = Atrial fibrillation. HF det. = Heart failure deterioration. MI = Myocardial infarction. Renal dys. = Renal dysfunction. VA = Ventricular arrhythmia. The ROC figures show the mean ROC curve as a solid grey line and the mean $\pm$ the standard deviation of the ROC curves as the grey shaded area.

Using patient-specific (ψsbp , ψdbp and ϕdbp) markers as opposed to the unscaled mean and standard deviation markers did not improve the performance of the classifier on the test set.

For comparison, results for the guideline-based method using established thresholds of weight, blood pressure and NYHA values as its inputs are shown in Figure 4.13. The method achieved an average AUC of 0.51 on the test set data across 100 random partitions of training and test sets, suggesting no ability to identify days followed by HF-related adverse events (i.e. no better than random chance).

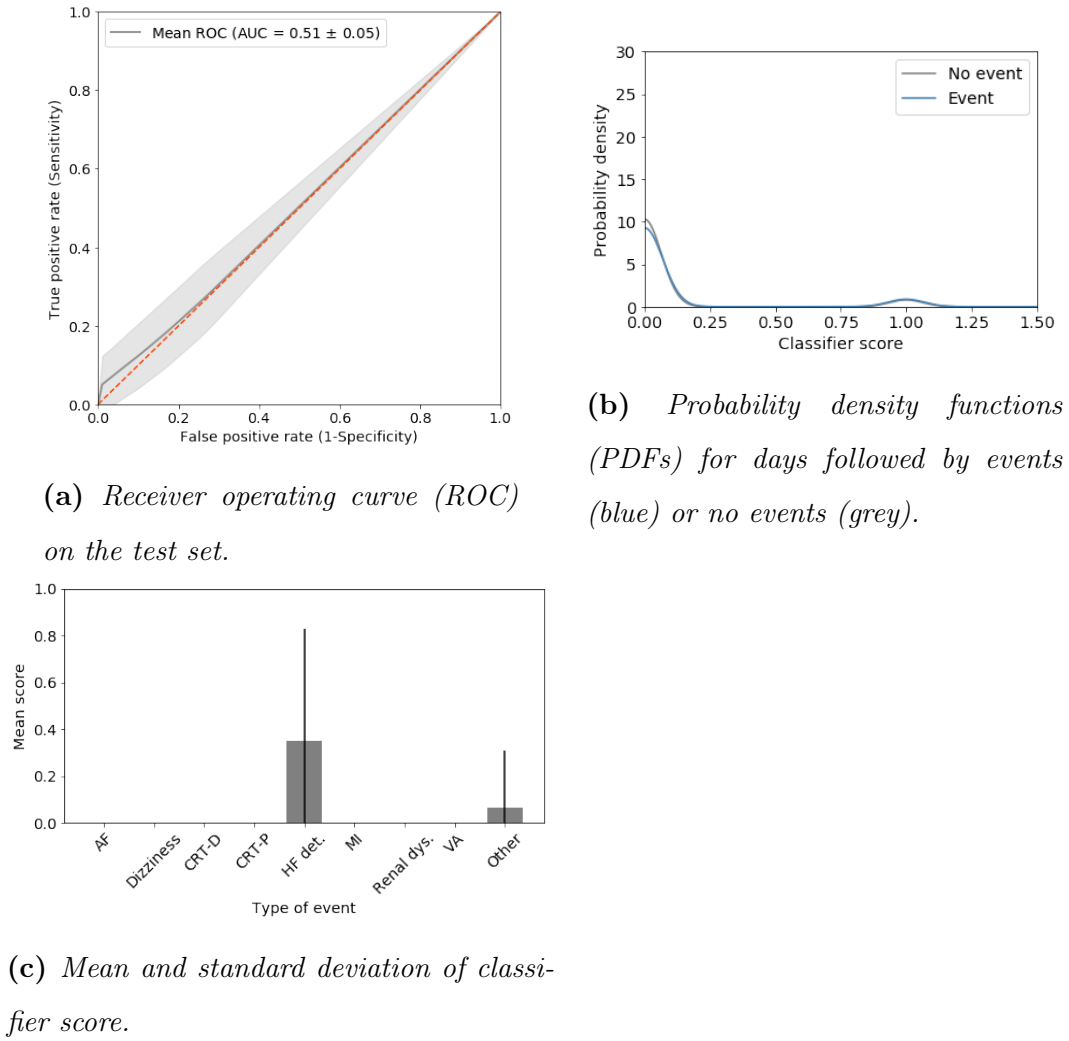
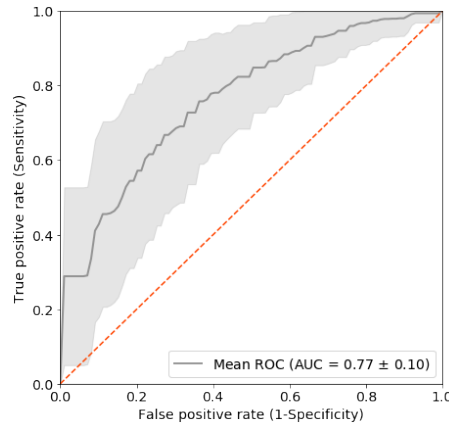


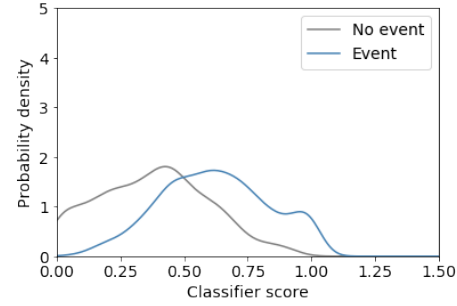
Figure 4.13: Summary of performance for the guideline-based model. AUC = Area under the receiver operating curve. AF = Atrial fibrillation. HF det. = Heart failure deterioration. MI = Myocardial infarction. Renal dys. = Renal dysfunction. VA = Ventricular arrhythmia. The ROC figure shows the mean ROC curve as a solid grey line and the mean $\pm$ the standard deviation of the ROC curves as the grey shaded area.

In order to study the performance of the optimal three-day change model (using the σ_{sbp} , σ_{dbp} and δ_{dbp} markers) and guideline-based models on the known types of HF-related adverse events, the analyses were repeated removing all events classed as 'other' from the dataset, leaving a total of 15 adverse events. Using the 70:30 partitioning into training and test sets, this meant there were on average ten adverse events in the training and five adverse events in the test set. The results for the guideline-based methods and

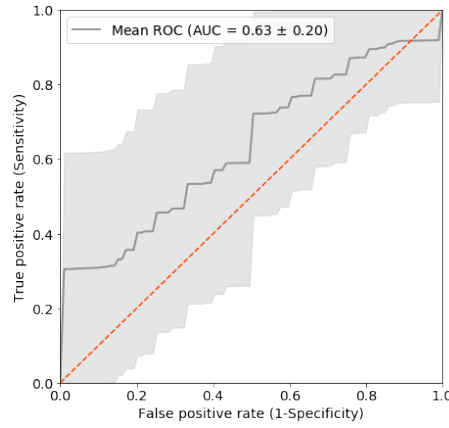
the best-performing logistic classifiers using three-day input vectors are shown in Figure 4.14. The mean AUC on the test set for the classifier-based methods improved, however at the cost of spread of values across the 100 iterations (see Figure 4.14c). The removal of half the event days makes the models' performance more susceptible to which subgroup of participants were included in the training and test sets and whether these participants have had adverse events. The mean AUC achieved using the logistic classifier was 0.63 and again the guideline-based thresholding model, which achieved an AUC of 0.51, did no better than random classifications.



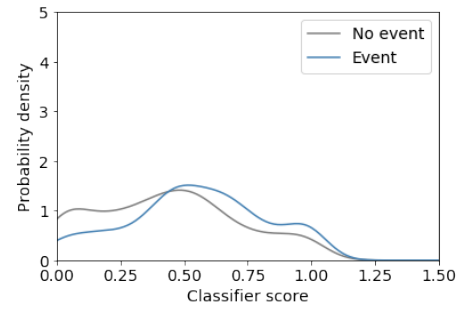
(a) ROC of logistic classifier with three-day input variables σsbp , σdbp and δdbp for the training set.



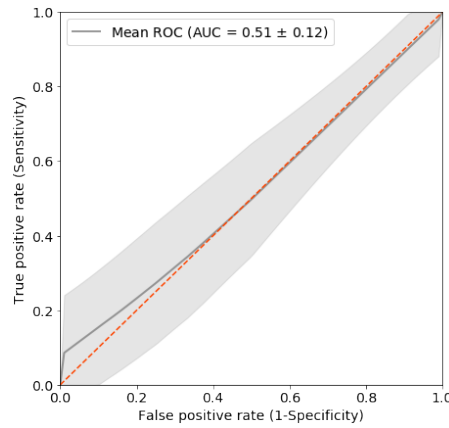
(b) PDFs of logistic classifier with three-day input variables σsbp , σdbp and δdbp for the training set.



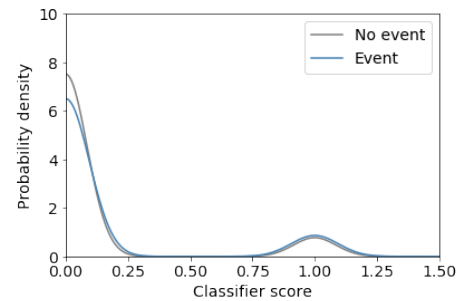
(c) ROC of logistic classifier with three-day input variables σsbp , σdbp and δdbp for the test set.



(d) PDFs of logistic classifier with three-day input variables σsbp , σdbp and δdbp for the test set.



(e) ROC of guideline-based model.



(f) PDFs of guideline-based model.

Figure 4.14: Receiver operating curves (ROC) and probability density functions (PDFs) for the methods excluding adverse event category 'other'. The ROC figures show the mean ROC curve and the mean $\pm$ the standard deviation of the ROC curves (grey shaded area).

4.5 Discussion

During the SHF2 trial, a total of 38 out of 202 patients experienced HF-related adverse events. 30 adverse events were preceded by sufficient data to be included in the prediction analysis. The variable of interest for this analysis was chosen to be the three-day change in the different physiological measurements and NYHA values collected as part of the SHF2 study. A combination of markers based on the average and standard deviation of three-day changes in measurements achieved the best performance in predicting HF-related adverse events one day before their occurrence, outperforming the model based on guideline-based thresholds, which did no better than a random prediction. Including patient-specific input markers did not improve the performance of the logistic classifier. The results suggest that meaningful information on the occurrence of HF-related adverse events is to be found in the variability of three-day changes in systolic and diastolic blood pressure as well as in the average three-day changes of diastolic blood pressure.

Major challenges faced during the analysis were the size of the dataset and the low number of adverse events. Despite the fact that the SHF2 study collected data from 202 patients and monitored them for a period of approximately six months, a total of only 44 HF-related adverse events were recorded, and only a subset of these were linked with sufficient self-monitored data to be included in the analysis.

Furthermore, nearly half of all registered HF-related adverse events were classified as '*other*'. This category ranged from elective check-up events to ensure the correct working of pacemakers to sudden death. The latter can be caused by a range of factors, including myocardial infarction or arrhythmias, and sudden death was therefore assigned to the group '*other*'. An attempt to remove the category '*other*' from the analysis improved the performance of the method, however at the cost of increasing variability in performance.

The large difference in the number of days with and without HF-related adverse events required the dataset to be balanced prior to training. As part of the balancing process each day with an adverse event was matched with a randomly selected day without adverse event. Whilst balancing results in a better discrimination between the two classes, the performance of the method is however highly dependent on the choice of randomly selected non-event days and how representative they are of non-event days in general.

It should also be noted that a median 10% of recorded days were missing. Missing values may occur at random but may also be correlated with disease deterioration. Future studies should consider including the occurrence of missing values as an input marker to a risk-prediction model and also study the time from missed measurements to the occurrence of adverse events. For the purpose of the analysis described in this chapter it was decided that including additional markers in the model would be inappropriate with such a small number of adverse events. This is however a shortcoming of the analysis. In addition to the non-randomness of missing values, another factor to consider is the impact of patient compliance on the performance of tele-monitoring systems.

4.6 Conclusion

The overall moderate performance achieved using the models described in this chapter is in agreement with previous findings on the included HF measurements [26][27]. Whilst there is some valuable information contained within the monitored measurements, this does not appear to be sufficient to provide an accurate tool for repeated-measure risk-prediction. Advanced risk markers, such as the standard deviation of three-day markers described in this chapter or the repeated down-sampling markers proposed by Koulaouzidis *et al.* should be further evaluated in future studies with far greater numbers of adverse events, however it is likely that additional markers will be required.

Automated Sleep-boundary Detection using Wrist-worn Accelerometers

5.1 Introduction

The previous chapter described and evaluated methods to predict HF deterioration using a combination of traditional repeated-measures markers, including weight, blood pressure and symptom scores. However, these methods only achieved moderate accuracy in adverse event prediction and as a result alternative approaches are needed. One such approach involves the inclusion of additional informative markers to enhance the performance of predictive algorithms.

Accelerometer-derived markers have previously been shown to independently predict long-term hospitalisations and deaths in HF patients [111][37]. Low accelerometer-derived PA levels in HF patients have been associated with long-term risks of hospitalisation and death [37]. As well as PA data extracted from accelerometers, information on a patient's duration and quality (e.g. fragmentation) of sleep may provide meaningful insights into their disease status [112]. Nearly 50% [113] of HF patients are believed to be affected by sleep disordered breathing and 33% of HF patients are believed to suffer from insomnias [114]. Such sleep conditions have shown to directly impact the long-

term prognosis of patients with HF [112]. In addition to this, reversal of sleep apnea has shown to have a positive impact on other physiological HF markers, such as blood pressure and systolic contraction [59].

Despite the potential benefits of accelerometer-derived PA and sleep markers, research on the ability of such markers to predict short-term HF deterioration remains limited [47][59][48]. The current chapter will lay the foundations for this analysis by addressing a major technical limitation of accelerometer analysis: the need for self-reported sleep-onset and wake-up information.

The principal contributions of this chapter are:

- The development of (a) a PDF-based, (b) a random-forest classifier and (c) a change-point detection method for the automatic detection of sleep-boundary times.
- The evaluation of all algorithms against self-reported sleep-boundary data using mean absolute differences in sleep-boundary times, kappa statistics and Bland-Altman plots.

5.2 Background

5.2.1 Accelerometer-derived Sleep Analysis

Polysomnography (PSG) is the gold standard for sleep analysis [115]. During a PSG assessment, an individual's neural, ocular and cardiac activity as well as other body functions are monitored continuously overnight. The analysis is typically conducted in a sleep laboratory. PSG is a costly and time-intensive analysis that is not suitable for large population studies [116]. As a result, other, more appropriate means for sleep analysis have been considered.

Wrist-worn accelerometers provide a suitable tool for sleep detection, where long-term, large-scale or affordable recordings of sleeping behaviour are required [116]. Existing algorithms are typically based on the identification of periods of low movement measured as low *count* values or small standard deviations of acceleration across one or multiple axes [66][117]. *Counts* are hereby defined in an often proprietary way as aggregating acceleration information within specified time intervals, also known as *epochs*. Once sleep and wake periods have been identified, descriptive markers such as total sleep duration, sleep latency (defined as the time between going to bed and falling asleep), number of sleep interruptions (fragmentation) and sleep efficiency (defined as the percentage of time in bed (t_{bed}) spent asleep (t_{asleep})) (Equation 5.1) can be derived [66].

$$Sleep_efficiency = \frac{t_{asleep}}{t_{bed}} \quad (5.1)$$

These markers may be meaningful markers for HF risk prediction. Table 5.1 lists commonly cited sleep-detection algorithms and provides information on their performance in comparison to PSG.

Cole *et al.* [117] developed a technique for the Motionlogger Actigraph (Ambulatory Monitoring, Inc, New York, U.S.A.), a wrist-worn accelerometer set to collect uni-axial accelerometer data at an epoch length of 2s. The authors collapsed the two-second epochs into 60s epochs (by taking the maximum ten seconds of acceleration within each epoch) and aggregated the weighted (W) sums of consecutive 60s epoch counts (E) over a seven-minute period (four minutes before to two minutes after the epoch of interest). The final sum is scaled by a scaling factor P as shown in Equation 5.2.

$$D_n = P(E_{n-4}W_{-4} + E_{n-3}W_{-3} + E_{n-2}W_{-2} + E_{n-1}W_{-1} + E_nW_0 + E_{n+1}W_{+1} + E_{n+2}W_{+2}). \quad (5.2)$$

If D_n is smaller than 1.0, the subject is classed as being asleep, otherwise the subject is classed as awake. According to the authors, typical weightings and scaling factors are: $P = 0.0001$, $W_{-4} = 404$, $W_{-3} = 598$, $W_{-2} = 326$, $W_{-1} = 441$, $W_0 = 1408$, $W_{+1} = 508$, $W_{+2} = 350$. These may however differ depending on accelerometers and patient groups.

Author	Input (epoch length)	Type of algorithm	Dataset	Accuracy
Cole	Proprietary count (60s)	Weighted sum	21 adults	88%
Kushida	Proprietary count (30s)	Weighted sum	100 subjects with sleep disorders	77%
Sadeh	Proprietary count (60s)	Weighted sum	26 adults	94%
Borazio	Raw data (1s)	Rule-based	42 sleep lab patients	78%
Van Hees	Raw data (5s)	Rule-based	21 adults	83%

Table 5.1: Overview of sleep-detection algorithms commonly referred to in the literature. Accuracy refers to the mean percentage of correctly identified sleep and wake labels across participants as shown in Equation 5.3. An exception is the accuracy reported by Borazio *et al.*, which is referring to the median across participants. It is worth noting that algorithms were developed and evaluated using different epoch lengths, which may affect the accuracies reported.

In order to further improve the performance of their algorithm, Cole *et al.* proposed the use of a set of rules previously developed by Webster *et al.* [118]:

- After at least 4 minutes scored as wake, the next 1 minute scored as sleep is rescored wake.
- After at least 10 minutes scored as wake, the next 3 minutes scored as sleep are rescored wake.
- After at least 15 minutes scored as wake, the next 4 minutes scored as sleep are rescored wake.
- 6 minutes or less scored as sleep surrounded by at least 10 minutes (before and after) scored as wake are rescored wake.
- 10 minutes or less scored as sleep surrounded by at least 20 minutes (before and after) scored as wake are rescored wake.

Cole *et al.* [117] validated their algorithm on a group of 21 subjects against PSG. They found that their algorithm achieved an overall accuracy of 88% with a tendency to overestimate sleep time. Accuracy was defined as the number of correctly detected 'sleep' (*true positives* (tp)) and 'wake' (*true negatives* (tn)) epochs over the total number of epochs (N_{epoch}).

$$Accuracy = \frac{tp + tn}{N_{epoch}}. \quad (5.3)$$

Kushida *et al.* [119], based on the work of Oakley *et al.* [120], suggested a similar approach using a two-minute time frame before and after the epoch of interest (Equation 5.4). Kushida *et al.*'s [119] algorithm was developed for the AW4 (Mini Mitter Co., Inc., Oregon, U.S.A.), a wrist-worn accelerometer with a sampling frequency of 32Hz, set to aggregate data into 30-second epochs. The authors applied a nine-epoch moving average filter to the device's proprietary count values as shown in Equation 5.4.

$$D_n = 0.04 * E_{n-4} + 0.04 * E_{n-3} + 0.2 * E_{n-2} + 0.2 * E_{n-1} + 2 * E_n + 0.2 * E_{n+1} + 0.2 * E_{n+2} + 0.04 * E_{n+3} + 0.04 * E_{n+4}. \quad (5.4)$$

Kushida *et al.* proposed three wake thresholds, high (80), medium (40) and low (20) in order to adjust for sleep disorders that lead to increased levels of movement during sleep; if the total activity count D_n exceeds the chosen limit, the time of interest is classed as wake time, otherwise it is classed as sleep. The authors tested their sleep-scoring algorithm on 100 subjects with sleep disorders against PSG and found that it achieved an accuracy of 77% with a sensitivity (defined as detecting an epoch labelled as 'sleep' by the PSG as 'sleep' using the accelerometer data) of 92% and a specificity (defined as detecting an epoch labelled as 'wake' by the PSG as 'wake' using the accelerometer data) of 48%.

Sadeh *et al.* [121] developed another proprietary-count-based algorithm for the wrist-worn AMA-34 (Ambulatory Monitoring, Inc., New York, U.S.A.) device at an epoch length of 1 minute, but included additional information on the activity around the time

of interest into their scoring calculation . The algorithm is shown in Equation 5.5:

$$PS_n = 7.601 - 0.065 * MW5_n - 1.08 * NAT_n - 0.056 * SD6_n - 0.072 * \ln(ACT_n) \quad (5.5)$$

$MW5$ is the average number of proprietary counts during the window five minutes before and after the scored minute, NAT is the number of minutes with proprietary activity counts between 50 and 100 in a window of eleven minutes around the scored minute, $SD6$ is the standard deviation in the scored minute and five minutes before and ACT is the number of activity counts in the scored minute plus one. If the probability of sleep (PS) is greater than 0, the minute is scored as sleep. Sadeh *et al.* [121] validated their algorithm on 26 subjects against PSG and found that it achieved an accuracy of >94%.

Cole's [117], Kushida's [119] and Sadeh's [121] sleep-scoring algorithms make use of accelerometer counts and were created for specific accelerometer devices. Different devices use different pre-processing algorithms (e.g. cleaning and filtering) and aggregate markers in specific ways (e.g. summed over epoch versus maximum value during epoch, etc.). Cross-calculations based on sleep algorithms developed for specific devices have been proposed, but require the implementation of individual comparison studies [65].

Borazio *et al.* [116] developed a more general approach to sleep detection based on the standard deviations of acceleration g along the z -axis of the accelerometer. In their approach, the authors calculate the standard deviation across 1s-intervals (Equation 5.6).

$$S_n = \begin{cases} wake, & \text{if } \sqrt{\frac{1}{99} \sum_{i=1}^{100} (z_i - \bar{z})^2} > \delta \\ sleep, & \text{otherwise} \end{cases} \quad (5.6)$$

S_n is the sleep state across each 1s-interval n , z_i is the acceleration along the z -axis measured at ms i and $\bar{z}$ is the mean acceleration along the z -axis during the 1s-interval. If the standard deviation has not exceeded threshold δ for a pre-defined duration, the interval is classified as 'sleep'. Borazio *et al.* [116] tested their sleep-scoring algorithm on 41 subjects against PSG. Their algorithm achieved a median accuracy of 78%. It is worth

noting that the authors chose to report the median rather than the mean performance across participants, in contrast to other authors, without further explanation.

Van Hees *et al.* [66] developed another general sleep-scoring approach based on angular arm movement. Their algorithm requires the change in median arm angle (Equation 5.7) between five-second windows to be below five degrees for a period of at least five minutes in order for a time-period to be classified as ‘sleep’.

$$angle = \left(\arctan \frac{a_z}{\sqrt{a_x^2 + a_y^2}} \right) * \frac{180}{\pi}, \quad (5.7)$$

where a_x , a_y , a_z are median values of acceleration (g) across a five-second time window.

The authors validated their algorithm against PSG on 28 individuals and found that it achieved an accuracy (defined in Equation 5.3) of 83% with a sensitivity of 91% and a specificity of 45%.

All of the listed sleep-detection algorithms were developed in a laboratory environment and rely on self-reported sleep-boundary information. This information is typically collected through means of a sleep diary [66] as shown in Figure 5.1. Subjects are asked to record lights off, asleep (sleep-onset) and wake-up times for the duration of the accelerometer recording. In large-scale studies, such as the UK Biobank, it may however not be feasible to collect self-reported sleep-boundary times for all participants and therefore a need exists for an automated way to extract sleep-onset and wake-up times.

In this chapter we therefore set out to develop a fully automated method for sleep-boundary detection from wrist-worn accelerometer data.

5.3 Methodology

Three approaches for automatic sleep-boundary detection were employed: (i) a data-driven-threshold approach based on the intersection of the PDFs of sleep and wake data,

What time did you: *Please use the 24-hour clock*

- Turn off the light last night (ready to go to sleep)? __ : __
- First fall asleep last night? __ : __
- Wake up today (eyes open, ready to get up)? __ : __

Did you wake up during the night? ☐ Yes ☐ No

If yes, for how long were you awake in total? __ hours __ minutes

Did you have a nap today? ☐ Yes ☐ No

If yes, please detail any naps:

	1 st Time	2 nd Time	3 rd Time	4 th Time
Time you fell asleep	__ : __	__ : __	__ : __	__ : __
Time you woke up	__ : __	__ : __	__ : __	__ : __

Figure 5.1: *Sleep diary used to collect sleep-onset and wake-up information.*

(ii) a machine learning technique using a random forest (RF) classifier trained on a variety of input features and (iii) a statistical technique for detecting change points in the accelerometer time series. All methods are illustrated in the flow diagram shown in Figure 5.2. Whilst the PDF and RF methods detect minute-wise labels that are then smoothed to provide sleep-boundary times, the change-point detection (CPD) method identified entire periods of high or low movement and is filtered only to remove short periods of activity during sleep and vice versa.

5.3.1 Database

Accelerometer recordings and self-reported sleep-onset and wake-up information from the CHA Exercise study were used for this analysis. The purpose of the CHA Exercise study was to determine the effect of an anaerobic exercise intervention on MRI brain measures. Inactive adults between the ages of 60 and 80 years were eligible to take part in the study. This population is similar in age to typical HF populations. After providing informed consent, participants were asked to wear the GENEActiv accelerometer on their non-dominant wrist for a period of nine days on three occasions during the study

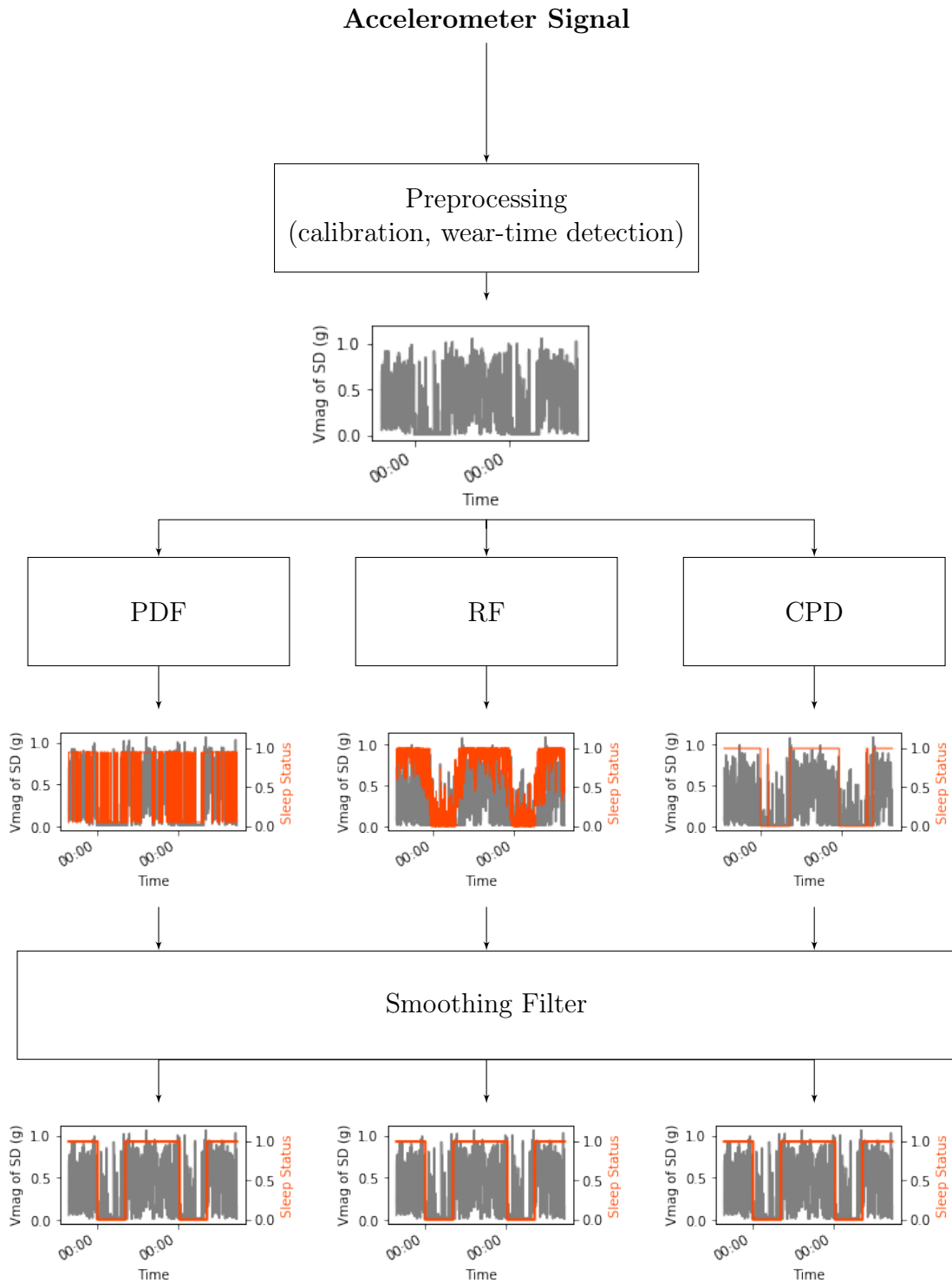


Figure 5.2: Flow diagram providing an overview of the three automated sleep-boundary-detection methods described in this chapter. PDF = Probability density function, and refers to the threshold-based approach based on the intersection of accelerometer PDFs during sleep and wake times. RF = Random forest. CPD = Change-point detection.

(weeks 0, 12 and 24). The GENEActiv is a triaxial accelerometer that has previously been validated for sleep research [66]. Acceleration data were recorded at a sampling frequency of 87.5Hz. Participants also recorded daily lights-off, asleep and wake-up times using a sleep-diary for the duration of the recordings.

5.3.2 Data Preparation

5.3.2.1 Calibration

Accelerometers contain electromechanical transducers that convert mechanical force into electrical signals [122]. In order to ensure that accelerometers exhibit a standardised response to acceleration, individual devices are calibrated prior to utilisation. Where this is not possible, offline calibration algorithms can adjust *offset* and *gain* values post recording. For the purpose of this analysis, triaxial accelerometer traces were calibrated using the offline approach proposed by van Hees *et al.* [123]. This method is based on the principle that at rest, the vector magnitude of the force measured by an accelerometer should be that of the gravitational field of the earth, 1g.

Axis-wise means ($\bar{x}$, $\bar{y}$, $\bar{z}$) and standard deviations of consecutive ten-second windows of the accelerometer signal were extracted. Periods of *low movement* (likely to include data at rest) were identified automatically, defining *low movement* as 10s periods with standard deviations of $<0.013g$ along all three axes. The value of $0.013g$ is a heuristically chosen threshold based on noise measurements of wrist-worn accelerometers at rest (see [123]).

The mean acceleration along each axes for the selected 10s periods was scaled by the vector magnitude of the signal across all axes. This process provided a new set of *ideal* mean values ($\bar{x}_{\text{scaled}}$, $\bar{y}_{\text{scaled}}$, $\bar{z}_{\text{scaled}}$) that fulfill the condition of creating a vector magnitude of 1g, and therefore the expected acceleration experienced by the

device at rest (Equations 5.8- 5.13).

$$\bar{x}_{scaled_n} = \frac{\bar{x}_n}{\sqrt{\bar{x}_n^2 + \bar{y}_n^2 + \bar{z}_n^2}} \quad (5.8)$$

$$\bar{y}_{scaled_n} = \frac{\bar{y}_n}{\sqrt{\bar{x}_n^2 + \bar{y}_n^2 + \bar{z}_n^2}} \quad (5.9)$$

$$\bar{z}_{scaled_n} = \frac{\bar{z}_n}{\sqrt{\bar{x}_n^2 + \bar{y}_n^2 + \bar{z}_n^2}} \quad (5.10)$$

Finally, a linear regression was performed in order to find the *gain* and *offset* values associated with the best fit between the original and scaled accelerometer values along each axis. This process was repeated until the error between the *gain* and *offset*-adjusted values ($\bar{x}_{new}$, $\bar{y}_{new}$, $\bar{z}_{new}$) (Equations 5.11-5.13) and the scaled values was below a set threshold error.

$$\bar{x}_{new_n} = gain_x * \bar{x}_n + offset_x \quad (5.11)$$

$$\bar{y}_{new_n} = gain_y * \bar{y}_n + offset_y \quad (5.12)$$

$$\bar{z}_{new_n} = gain_z * \bar{z}_n + offset_z \quad (5.13)$$

5.3.2.2 Wear-time Detection

Non-wear periods at the beginning, end or during the accelerometer recording may result in a misinterpretation of the data. In order to remove these periods prior to the analysis, periods of wear and non-wear time were detected using the approach suggested by van Hees *et al.* [124] [125]. 60-minute periods were extracted from the calibrated accelerometer data, with each 60-minute period P_A overlapping the previous period by 15 minutes. Any period (m) in which the standard deviation (Equation 5.14) was below 0.003g for at least two of the three axes or the value range (Equation 5.15) was below 0.05g for at least two of the three axes was classified as a non-wear period. 0.003g was chosen based on the experimentally-extracted standard deviation due to noise during non-wear times [123]. The longest continuous wear time was selected for further analysis.

$$x_std_m = \frac{1}{(P_A * 60 * f_sample) - 1} \sum_{n=i*60*f_sample}^{(i+60)*60*f_sample} (x_n - \bar{x}), \quad (5.14)$$

for $i = 0, 15, 30, \dots$, f_sample = accelerometer sampling frequency and $\bar{x}$ is the mean acceleration during the corresponding 60-minute period.

$$x_range_m = \max([x_{i*60*100} : x_{(i+60)*60*100}]) - \min([x_{i*60*100} : x_{(i+60)*60*100}]) \quad (5.15)$$

5.3.3 Data Analysis

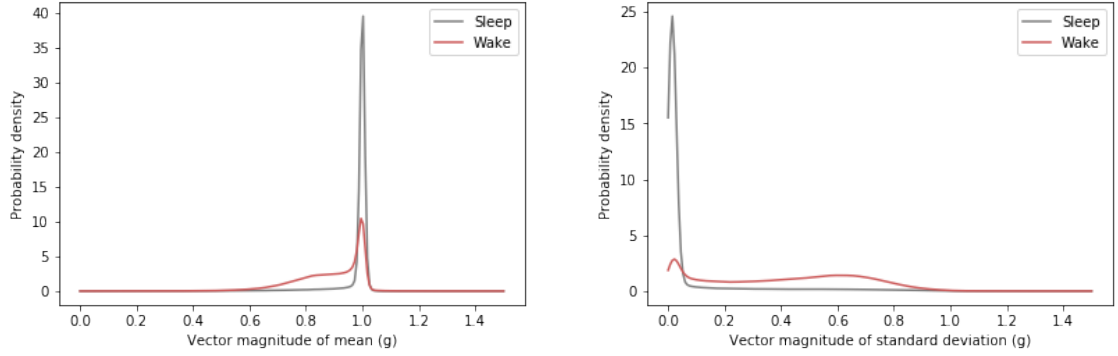
Three approaches for sleep-boundary detection were studied: a PDF-, a random forest and a change-point-detection-based method. Self-reported sleep-onset and wake-up times were used as the ground-truth for the development of the algorithms. All methods were trained on data from a random subset of 70% of participants and tested on data from the remaining 30% of participants. This process was repeated 30 times with different random partitions of training and test sets in order to obtain a robust estimate of the performance of the algorithms, with each of the participants likely to have been within the testing set ten times. All three methods used accelerometer-derived features with an epoch length of one minute as their input, corresponding to the resolution at which self-reported sleep times were reported.

In order to identify a suitable input variable for the threshold and change-point-detection method, the degree to which the accelerometer signals for wake and sleep times differed was studied using PDFs. PDFs represent the likelihood of a random variable taking on any specific value in the variable-space. They can be extracted by applying a smoothing kernel over individual values of the variable and aggregating the kernels into one function. For the PDFs of sleep and wake times, a Gaussian kernel with a bandwidth (bw) (Equation 3.2) set by Scott's rule of thumb ([71]) was used (see

Section 3.2 for more detail). Figures 5.3a and 5.3b show the PDFs of the $vmag_{mean}$ (Equation 5.16) and the $vmag_{std}$ (Equation 5.17) variables during sleep and wake times.

$$vmag_mean_m = \sqrt{x_mean_m^2 + y_mean_m^2 + z_mean_m^2}. \quad (5.16)$$

$$vmag_std_m = \sqrt{x_std_m^2 + y_std_m^2 + z_std_m^2}. \quad (5.17)$$



(a) PDFs of the $vmag_{mean}$ during sleep (solid grey line) and wake (solid red line) times. (b) PDFs of the $vmag_{std}$ during sleep (solid grey line) and wake (solid red line) times.

Figure 5.3: Probability density functions (PDFs) for the vector magnitude of the mean ($vmag_{mean}$) and the vector magnitude of the standard deviation ($vmag_{std}$).

It can be seen that differences between wake and sleep times are more pronounced for $vmag_{std}$ than they are for $vmag_{mean}$. A mathematical way to quantify the difference between two PDFs is the Kullback-Leibner divergence (D_{KL}) (see Section 3.2). The D_{KL} of the distributions for $vmag_{mean}$ and $vmag_{std}$ were 0.80 and 1.35 respectively, confirming that differences in $vmag_{std}$ during wake and sleep times are more pronounced than in $vmag_{mean}$. The threshold and change-point detection based methods were therefore developed using the vector magnitude of standard deviation ($vmag_{std}$) as their input. $vmag_{std}$ for a typical subject from the CHA Exercise study are shown in Figure 5.4.

The input to the random-forest classifier was an array of 41 time- and frequency-domain features, containing single-axis features computed using the vector magnitude of acceleration as well as cross-axial features (Table 5.2) extracted from one-minute epochs.

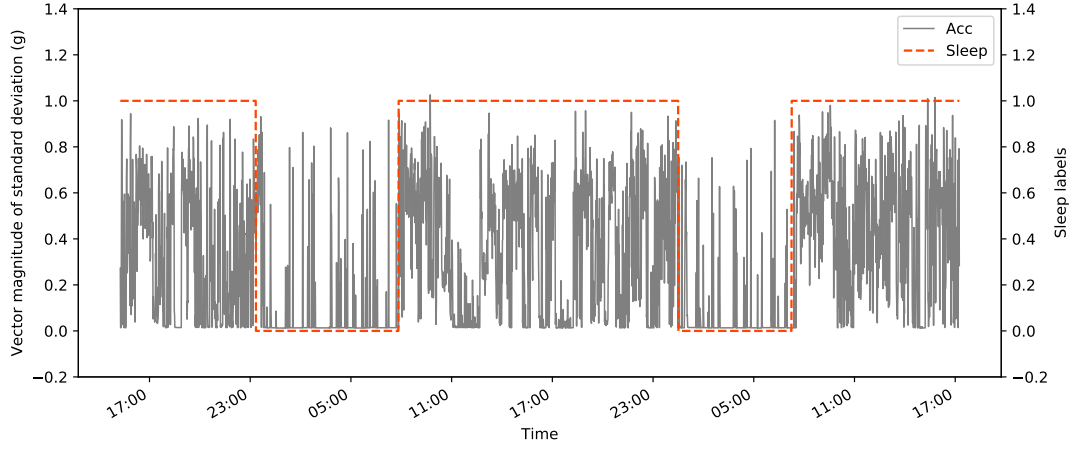


Figure 5.4: $vmag_{std}$ (grey solid line) and self-reported (orange dashed line) sleep labels for a typical subject from the CHA Exercise study (for 48 hours' worth of data).

Features were chosen based on a previous selection by Ellis *et al.* [126], who aggregated features commonly used for activity analysis and found they performed well in classifying different PA types.

5.3.3.1 Sleep Detection using a Data-driven Threshold

Amplitudes of $vmag_{std}$ are on average larger during wake times than during sleep times (see Figure 5.4). As a result, it may be possible to differentiate between the two states simply by using an appropriate threshold. The first method developed for automatic sleep-boundary detection was based on a data-driven-threshold to assign minute-by-minute sleep labels coupled with a smoothing filter.

The optimal threshold to differentiate between sleep and wake times was found using probability-adjusted crossing-points of PDFs during wake and sleep times. An example set of PDFs for wake and sleep times is shown in Figure 5.5a. In order to find the optimal threshold ($\tau_{i,j}$) for each recording in the training set, the PDFs of sleep and wake times were adjusted by their relative probabilities-of-state. Assuming an average amount of eight hours sleep a night for 24 hours worth of data (see Figure 5.4), a wake-to-sleep

Single-axis Features	Multi-axes Features
Mean	Pearson correlation
Standard deviation	Covariance
Coefficient of variation	
Skewness	Mean roll
Kurtosis	Mean pitch
Median	Mean yaw
Interquartile range	Standard deviation of roll
Minimum	Standard deviation of pitch
Maximum	Standard deviation of yaw
Power contained in 1-15 Hz band	
Dominant frequency	
Power in dominant frequency	
Entropy	
Autocorrelation	

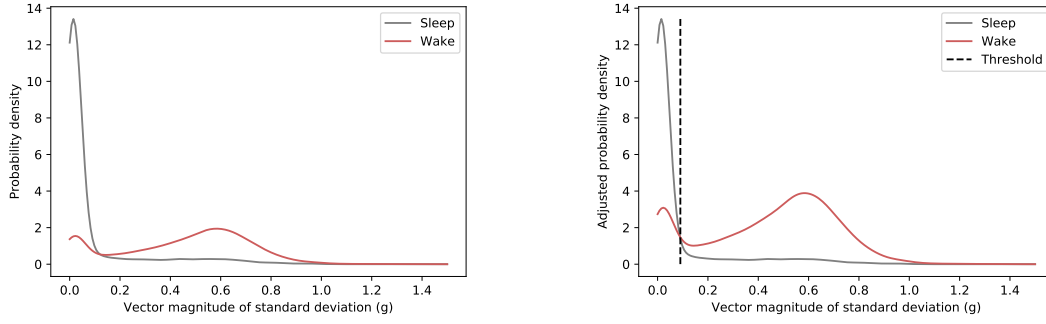
Table 5.2: Overview of features used for random forest classification. Features were chosen based on a previous selection by Ellis et al.[126].

ratio of 2:1 was used. The intersection between the adjusted wake and sleep PDFs were found and the corresponding $vmag_{std}$ values provided $\tau_{i,j}$ (Figure 5.5b). The PDF-based thresholds were averaged across the recordings (j) for each participant and consecutively across all participants (i) in the training set in order to find the final threshold τ_{PDF} (see Equation 5.19). For the PDF-based sleep classification, all values of $vmag_{std}$ above τ_{PDF} were classified as wake time (1) and all others as sleep time (0) (Equation 5.19).

$$\tau_{PDF} = \frac{1}{I} \sum_{i=0}^I \frac{1}{J} \sum_{j=0}^J \tau_{i,j}, \quad (5.18)$$

where I is the total number of participants in the training set and J is the total number of recordings for participants i .

$$state_m = \begin{cases} 1, & \text{if } vmag_std_m \geq \tau_{PDF} \\ 0, & \text{otherwise} \end{cases} \quad (5.19)$$



(a) PDFs of the $vmag_{std}$ at sleep (solid grey line) and wake (solid red line) times.

(b) Probability-adjusted PDFs of the $vmag_{std}$ at sleep and wake times and the $vmag_{std}$ at which the two PDFs intersect (black dashed line).

Figure 5.5: Extraction of PDF-based thresholds for sleep detection. PDF = Probability Density Function.

5.3.3.2 Sleep Detection using a Random Forest Classifier

The previously described PDF approach only used a single input variable; however, additional information may be contained in other time- or frequency-domain variables. A random forest based method was developed using a range of input features to study their combined ability to correctly identify sleep and wake times. The Sklearn Ensemble toolbox [75] in Python was used for this analysis.

Random forest classifiers are ensemble classifiers that combine multiple, individually weak trees into one stronger classifier. The advantage of a random forest classifier over other classifiers is that it creates multiple randomly selected subsets of the training data and therefore reduces the impact of hierarchical structures within the dataset. Hierarchical structures are likely to be found in time-series data such as accelerometer data, where within subject readings may be correlated over time.

For each participant self-reported sleep and wake times were extracted. The minute-wise sleep and wake labels provided the ground truth for the learning-process of the

classifier. The feature vector for each labelled minute contained the features listed in Table 5.2.

Participants in the training set were randomly divided into a training (Tr_ml) and a testing (Te_ml) set using a 70:30 split. In order to avoid a classification bias towards wake classifications (as wake minutes occur more frequently during a 24-hour period), the training set Tr_ml was balanced to contain as many sleep as wake labels. For each participant, the difference between the total number of wake and sleep epochs was identified; a randomly selected subset of sleep epochs equal to that difference was then added to Tr_ml . The balanced Tr_ml dataset was normalised using the mean and standard deviation of each of the features contained within that dataset (Equation 5.20 shows this for $vmag_{std}$) and fed into the random forest classifier.

$$vmag_norm_n = \frac{vmag_n - \overline{vmag_std_train}}{std(vmag_std_train)} \quad (5.20)$$

To evaluate the performance of the classifier, the features in the Te_ml dataset were normalised using the mean and standard deviation of the features in the Tr_ml dataset and fed into the trained classifier. The area under the curve (AUC) of the receiver operating curve (ROC) was recorded as the performance measure. All classification steps are shown in Figure 5.6.

The number of trees used in, and branch depth of, a random forest classifier directly impact on its classification performance. Having a deeper branch structure can for example mean that more detailed relationships between variables are learned by the classifier. Too deep a structure on the other hand may cause overfitting on the training set resulting in poorer performance on the testing set. In order to find the optimal random forest structure for the sleep-wake classification, a range of tree numbers and minimum number of samples per branch (related to the branch depth) requirements were tested. The best combination of the two based on its AUC value was then used for

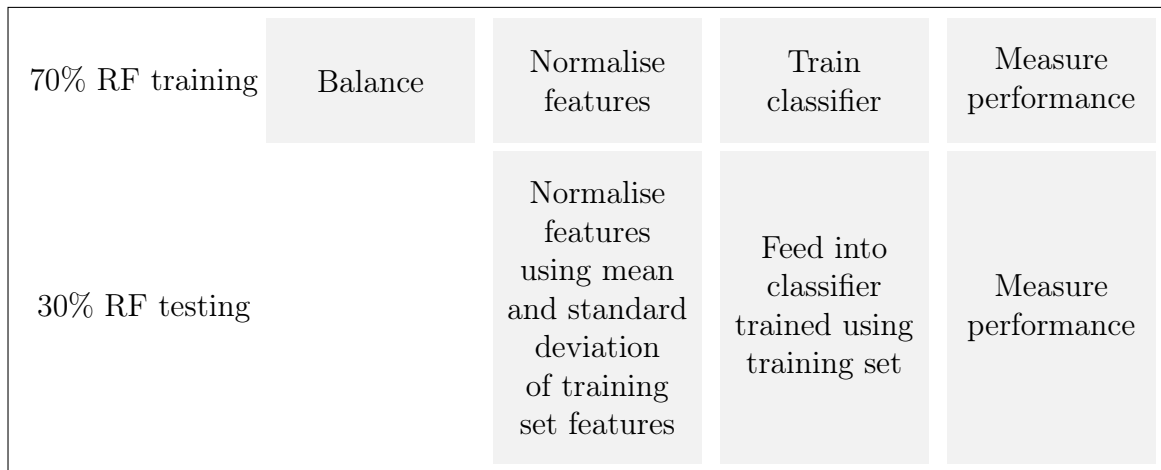


Figure 5.6: *Flow diagram summarising random forest training and testing steps.*

further analysis.

Including too many potentially correlated variables in the random forest classifier may lead to a combination of them being assigned to trees in an over proportional amount. This in turn can lead to an overestimation of the weight of these variables and therefore to suboptimal classifier performance. In order to evaluate whether this effect was taking place, the random forest classifier was run 30 times using all of the features and 30 times with only the five most important features. The ROC curves were consecutively plotted and compared, to study the variation in ROC performance with partitioning of the dataset.

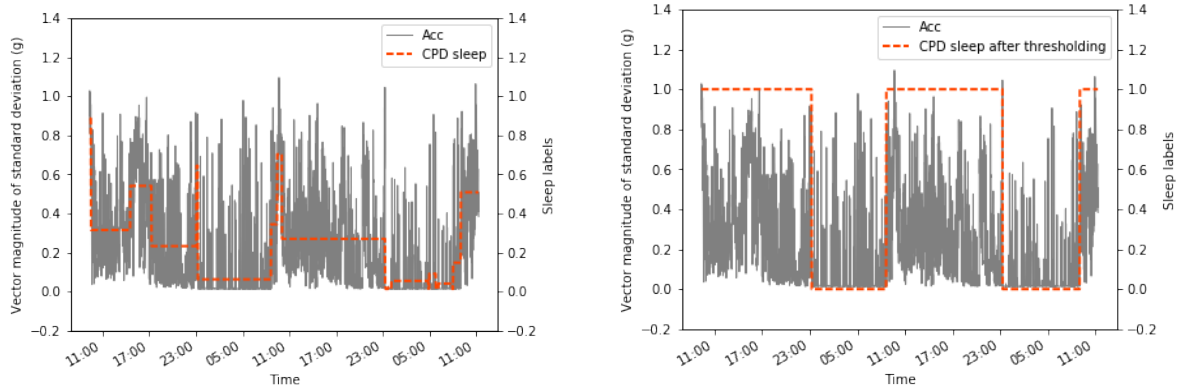
5.3.3.3 Sleep Detection using Change-point Detection

The PDF- and RF-based sleep detection approaches focused on characteristics of individual samples of the accelerometer signal ($vmag_{std}$), with no information on the context in which they occur. An alternative approach to identify sleep-onset and wake-up times may be to detect changes in the time-series signal. The final sleep-boundary detection method tested was based on change-point detection (CPD), which in contrast to the previously described methods, identified periods of sleep or wakefulness rather than minute-by-minute labels.

The CPD technique developed by Killick *et al.* [76], implemented in R package '*changepoint*' [77] was employed to identify changes in the mean or variance of $vmag_{std}$. This technique is based on maximum log-likelihood estimations (MLE), which provide a measure of the integrity/agreement within a signal. If the sum of the MLEs of the segments of a signal prior to and after a time n is larger than the MLE of the combined signal, n is marked as a change point within the signal (see 3.5 for more detail). In order to identify multiple change points within the signal, a binary-segmentation approach was employed, whereby the dominant change point was identified first and samples either side of the change point were consecutively scanned for additional change points. The method stops searching for change points after N_{max} points have been identified. For the purpose of this analysis, N_{max} was set to ten times the number of days in the recording in order to accommodate for change points due to sleep and wake stages as well as changes in activity during sleep or wakefulness.

Whilst differences in acceleration between sleep and wake stages are expected to create dominant change points, there may also be changes in accelerometer patterns throughout the day or night that result in false positives, e.g. a subject may show predominantly active behaviours in the morning and then become more sedentary in the afternoon (Figure 5.7a). In order to improve the performance of the change-point-detection based sleep classification method, a threshold (τ_{CPD}) was introduced after change-point detection. Whenever the mean $vmag_{std}$ of a section identified by the change-point-detection method increased above τ_{CPD} , the section was classified as wakefulness. Whenever the mean $vmag_{std}$ decreased below τ_{CPD} , the section was classified as sleep (Figure 5.7b).

In order to find the optimal value of τ_{CPD} a range of thresholds were applied to the accelerometer recordings in the training set. The limits of the threshold range were chosen based on the range of intersection points between $vmag_{std}$ values during wake-



(a) Visualisation of false positive change point. The figure shows the $vmag_{std}$ (grey line) and the mean $vmag_{std}$ values for the segments identified by the change-point detection (orange dashed line).

(b) $vmag_{std}$ (grey line) and sleep (0) and wake (1) classes (orange dashed line) after thresholding.

Figure 5.7: Change-point-detection method.

fulness and sleep found for the training set (as described for the PDF-based method). The threshold that resulted in the smallest mean absolute error (MAE) (Equation 5.21) between automatically-derived and self-reported sleep-onset and wake-up times on the training set was chosen for further evaluation on the testing set.

5.3.3.4 Post-processing using a Smoothing Filter

A rolling-average smoothing filter (mean-filter with equal weights and length N) was applied to the outputs of all three methods and its output was hardlimited ($y_h = 1.0$ if $y > 0.5$ and $y_h = 0.0$ if $y \leq 0.5$). Whilst the purpose of this filter was to turn minute-by-minute labels into sleep/ wake periods in the case of the PDF and RF methods, it was also applied to the CPD-derived sleep periods to remove short episodes of wake time during sleep and vice versa. After applying the hardlimiter, the start and end times of the M largest sleep periods were selected as the sleep-boundary times identified by the change-point detection method, where M was the number of days (defined as number of 4:00 am instances) contained within a recording.

The length of the filter, N (N_{PDF} , N_{RF} , N_{CPD}) was learned from the training set by minimising the mean absolute error (MAE) between self-reported and automatically-derived sleep-onset and wake-up times post filtering. The MAE was calculated as shown in Equation 5.21, where $tan_{i,j,k}$ is the self-reported and $tal_{i,j,k}$ is the algorithm-derived sleep-onset or wake-up time, K is the combined number of sleep-onset and wake-up times in a specific recording, J the number of recordings for a specific individual and I the number of individuals in the training set.

$$MAE = \frac{1}{I} \sum_{i=0}^I \frac{1}{J} \sum_{j=0}^J \frac{1}{K} \sum_{k=0}^K \sqrt{(tan_{i,j,k} - tal_{i,j,k})^2} \quad (5.21)$$

5.3.4 Evaluation

All sleep-detection algorithms were evaluated on the accelerometer signals from the participants in the testing set. The primary performance measure was the mean absolute error (MAE) between the algorithm-derived and self-reported sleep-onset and wake-up times. This was averaged over all individuals (Equation 5.21) and over all nights (Equation 5.22).

$$MAE_{night} = \frac{1}{N_{nights}} \sum_{i=0}^I \sum_{j=0}^J \sum_{k=0}^K \sqrt{(tan_{i,j,k} - tal_{i,j,k})^2}, \quad (5.22)$$

where N_{nights} is the total number of nights in the testing set.

As secondary performance markers, the accuracy, sensitivity, specificity and kappa statistic of the algorithms were calculated on a minute-by-minute level (see Section 3.6 for more details). Whilst the performance of sleep-detection algorithms is mainly recorded in the literature using accuracy, accuracy may be affected by the frequency of occurrence of each state, e.g. if there are significantly more wake minutes and an algorithm classifies all minutes as wake, a method may achieve a high accuracy despite having a poor sensitivity. The kappa statistic incorporates the relative probability of chance agreement (P_e) and may therefore be a more appropriate statistic (see 3.6 for more detail on the kappa statistic). Confusion matrices were computed to evaluate

whether algorithms were more prone to any particular type of misclassification (Table 5.3).

		Self-reported		Total
		Positive (Sleep)	Negative (Wake)	
Automated	Positive (Sleep)	t_p	f_p	$t_p + f_p$
	Negative (Wake)	f_n	t_n	$f_n + t_n$
Total		$a + c$	$b + d$	N

Table 5.3: Example of a confusion matrix, where t_p = correctly identified sleep minutes, t_n = correctly identified wake minutes, f_p = falsely identified sleep minutes, f_n = falsely identified wake minutes.

In order to visualise the agreement between the self-reported and algorithm-derived sleep-boundary times, Bland-Altman plots were created both on a per-night and per-individual basis. Bland-Altman plots are a tool for visualising the agreement between two sources of measurement, displaying the average value of the two sources on the x -axis and the difference between them on the y -axis (Figure 5.8).

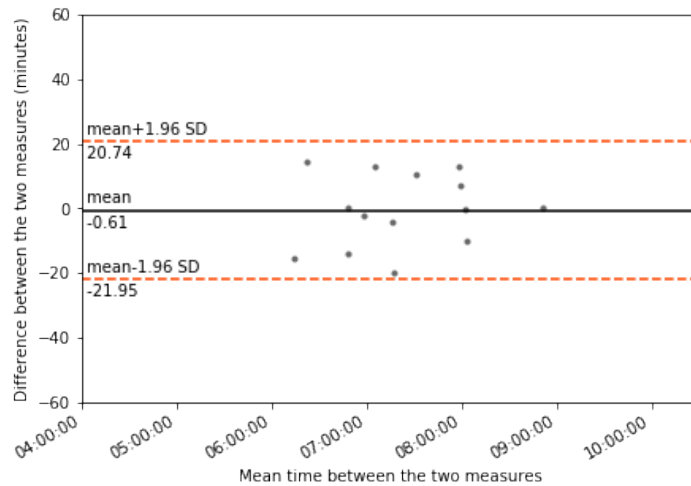


Figure 5.8: Typical Bland-Altman plot with data points shown in grey, the mean agreement shown in black and confidence intervals of the mean shown in orange.

Key performance markers were extracted for 30 randomly-selected partitions of the training and testing sets, and averaged over all iterations in order to obtain a robust

evaluation of the performance of the algorithms.

5.4 Results

5.4.1 Data Collection

Data from 46 participants in the CHA Exercise study were analysed. Participants had a median age of 65 years (62-70 years); 63% were female and 37% male. Participants recorded self-reported wake-up and sleep information for a median (IQR) duration of 18 days (17-27), with a median sleep-onset time of 23:30pm (23:00-00:15pm) and a median wake-up time of 7:15am (6:35-8:00am). Table 5.4 lists the summary characteristics for all participants included in the analysis. Figure 5.9 shows the distributions of the sleep-onset and wake-up times for all participants.

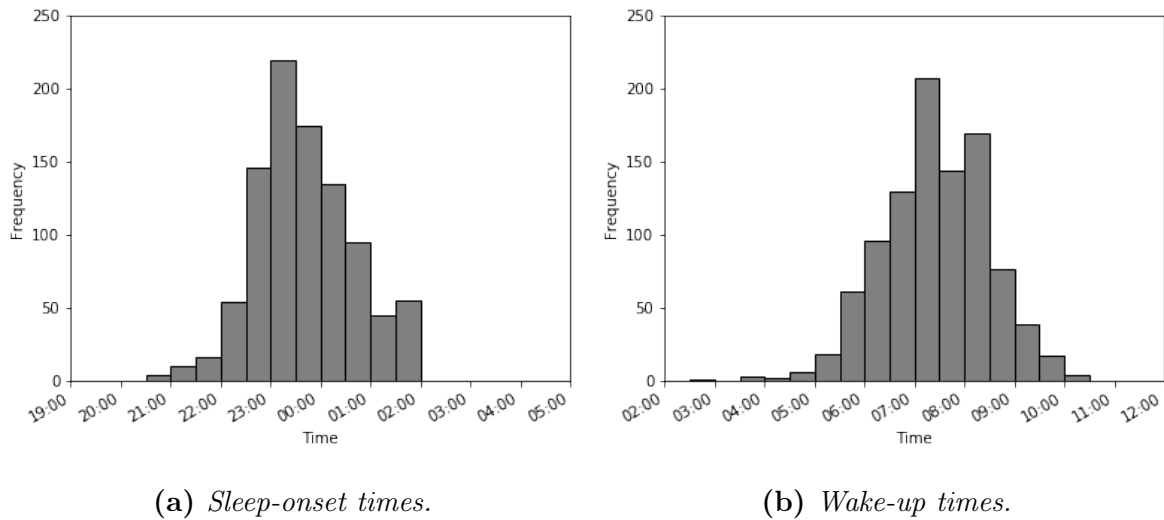
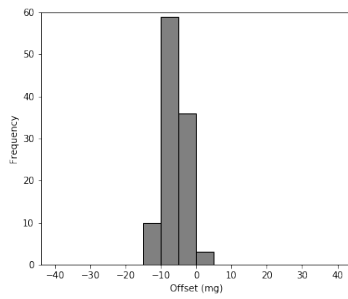


Figure 5.9: *Distribution of self-reported sleep-onset and wake-up times for participants in the CHA Exercise study.*

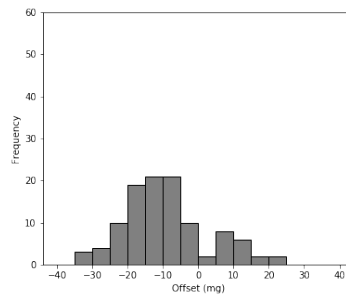
5.4.2 Data Preparation

5.4.2.1 Calibration

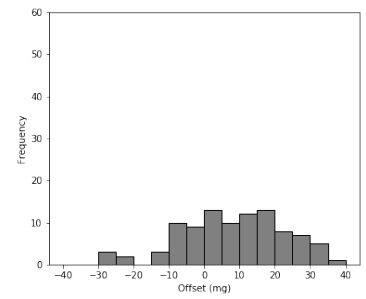
The distributions of calibration coefficients (see Section 5.3.2.1 for more detail) for all participants included in the analysis are presented in Figure 5.10. Median offsets along the x-, y- and z-axis were -6.3mg, -10.1mg and 11.6mg respectively. The median calibration slopes were 0.9998, 0.99996 and 1.0029 respectively (Table 5.4).



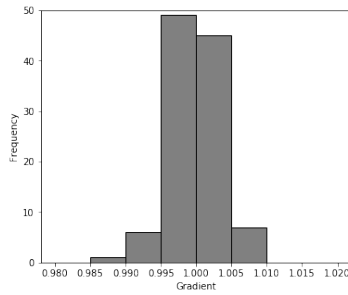
(a) *Distribution of x-offset values.*



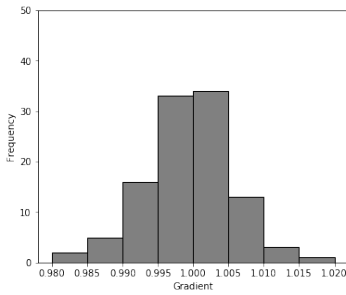
(b) *Distribution of y-offset values.*



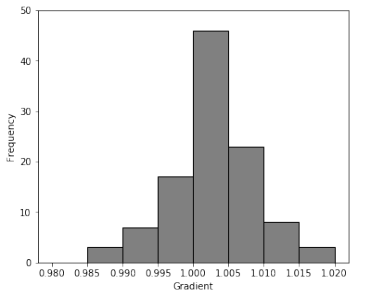
(c) *Distribution of z-offset values.*



(d) *Distribution of x-slope values.*



(e) *Distribution of y-slope values.*



(f) *Distribution of z-slope values.*

Figure 5.10: *Histograms of calibration offset and slope values.*

5.4.2.2 Wear-time Detection

After removing non-wear periods from the accelerometer recordings, a total of 967 days were included in the dataset, with corresponding self-reported sleep-onset and wake-up labels available for 947 days. The final selection of days with accelerometer recordings

and corresponding self-reported data was obtained from 22 participants with three sets of recordings, 18 participants with two sets of recordings and six participants with a single recording.

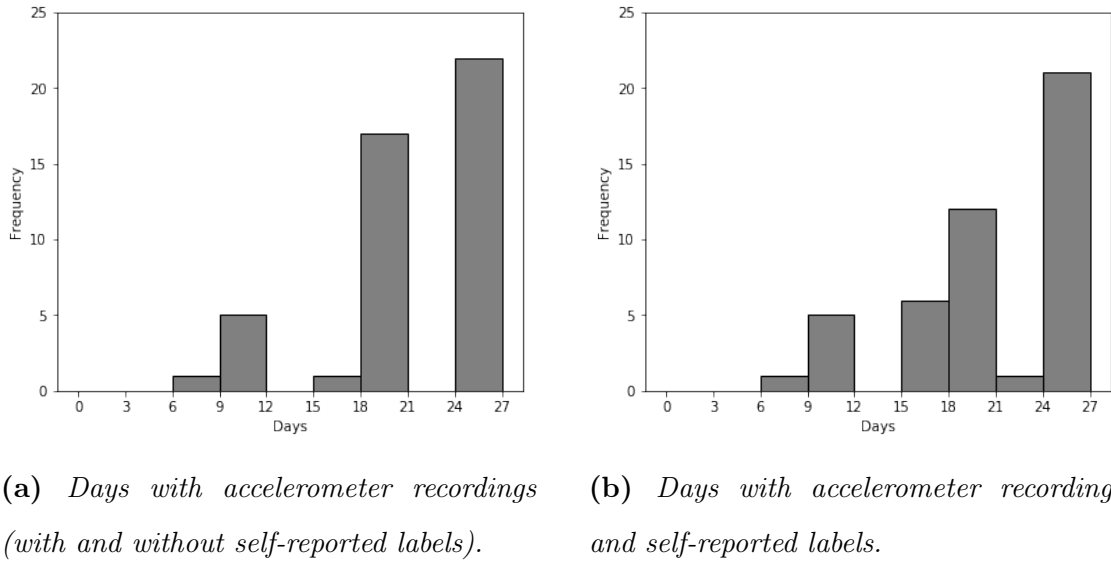


Figure 5.11: Histograms of total accelerometer wear days (left) and those with corresponding self-reported sleep-onset and wake-up labels (right). 'Frequency' refers to the number of participants with corresponding numbers of days with available data.

	CHA Participants (N = 46)
Age	65 (62-70)
Sex (% female)	63%
Self-reported (days)	18 (17,27)
Calibration	
x-offset (mg)	-6.3 (-8.5,-4.1)
y-offset (mg)	-10.1 (-8.5,-41)
z-offset (mg)	11.6 (1.0,23.9)
x-slope	0.9998 (0.9977, 1.0018)
y-slope	0.9996 (0.9957, 1.0032)
z-slope	1.0029 (1.0000, 1.0068)
Wear-time	
All (days)	18 (18,27)
With labels (days)	18 (17,27)

Table 5.4: *Baseline and preprocessing information for participants included in the sleep-boundary-detection analysis. Data shown as median (inter-quartile range).*

5.4.3 Data Analysis

Minute-by-minute sleep and wake labels derived from the self-reported sleep-onset and wake-up times were used as the ground truth for the development and evaluation of all algorithms. Figure 5.12 shows a typical accelerometer trace, from participant P082, with sleep and wake labels shaded in grey and red respectively.

5.4.3.1 Sleep Detection using Data-driven Thresholds

Results are shown for an example iteration of one of the training and testing set partitions. The PDF-based $v_{mag_{std}}$ thresholds (τ_i), which best differentiated between sleep and wake times for individuals in the training set are presented in Figure 5.13a. The orange dashed line shows the averaged threshold across the participants ($\tau_{PDF} = 0.074$ g) selected for further analysis. Figure 5.14a shows the PDF-based sleep labels (grey solid line) as well as the self-reported sleep and wake times (orange dashed line) for a typical participant from the training set (P082). The output of the classifier was noisy;

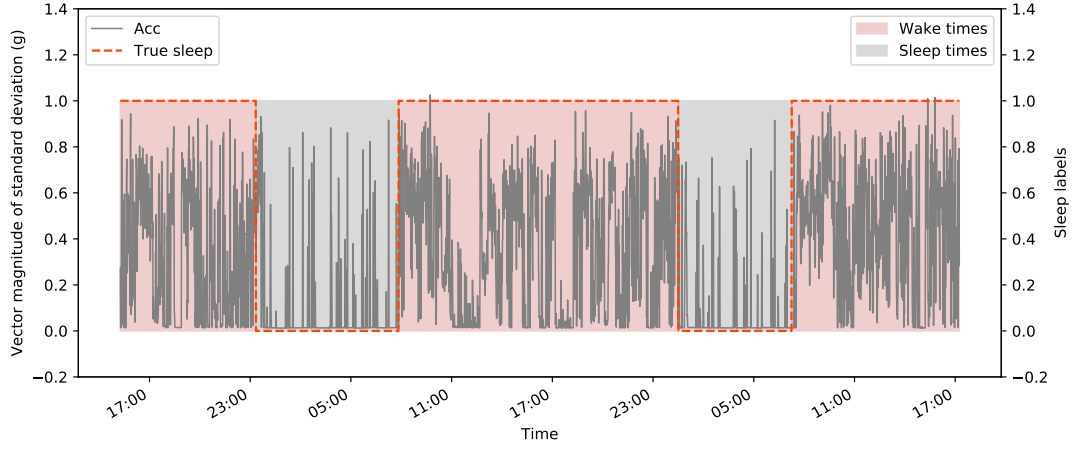
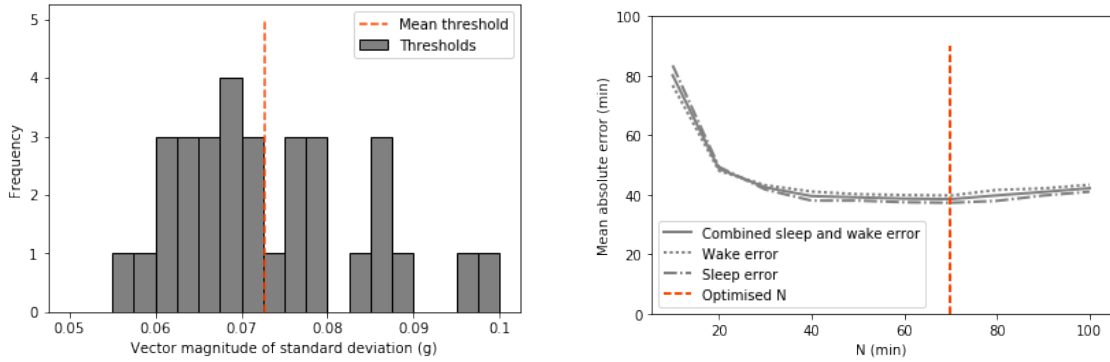


Figure 5.12: $v_{mag_{std}}$ (grey solid line) and self-reported (orange dashed line) sleep labels for a typical subject from the CHA Exercise study. Red and grey shaded areas correspond to wake and sleep times based on self-reported sleep information.

however trends of sleep and wake times were identified correctly with a majority of wake classifications during wakefulness, and sleep classifications during sleep.

Figure 5.13b shows the mean absolute error (MAE) between the self-reported and PDF-derived sleep-onset and wake-up times for different smoothing filter lengths. The orange dashed line represents the filter length ($N_{PDF} = 70$ minutes) which resulted in the smallest MAE for the training set. Figure 5.14b illustrates the PDF-based sleep-onset and wake-up times after applying the rolling-average smoothing filter with hardlimiter. In this particular example, sleep and wake periods were correctly classified with the exception of short periods during wake times. These misclassifications may have been caused by restful periods during the day. Figure 5.14c shows the PDF-based sleep and wake classifications for participant P082 after selecting the longest continuous sleep durations as final sleep periods. The PDF-based method identified the sleep-onset time in the first night earlier than the self-reported time and finds a slightly later wake-up time than stated by the participant on the second day.



(a) Histogram of the PDF-based thresholds for sleep and wake classification (grey solid line), and mean threshold across all individuals in the training set (orange dashed line).

(b) Mean absolute error between self-reported and PDF-based sleep-onset and wake-up times for different rolling-average smoothing filter lengths (grey solid line), and best performing filter length (orange dashed line).

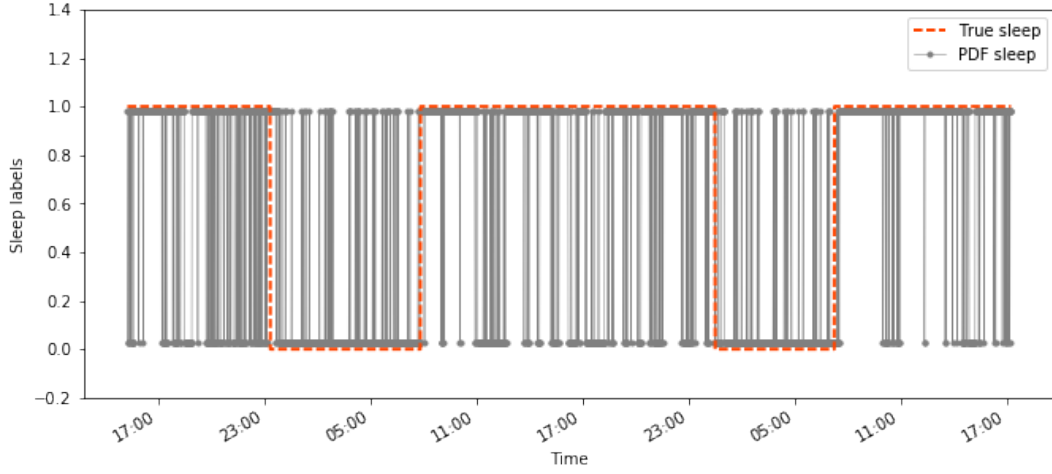
Figure 5.13: Parameters for PDF-based method.

The Bland-Altman plots in Figure 5.15 illustrate the agreement between the self-reported and PDF-based sleep-boundary times on the test set of the example partition (MAE on the per-night basis = 33 ± 14 minutes, MAE on the per-participant basis = 34 ± 10 minutes). On average, the PDF-based sleep-boundary detection approach underestimated total sleep times based on self-reported information. On the per-night criterion, the method identified sleep-onset times 15 minutes late, wake-up times on average two minutes late. On the individual basis, the PDF-based sleep-boundary detection algorithm (with rolling-average smoothing filter and hardlimiter) found sleep-onset times on average 17 minutes later than the self-reported sleep-onset time and wake-up times on average zero minutes later. When evaluating the algorithm's minute-by-minute agreement with self-reported sleep times (on the example partition), the PDF-based method achieved a mean accuracy of 95% and a kappa statistic of 0.89. The confusion matrix for the algorithm is shown in Figure 5.5 and shows that 3% of misclassifications were due to wake times being wrongly identified as sleep times and 2% due to actual sleep times being classified as wake times. The performance characteristics of the PDF-based

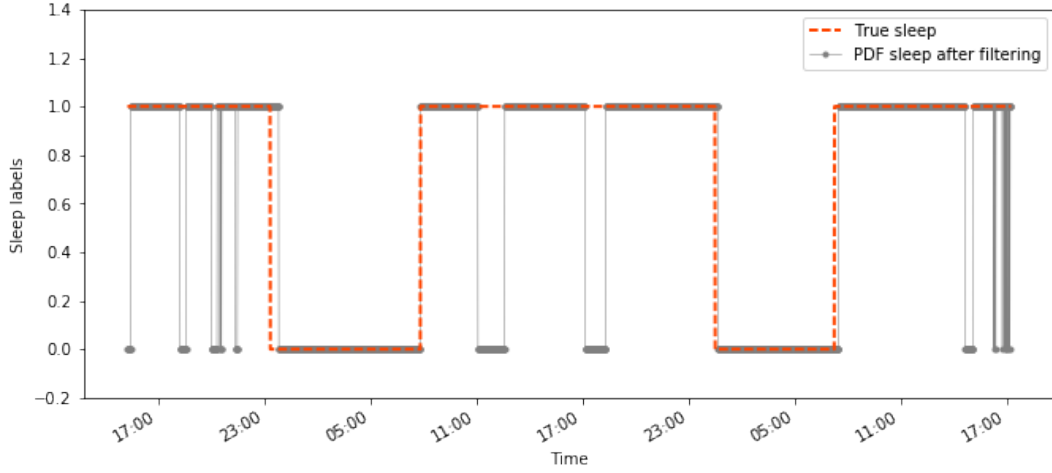
sleep-boundary detection method averaged across all 30 random partitions of training and test sets are summarised in the conclusion section of this chapter (Table 5.8).

		Self-reported		Total
		Positive (Sleep)	Negative (Wake)	
Automated	Positive (Sleep)	114, 575 32%	10, 472 3%	125, 047
	Negative (Wake)	6, 863 2%	220, 642 63%	227, 505
Total		121, 438	231, 114	N

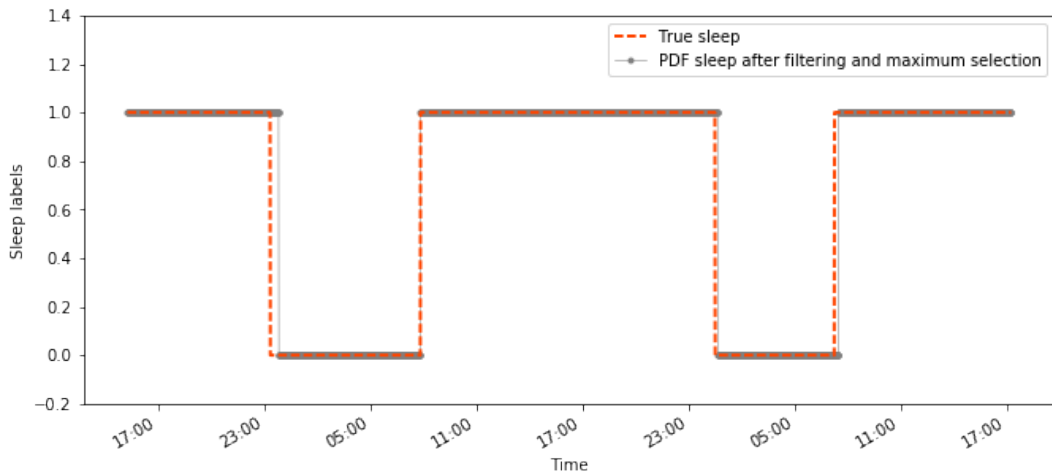
Table 5.5: *Confusion matrix of PDF-based classifications in the test set. Numbers represent the number of one-minute epochs in each category. Percentages represent percentage of one-minute epochs in each category versus overall number of one-minute epochs.*



(a) Sleep and wake labels after thresholding with $\tau_{PDF} = 0.074$ g (PDF classifications are reduced in amplitude for readability purposes).

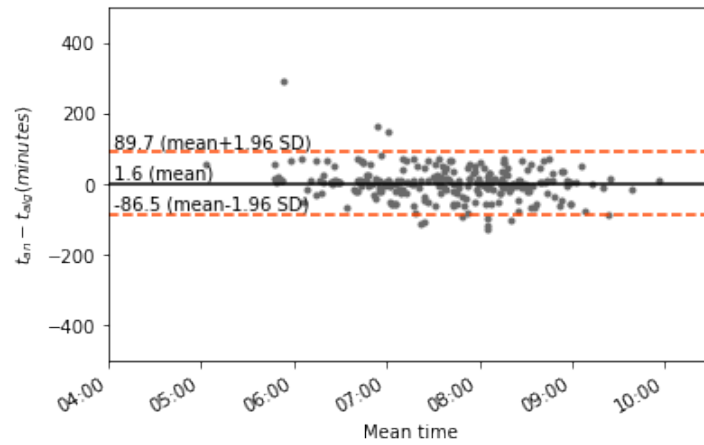


(b) Sleep and wake labels after thresholding with $\tau_{PDF} = 0.074$ g and applying the rolling-average smoothing filter ($N_{PDF} = 70$ minutes) and hardlimiter.

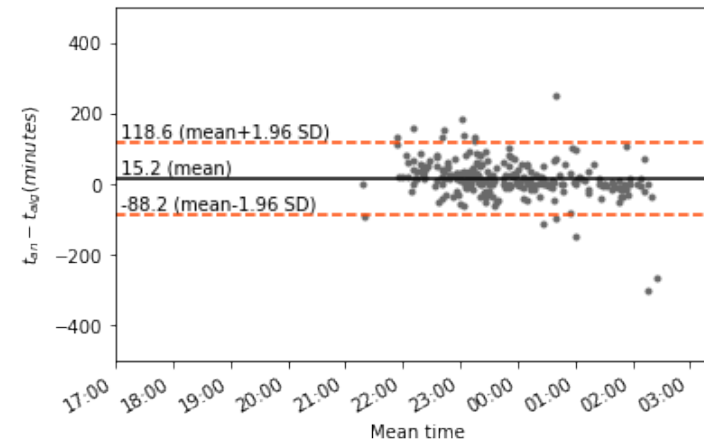


(c) Sleep and wake labels after thresholding ($\tau_{PDF} = 0.074$ g), filtering ($N_{PDF} = 70$ minutes) and maximum selection.

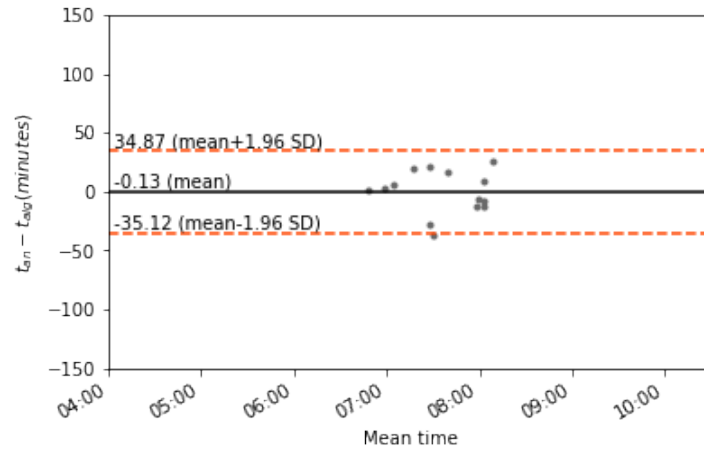
Figure 5.14: Self-reported (orange dashed line) and PDF-based (grey solid line) sleep-onset and wake-up times.



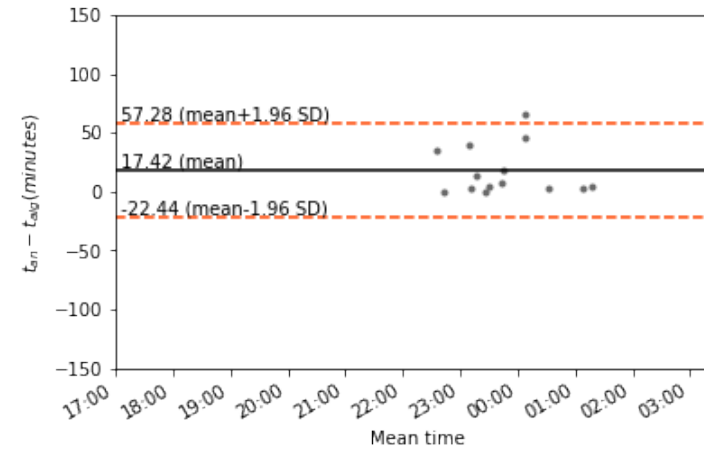
(a) Wake-up times for nights in the test set.



(b) Sleep-onset times for nights in the test set.



(c) Wake-up times for individuals in the test set.



(d) Sleep-onset times for individuals in the test set.

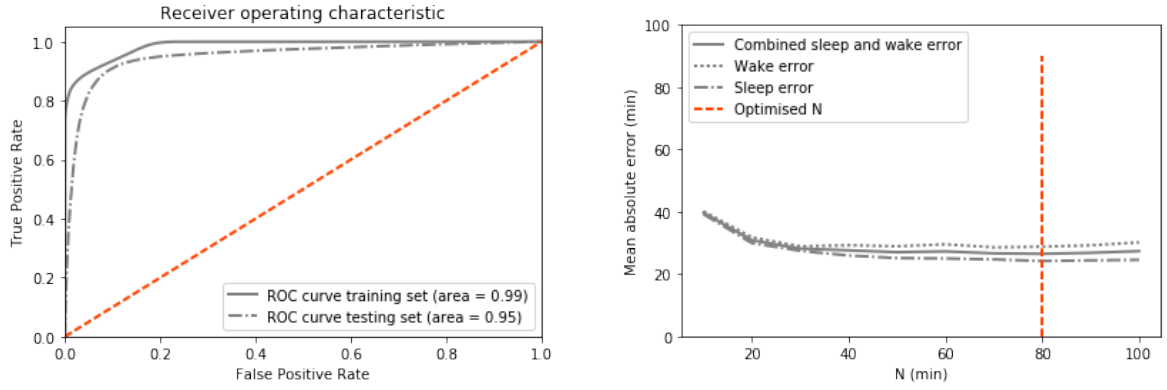
Figure 5.15: Bland-Altman plots of agreement between self-reported (t_{an}) and PDF-based (t_{alg}) sleep-onset and wake-up times.

5.4.3.2 Sleep Detection using a Random Forest Classifier

The RF classifier used a combination of time- and frequency-domain variables to classify minute-wise sleep and wake labels. The assigned labels were subsequently post-processed in a similar manner to the PDF-based method to obtain continuous periods of sleep and wakefulness and hence sleep- onset and wake-up times. Figure 5.16a summarises the performance of the trained random forest classifier on a typical *Tr_ml* training set (grey solid line) and *Te_ml* testing set (grey dashed line) using ROC curves. The area under the curve scores for the classification were 1.00 (rounded) and 0.93 respectively, suggesting that the classifier performed well on both sets of data. Varying the number of trees and branch depth (via minimum number of samples required to split into a new branch) between 10 and 400, and 10 and 1000 respectively, only resulted in minor changes in AUC scores (all values fell between 0.93 and 0.94). For computational cost reasons the best performing setting with the lowest number of trees was chosen for further analysis (number of trees: 10, minimum split size: 10). The ranking of the features as determined over 30 iterations of algorithm training is shown in Figure 5.17. The most important features identified were the 75th percentile, 25th percentile, median and entropy of the one-minute epoch period.

Figures 5.18a and 5.18b show ROC curves obtained over 30 iterations of RF classifier training using all, and the five most important features respectively. It can be seen that the performance of the classifier using all input features did not vary more than when using only five features. As a result, the remainder of the random forest based sleep-boundary detection was conducted using a classifier with all input features.

The random forest based sleep and wake classifications for a typical recording (P082) from the CHA study are presented in Figure 5.19a. The classifier assigned probabilities of being asleep or awake rather than binary labels. From the figure, it can be seen that the classifier performed well on a minute-to-minute basis.



(a) Receiver operation curves (ROC) of random forest classifier on Tr_ml (red) and Te_ml (blue) set.

(b) Mean absolute error between self-reported and random forest based sleep-onset and wake-up times for different filter lengths (grey solid line) and best performing filter length (orange dashed line).

Figure 5.16: Parameters for random forest based sleep-boundary-detection. '25thp' and '75thp' refer to the 25th and 75th percentile, 'sd' to the standard deviation, 'corr' to the correlation, 'avg' to the average and 'fft0' to 'fft14' to the extracted frequency bands.

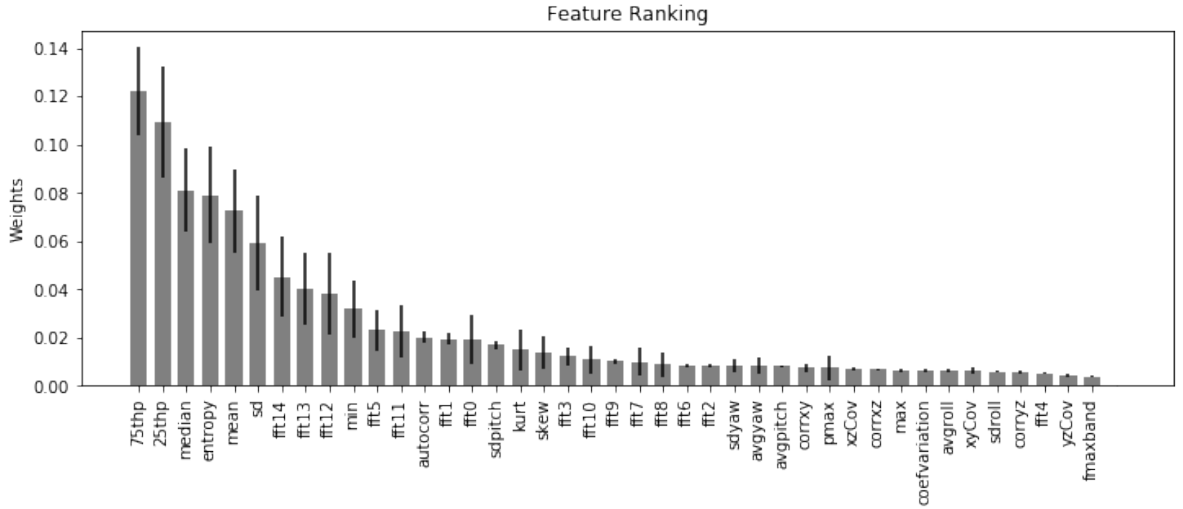
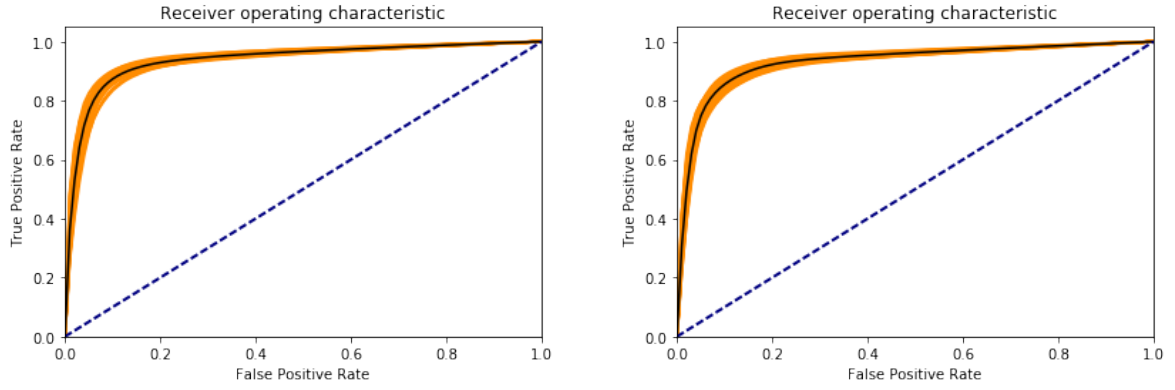


Figure 5.17: Random forest classifier feature importance based on 30 iterations.

After training the random forest classifier, different rolling-average smoothing filter lengths N_{rf} were tested in order to optimise the performance of the sleep-boundary-detection algorithm. Figure 5.16b shows the MAE of the sleep-boundary-detection al-



(a) Receiver operation curves (ROC) of performance of random forest classifier on *Tr_ml* (red) set over 30 iterations (orange lines). Mean across the iterations is shown in black.

(b) Receiver operation curves (ROC) of performance of random forest classifier with five most important features on *Tr_ml* (red) set over 30 iterations (orange lines). Mean across the iterations is shown in black.

Figure 5.18: Evaluation of the effect of large number of input features.

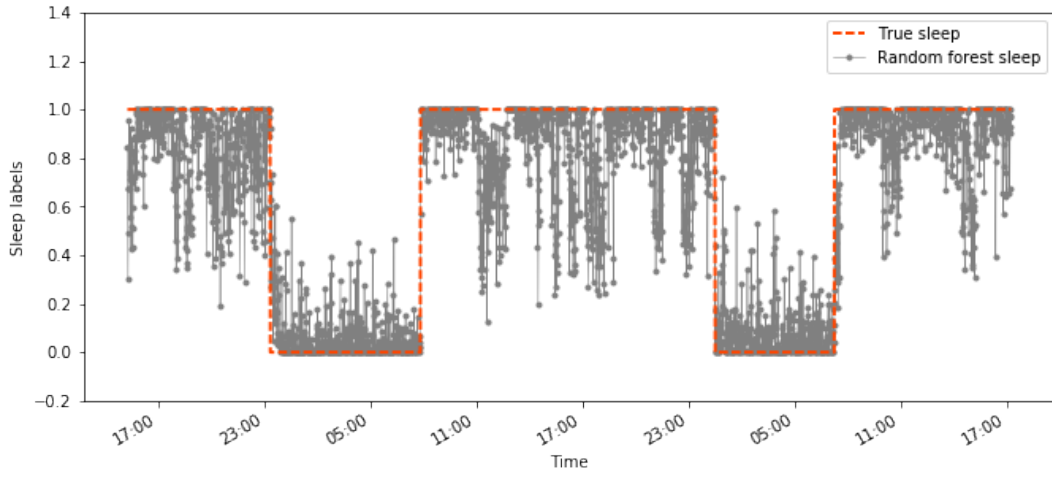
gorithm for different values of N_{rf} (grey solid line) and the optimal length N_{rf} of 80 minutes on the training set (for an example partition). The wake- and sleep-times after filtering (with hardlimiter) and after maximum selection are presented in Figures 5.19b 5.19c respectively.

In the example partition of the dataset, the random forest based sleep-boundary detection achieved a MAE of 31.4 ± 15.6 minutes on a per-night basis and a MAE of 32.3 ± 9.3 minutes (mean $\pm$ standard deviation) averaged across participants. On a per-night basis, the method closely estimated total sleep times, identifying wake-up times on average six minutes late and sleep-onset times on average four minutes late (Figure 5.20). On an individual basis, the RF-based method identified wake-up times on average five minutes late and sleep-onset times on average six minutes late. The minute-by-minute accuracy achieved by the RF-based method was 95% and the kappa score 0.89 for the example partition. The confusion matrix (Figure 5.6) shows that a similar proportion of misclassifications were due to sleep times being wrongly classified

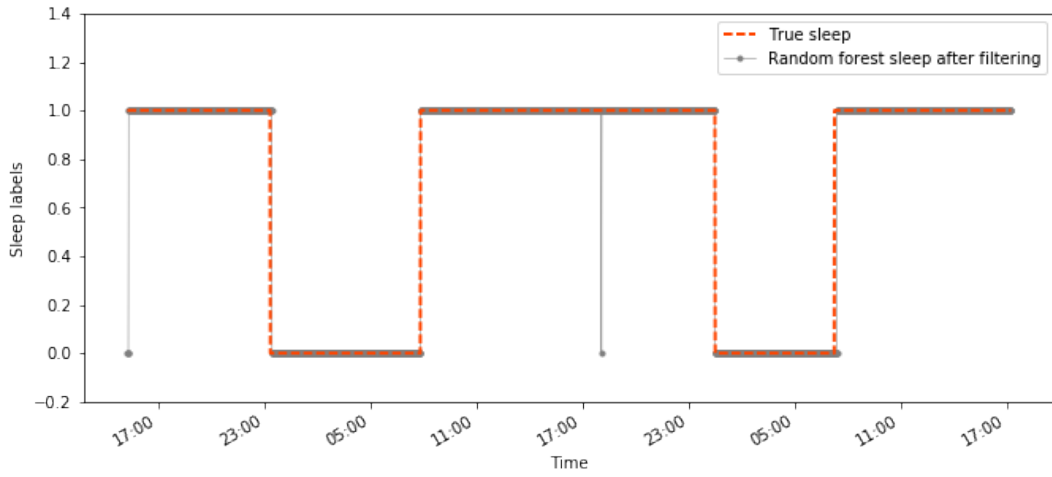
		Self-reported		
		Positive (Sleep)	Negative (Wake)	Total
Automated	Positive (Sleep)	112, 874 32%	7, 946 2%	120, 820
	Negative (Wake)	8, 564 2%	223, 168 63%	231, 732
Total		121438	231114	N

Table 5.6: *Confusion matrix of RF-based classification on the test set. Numbers represent the number of one-minute epochs in each category. Percentages represent percentage of one-minute epochs in each category versus overall number of one-minute epochs.*

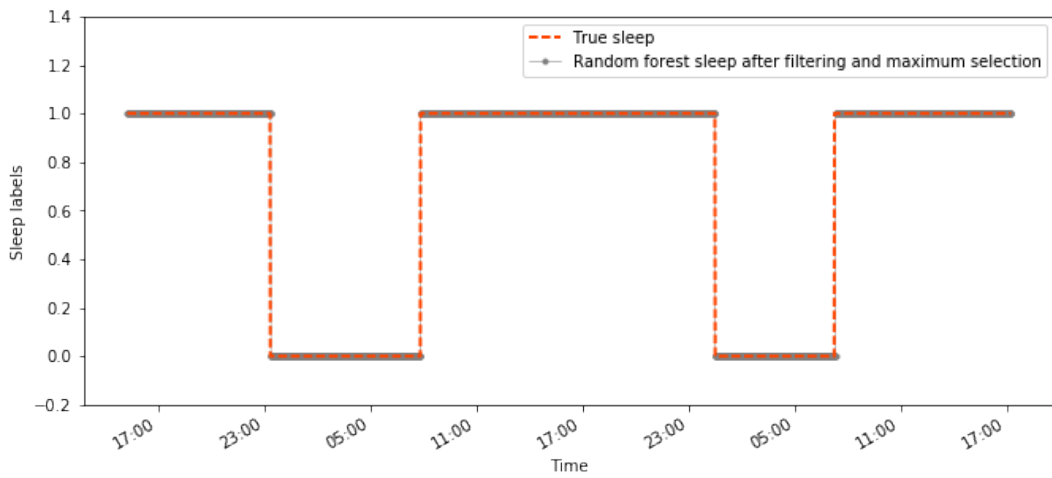
as wakefulness and vice versa. The performance characteristics of the RF-based method averaged across all 30 random partitions of training and test sets are summarised in the conclusion section of this chapter (Table 5.8).



(a) Sleep and wake labels from random-forest classifier.

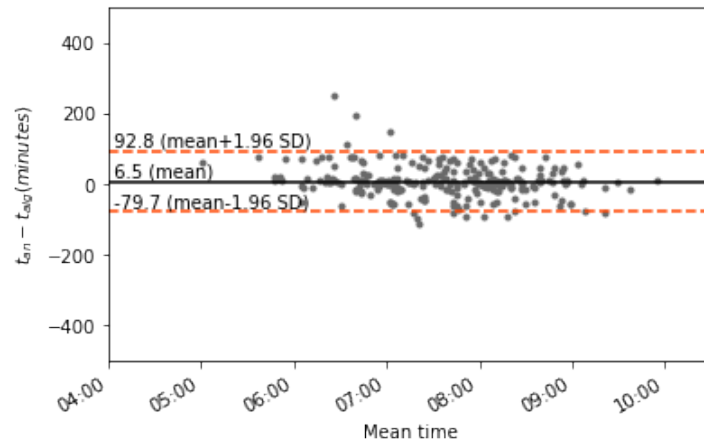


(b) Sleep and wake labels after applying rolling-average smoothing filter ($N_{rf} = 80$ minutes) and hardlimiter.

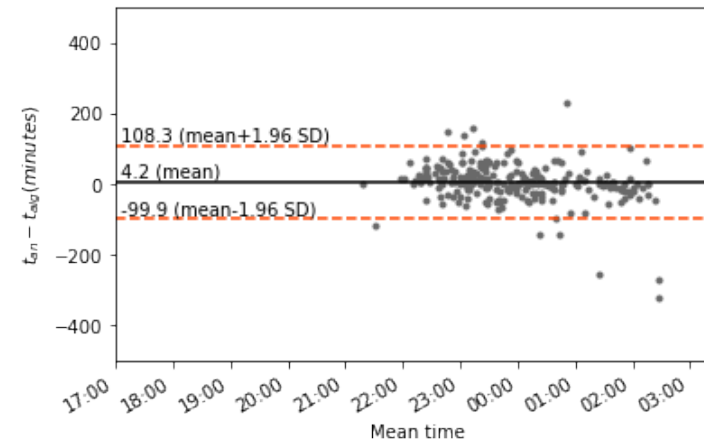


(c) Sleep and wake labels after thresholding, filtering ($N_{rf} = 80$ minutes, with hardlimiter) and maximum selection.

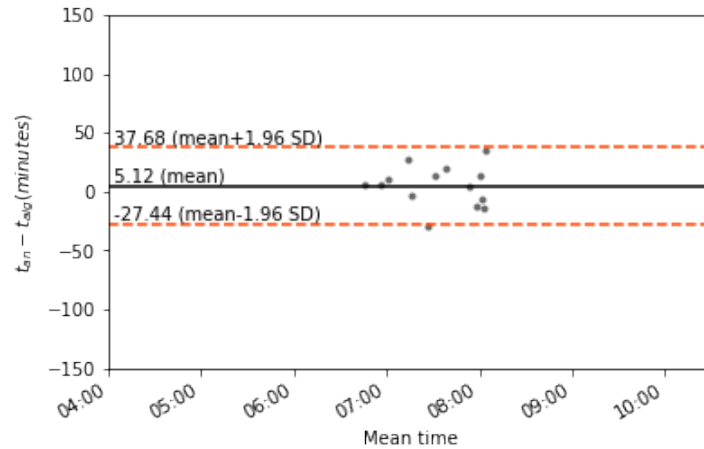
Figure 5.19: Self-reported (orange dashed line) and PDF-based (grey solid line) sleep and wake times.



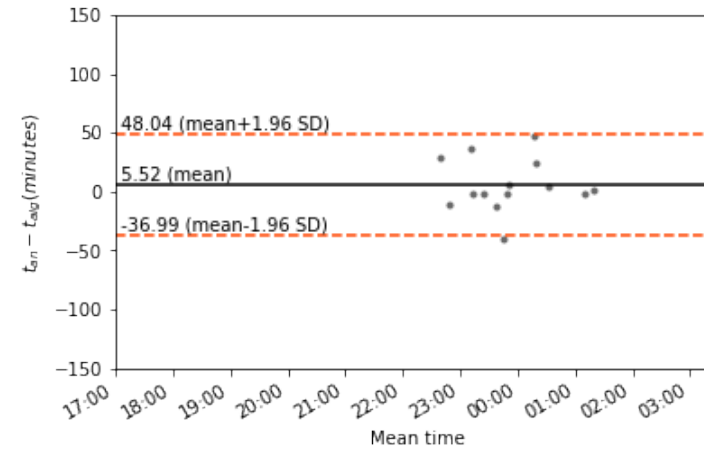
(a) Wake-up times for nights in the test set.



(b) Sleep-onset times for nights in the test set.



(c) Wake-up times for individuals in the test set.

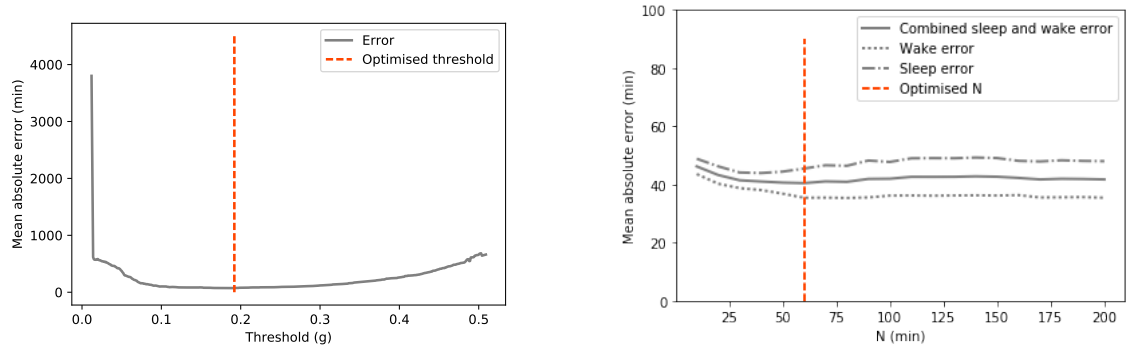


(d) Sleep-onset times for individuals in the test set.

Figure 5.20: Bland-Altman plots of agreement between self-reported (t_{an}) and random forest based (t_{alg}) sleep-onset and wake-up times.

5.4.3.3 Sleep Detection using Change-point Detection

The final method tested for sleep-boundary detection was the CPD approach. The optimal threshold τ_{CPD} to classify periods into sleep or wakefulness was found to be 0.192 g (see Figure 5.21a). These values were used for the further evaluation of the CPD-based algorithm on the test set.



(a) Identification of the best performing τ_{CPD} .

(b) Mean absolute error between self-reported and CPD-based sleep-onset and wake-up times for different filter lengths (grey solid line) and best performing filter length (orange dashed line).

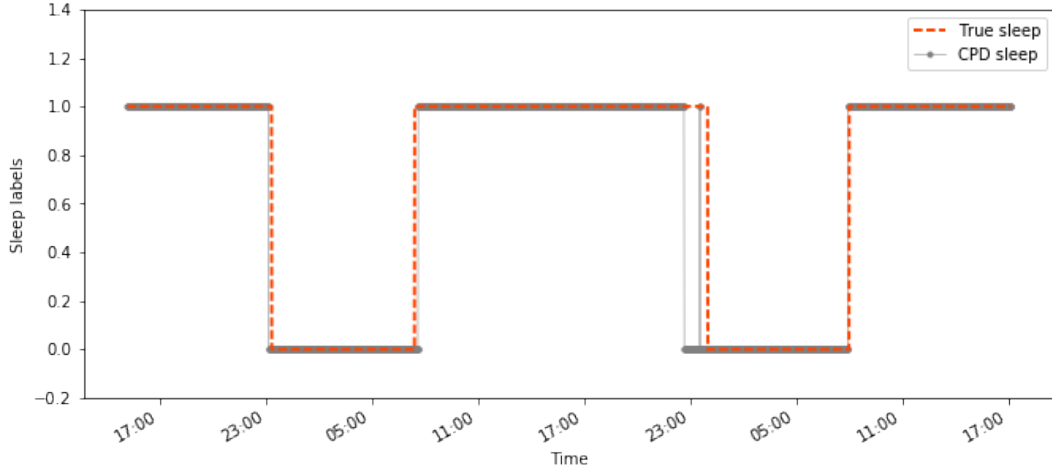
Figure 5.21: Parameters for the CPD based sleep-boundary-detection.

The CPD-based sleep (0) and wake (1) classifications on candidate P082 are presented in Figure 5.22a. The optimal rolling-average smoothing filter length N_{CPD} for the change-point-detection algorithm was found to be 60 minutes (Figure 5.21). Figures 5.22b and 5.22c show the output from the algorithm after filtering (with hardlimiter) and maximum sleep selection for the same participant (P082) as before. They demonstrate that CPD-based sleep-boundaries can be very accurate (as seen in the first night); however when the algorithm misclassifies the change from one state to another, mistakes can be very significant (as seen in the second night) as the algorithm classifies entire segments rather than individual minutes.

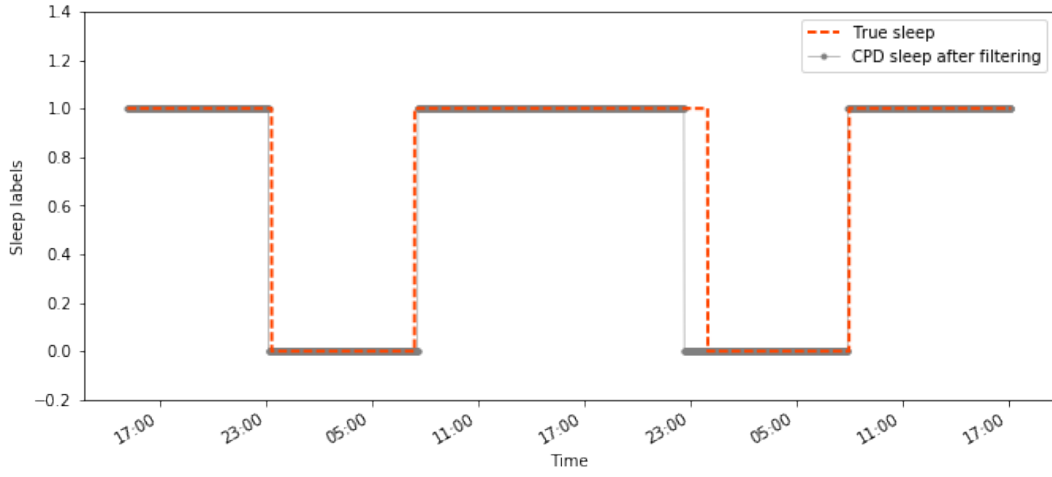
		Self-reported		Total
		Positive (Sleep)	Negative (Wake)	
Automated	Positive (Sleep)	114,269 32%	10,119 3%	136035
	Negative (Wake)	7,169 2%	220,995 63%	245392
Total		121,438	231,114	N

Table 5.7: *Confusion matrix of CPD-based classifications in the test set. Numbers represent the number of one-minute epochs in each category. Percentages represent percentage of one-minute epochs in each category versus overall number of one-minute epochs.*

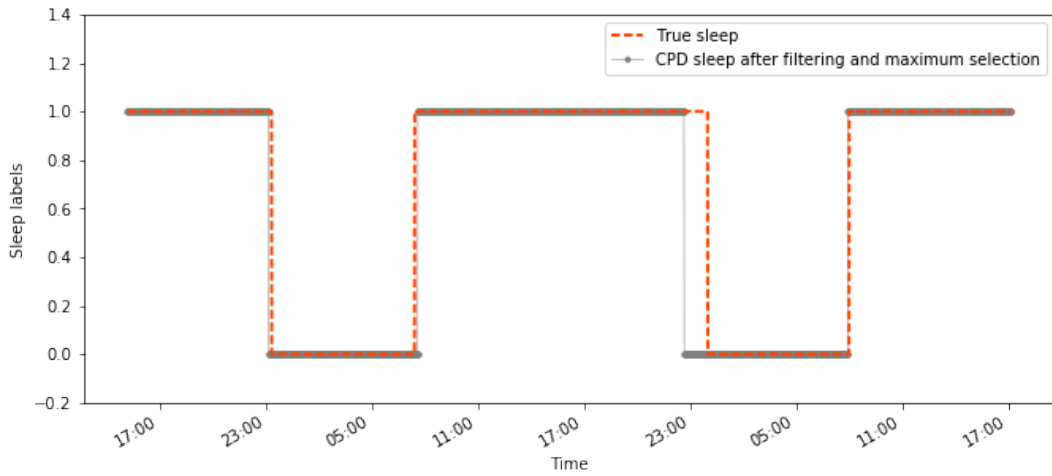
The Bland-Altman plots provided in Figure 5.23 summarise the performance of the algorithm on the test set of the example partition. On the per-night level, the CPD-approach showed a slight bias towards identifying sleep-onset times before the self-reported sleep-onset and wake-up times later than the self-reported times. On the individual-level analysis, the algorithm identified both sleep-onset and wake-up times on average later than the self-reported times. The minute-by-minute accuracy achieved by the CPD-based method was 94% and the kappa score 0.87. The confusion matrix (Figure 5.7) shows that a slight majority of misclassifications by the CPD algorithm were due to false positives (wrongly identified sleep times).



(a) Sleep and wake labels after thresholding with $\tau_{CPD} = 0.192$.

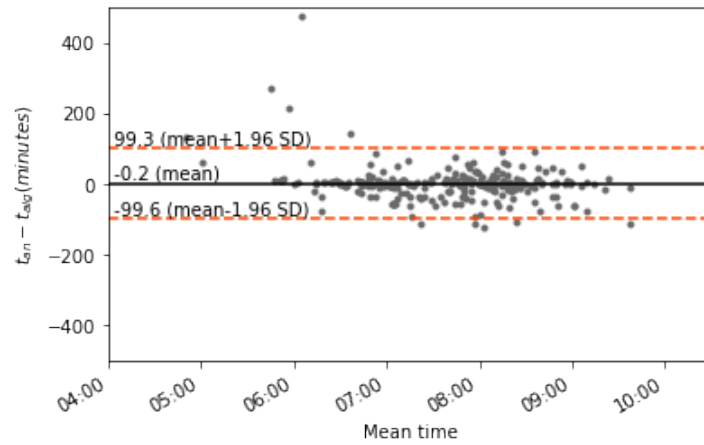


(b) Sleep and wake labels after thresholding with $\tau_{CPD} = 0.192$ and filtering ($N_{CPD} = 60$, with hardlimiter).

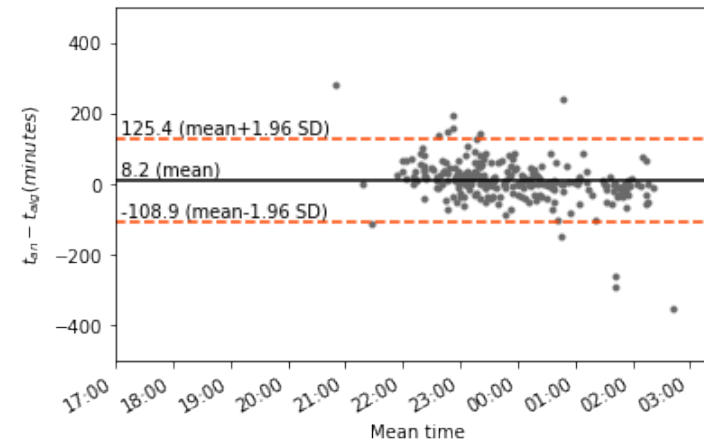


(c) Sleep and wake labels after thresholding ($\tau_{CPD} = 0.192$), filtering ($N_{CPD} = 60$, with hardlimiter) and maximum selection.

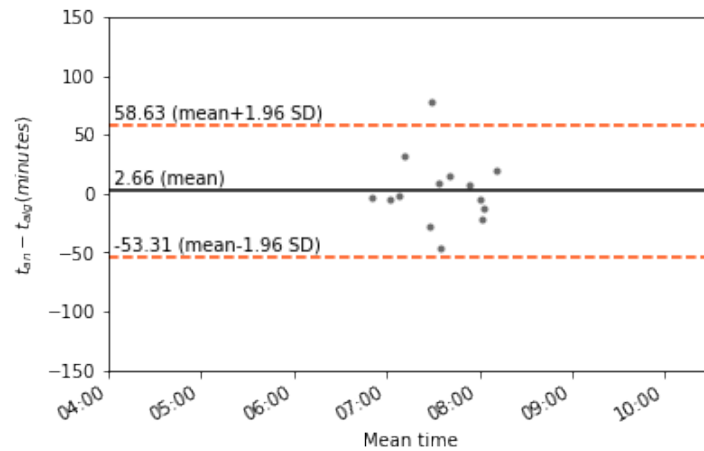
Figure 5.22: Self-reported (orange dashed line) and CPD-based (grey solid line) sleep and wake times.



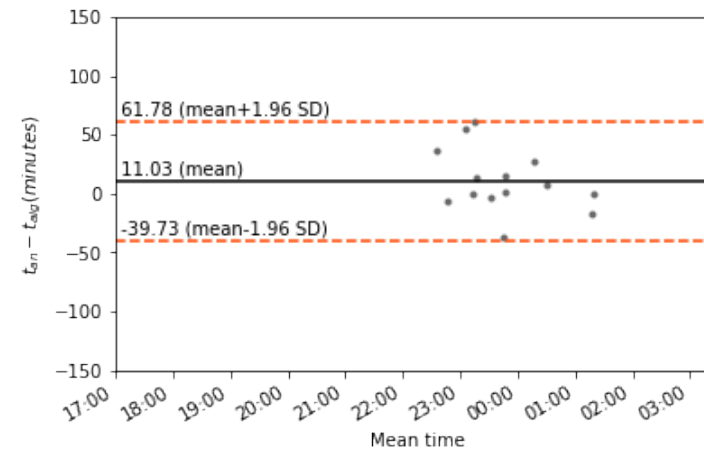
(a) Wake-up times for nights in the test set.



(b) Sleep-onset times for nights in the test set.



(c) Wake-up times for individuals in the test set.



(d) Sleep-onset times for individuals in the test set.

Figure 5.23: Bland-Altman plots of agreement between self-reported (t_{an}) and CPD-based (t_{alg}) sleep-onset and wake-up times.

5.4.4 Evaluation

The performance of all three sleep-boundary detection algorithms averaged over 30 iterations is summarised in Table 5.8. The kappa statistic and accuracy refer to the algorithms' ability to detect sleep and wake times on a minute-by-minute level. The MAE values refer to their performance in comparison to the self-reported sleep-boundary times. All algorithms achieved accuracies of 94% or above. Whilst the PDF and CPD methods had sensitivities of 93 % and specificities of 94% each, the RF method had a sensitivity (ability to detect sleep correctly) of 92% and a specificity (ability to detect wakefulness correctly) of 96%. The kappa statistics were very similar (0.88 to 0.89) with the PDF and RF algorithms marginally outperforming the CPD algorithm.

Per individual participant, the RF method performed best in identifying sleep onset times (MAE = 17.5 minutes) followed by the PDF and then the CPD approaches. The RF method also had the smallest MAE in identifying wake-up times (MAE = 19.7 minutes) followed by the CPD and then the PDF algorithms. Overall, the RF-based algorithm performed best with respect to the MAE averaged across all individuals (MAE = 35.5 minutes), with the PDF algorithm obtaining an average MAE of 37.7 minutes and the CPD algorithm achieving an average MAE of 40.7 minutes.

As can be seen in the Bland-Altman plots for each of the algorithms, all three methods were affected by outliers of sleep-onset/ wake-up differences, with some times differing by 300 minutes (5 hours) or more. Extreme detection errors (outliers in the figures) may be caused by misclassifications of restful periods during the day or active periods during sleep-time. For example, if an individual wakes up in the middle of the night and stays up for a long period of time before returning to sleep, this may result in their detected sleep duration for the night only being falsely classified as the time before or after the sleep interruption.

	PDF	RF	CPD
Accuracy	94% (1%)	95% (1%)	94% (1%)
Sensitivity	93% (1%)	92% (2%)	93% (1%)
Specificity	95% (1%)	96% (1%)	94% (1%)
Kappa-statistic	0.88 (1%)	0.88 (1%)	0.87 (2%)
MAE (per-night level)			
Overall	36.9 (4.2)	34.9 (3.9)	40.0 (4.8)
Sleep onset	37.2 (4.7)	34.4 (4.7)	46.0 (7.4)
Wake up	35.9 (4.8)	34.6 (4.5)	33.3 (4.8)
MAE (individual level)			
Overall	37.7 (4.1)	35.5 (3.8)	40.7 (4.5)
Sleep onset	19.7 (4.4)	17.5 (4.1)	23.8 (7.9)
Wake up	20.5 (5.0)	19.7 (4.5)	20.1 (4.5)

Table 5.8: Performance of sleep-boundary-detection algorithms versus self-reported data across 30 partitions of randomly selected training and test sets. MAEs are reported as mean (standard deviation) across the partitions.

5.5 Discussion

Three approaches for automatic sleep-boundary detection were developed and evaluated using data collected as part of the CHA Exercise study. Algorithms achieved high accuracies in identifying minute-by-minute sleep/ wake labels as defined by self-reported sleep-onset and wake-up times. In order to validate the algorithms' actual ability to identify minute-by-minute sleep state, additional validation against PSG would be required. Mean absolute errors (averaged across individuals) between self-reported and accelerometer-derived algorithms ranged from 35.5 to 40.7 minutes depending on the method used. The best performance was achieved using the RF method, which classified each minute of recording into sleep or wake. The minute-by-minute labels were then smoothed using a moving-average filter with hardlimiter in order to continuously detect periods of wakefulness and sleep, and hence sleep-onset and wake-up times. The performance of the RF method was closely followed by the performance of the basic PDF-based approach.

As the RF-based approach contained a wider range of features than the PDF method, it was no surprise that it performed better. However, there were only slight differences in MAEs of the RF and the PDF methods suggesting that $vmag_{std}$ contained most of the information required to detect sleep and wake times. Since the PDF method required less pre-processing and was therefore considered to be more robust across different devices and settings, it was considered most suitable for sleep-boundary detection.

Further investigation of the performance of the RF method revealed that the algorithm had a comparably large discrepancy between specificity and sensitivity suggesting that the classifier was better at identifying wake times correctly than at identifying sleep times correctly. Whilst the dataset was balanced before training, the duplication of a subset of sleep labels as part of the balancing process may have introduced sleep labels with a reduced variety of characteristics. This in turn may have affected the performance of the algorithm suggesting further improvements may be possible.

The CPD approach performed worst out of the three algorithms with larger MAEs at sleep onset and overall. The BA plots for the algorithm applied to individual recordings show that the CPD method was characterised by a majority of sleep-onset times with very little error (the individual points are distributed closer to the mean), however where there was an error, the error was comparably large. This was most likely due to the fact that the CPD algorithm identified entire periods of sleep or wakefulness rather than minute-wise labels. Further investigations of the performance of the CPD algorithm showed that the algorithm was much better at identifying wake-up times than at identifying sleep-onset times. This may be due to restful behaviour of participants before bed-time, which is difficult to differentiate from sleeping.

The MAEs of all algorithms were on average at least 35.5 minutes, with some outliers reaching MAE of 300 minutes and above. The analysis of outliers showed that two factors

introduced large differences between algorithm-derived and self-reported sleep-onset and wake-up times. First, restful periods during wakefulness or active periods during sleep were misclassified by the automated algorithms. Where these periods occurred for a long time, the smoothing filter could no longer adjust for them. Applying the maximum sleep selection in these cases resulted in significant errors in sleep-boundary times. Using only movement data, it is difficult, if not impossible, to differentiate between restful episodes and sleep. Additional data are required to reduce the misclassification rate of the algorithms, for example from different sensors.

The second main cause for misclassifications appeared to be potential errors in the self-reported data. Unlike previous sleep studies, this study was based on free-living data as opposed to data collected in a sleep laboratory. Such data provides more authentic information on a participant's natural sleep behaviour, however there is a risk of errors due to poor recall or rounding bias. Since even small differences in sleep duration have been shown to impact individual's health [127][128][129] and wellbeing [130], future work should focus on an improvement of the proposed algorithms, for example using training data extracted from PSG studies, which give objectively-recorded sleep-onset and wake-up times. In addition to this, the algorithms should be validated in a HF population, as sleeping behaviour of HF patients may be different to that of the elderly adults used for the development of the sleep-boundary detection methods. Finally, using PSG as a ground truth, the existing methods could be expanded to provide a new tool for the quantification of sleep-fragmentation (the number of short sleep interruptions a patient experiences during the night), a potentially valuable marker in HF patients [112]. The inclusion of additional sensor, such as heart rate or temperature sensors, may however be required in order to achieve accurate measurements of sleep fragmentation, as accelerometer data by itself cannot differentiate between wakeful rest and sleep.

In comparison to commercial devices (e.g. Fitbit), which may include functions to automatically detect sleep-boundary times, research-grade devices, such as the GENE-

Activ device used for this analysis, are generally better validated, provide access to raw accelerometer data and do not change their underlying filtering and pre-processing methods over time. As a result, latter have been preferred for medical research applications. The sleep-boundary detection methods described in this chapter are non-proprietary and useable across different devices, offering additional functionality for future accelerometer studies. An alternative approach to conduct accelerometer-based sleep analysis without the need of self-reported sleep-boundary times would be to use the 8 least active hours of the day. This approach is very simple, but does come with the disadvantage of potentially losing relevant information on an individual level (e.g. sleep durations of less or more than 8 hours). Future research could compare the two approaches in order to evaluate whether either approach is preferable.

5.6 Conclusion

This chapter proposed an automatic method to extract sleep-boundary times from wrist-worn accelerometer data. Three different approaches were implemented, evaluated and compared. A random forest method marginally outperformed simple PDF-based and change-point-detection approaches, however due to its simple and robust input variable, the PDF-based method was considered most suitable for sleep-boundary detection. For all methods, differences in algorithm-derived and self-reported sleep times may have been due to misclassifications of restful periods, particularly close to sleep times, as well as errors in self-reporting. Future work should therefore focus on merging self-reported sleep data with objectively-recorded polysomnography data in order to create a more reliable ground truth for classification algorithms.

The work described in this chapter fills a methodological gap in the processing of accelerometer data and lays the foundation for the accelerometer analysis of HF patients described in the following chapter. In particular, the sleep-boundary detection methods described in this chapter allow for the automatic detection of sleep-onset and

wake-up times, which will be used to extract PA markers during wake-times as well as accelerometer-derived estimated of subjects' sleep duration and sleep fragmentation. Both PA [37] and sleep [59] information may provide meaningful markers for HF risk prediction.

Heart Failure Classification using Wrist-worn Accelerometer Data

6.1 Introduction

The previous chapter explored different methods for automatic sleep-boundary detection using wrist-worn accelerometers and found a PDF-based method to be the most suitable to detect sleep-onset and wake-up times. This chapter used the sleep-boundary-detection technique alongside other feature extraction methods to explore accelerometer signals from HF patients and identify potential markers for repeated-measurement risk-stratification using accelerometer data.

Wearable accelerometers for human movement analysis can provide insights into both the PA and resting behaviour of individuals [131][132]. The growing availability of such devices makes them a promising tool for medical monitoring applications. These devices could prove particularly valuable for the management of chronic conditions such as HF. To date, research on accelerometer-derived movement patterns of HF patients has been limited [41][44] and the association between HF and changes in PA and sleep patterns remains little understood.

This chapter set out i) to study the ability of accelerometer-derived PA and sleep markers to differentiate between individuals with and without diagnosed HF and ii) to evaluate the added value of such markers when combining them with established HF markers. The analyses laid out in this chapter were conducted using baseline data and accelerometer recordings from the UK Biobank resource.

6.2 Background

6.2.1 Accelerometer-derived Markers for Physical Activity and Sleep

Accelerometer signals can be analysed by means of descriptive statistics (e.g. mean, standard deviation of acceleration) or derivative markers (e.g. intensity of activity or activity type). Descriptive statistics summarise the collected data in a structured and accurate manner but are not necessarily translatable into common concepts of PA analysis such as type or intensity of activity. Derivative markers on the other hand translate the accelerometer time-series signal into more common measures of body movement; this does however come at the cost of accuracy [133]. The analysis described in this chapter explored a range of both descriptive statistics and derivative markers that are further described in the methodology section. The following paragraphs provide an initial overview of common derivative markers used in accelerometer analysis.

Among the first researchers to report on the potential use of accelerometers for human movement analysis were Coates and Meade [134][135]. In 1960, the authors found that the acceleration of a waist-worn accelerometer, measured during the lift motion of the walking cycle, multiplied by the individual's body weight and step frequency, was highly correlated ($r = 0.98$) with oxygen consumption during walking. Based on these findings, Coates and Meade concluded that step counting is a viable method for measuring energy expenditure. In their 1981 paper, Wong *et al.* subsequently described

the development of a portable accelerometer-based device to measure energy expenditure in humans [136]. The association between measured acceleration from this device during different intensities of walking was later shown to be positively correlated with oxygen consumption [137][134].

Since the appearance of the first portable accelerometers for humans, various algorithms have been developed to convert often proprietary accelerometer summary statistics (also known as counts) to activity scores [131][132]. Activity is commonly quantified using metabolic equivalents of task (METs), a measurement of oxygen consumption: the average person uses 3.5 ml O₂/kg/min when seated at rest. By convention, this corresponds to 1 MET (1 kcal/kg/h). Algorithms to convert accelerometer data into METs differ dependent on device used, wearing location (e.g. waist, wrist, foot) and population studied. Thresholds for children, for example, are different to those for adults or elderly subjects.

In addition to MET scores, there has also been an interest in identifying particular activity types from accelerometer data [132][138]. Methods range from reference pattern recognition to machine-learning based classification algorithms trained on specific types of activities. As activity-detection methods become more specified, they tend to overfit the data to the activity or population on which they have been trained [138]. Overfitting indicates that the classifier has learnt the random variations within the training data rather than the underlying structure. There have therefore been attempts to develop so-called hierarchical activity classification algorithms, which detect different layers of activity with increasing detail albeit with reducing confidence the deeper one moves into the branches of the tree [139][138]. Mathie *et al.* [138], for example, developed an algorithm that uses threshold approaches to initially detect whether an individual is active or at rest and subsequently uses this information to identify the subtype of state in which an individual is (e.g. lying, sitting and intensity of activity). The authors developed the algorithm using wrist-worn accelerometer data from 26 participants and

found that their algorithm achieved a sensitivity of 98% and specificity of 99% on data collected in a laboratory environment.

For the analysis of the UK Biobank accelerometer data, only methods that use raw acceleration data from wrist-worn devices were of interest. In addition to this, algorithms developed using the GENEActiv (as introduced in the dataset chapter) device were of particular value as the Axivity (Axivity, Ltd., Newcastle, United Kingdom) armband used in the UK Biobank resource is based on the same sensor. Table 6.1 summarises algorithms that were considered for the analysis described in this chapter.

Lead author	Device	Sampling frequency	Input features	Classification type	Derivative marker
Bonomi	TracmorD	20 Hz	Time and frequency features	Decision tree	Activity type
Ellis	ActiGraph	30 Hz	Time and frequency features	Random forest	EE, Activity type
Esliger	GENEActiv	80 Hz	SVM_{gs}	Threshold	Sedentary, intensity
Hildebrand	GENEActiv	100 Hz	VM_{gs}	Threshold	EE, intensity
Rowlands	GENEActiv	100 Hz	x, y, z mean, SVM_{gs}	Threshold	Sedentary, intensity
Zhang	GENEActiv	80 Hz	SVM, FFT, DWT	Decision tree	Sedentary, household walking
Borazio	Custom	100 Hz	z standard deviation	Threshold	Sleep status
Van Hees	GENEActiv	85.7 Hz	Mean angle	Threshold	Sleep status

Table 6.1: Overview of algorithms for extracting derivative markers from raw accelerometer signals. SVM = signal vector magnitude. SVM_{gs} = gravity-adjusted signal vector magnitude. FFT = Fast Fourier Transform. DWT = Discrete Wavelet Transform. EE = Energy Expenditure measured in METs.

The first extraction of MET scores from the GENEActiv accelerometer was completed by Esliger *et al.* in 2011 [140]. They conducted a validation study of the GENE-

Activ accelerometer and derived acceleration thresholds for light, moderate and vigorous physical activities as well as sedentary behaviour. In the study, 60 adults wore accelerometers, one on each wrist, in addition to one on the waist and completed different intensity tasks whilst being monitored by a portable metabolic gas analyser. The authors subsequently used the gravity-subtracted signal vector magnitude (SVM_{gs}) (Equation 6.1) summarised over 60s epochs and the output from the gas analyser to identify suitable thresholds for the different PA intensities.

$$SVM_{gs} = \sum_{n=1}^M |\sqrt{x_n^2 + y_n^2 + z_n^2} - g|, \quad (6.1)$$

where M is the number of measurements per epoch, x , y , and z refer to the accelerometer axes and g is the gravitational force of the earth, $1g$. The final thresholds shown in Table 6.2 achieved sensitivities of 78% and above and specificities of 56% and above in the identification of the activity types.

Rowlands *et al.* further developed the work of Esliger *et al.* to include posture information using an approach that they named the sedentary sphere [133]. Activities were classified using the SVM_{gs} over 15s epochs as well as the mean arm angle ($angle_{mean}$) during each 15s epoch. A slightly adapted version of Esliger *et al.*'s thresholds was used: rather than using one threshold for each wrist, the authors combined the thresholds into generic cut-off values for both wrists. For the moderate and vigorous activity thresholds, Rowlands *et al.* used Esliger *et al.*'s left wrist thresholds. For the light activity threshold, the authors chose the mean of the right and left wrist thresholds found by Esliger *et al.*. They further adjusted the thresholds to cope with changes in frequency and epoch length, as shown in Equation 6.2 below.

$$thresh_{100Hz,15s} = thresh_{80Hz,60s} * \frac{100}{80} * \frac{15}{60} \quad (6.2)$$

	Left wrist	Right wrist	Non-dominant wrist
Eslinger <i>et al.</i> ($SV M_{gs}$ @80Hz, 60s)			
sedentary	<217	<386	
light	217-644	386-439	
moderate	645-1,810	440-2,098	
vigorous	>1,810	>2,098	
Rowlands <i>et al.</i> (VM_{gs} @100Hz, 15s)			
sedentary	<95	<95	
light	95-202	95-202	
moderate	203-566	203-566	
vigorous	>566	>566	
Hildebrand <i>et al.</i> (VM_{gs} @100Hz, 1s)			
sedentary			
light			<93.2
moderate			93.2-418.3
vigorous			>418.3

Table 6.2: Overview of intensity-scoring thresholds identified for the GENEActiv activity monitor. The table shows the lead author's name (method input marker, @sampling frequency, epoch length) and their identified threshold for sedentary, light, moderate and intense activities. $SV M_{gs}$ = gravity-adjusted summed vector magnitude (Equation 6.1). VM_{gs} = gravity-adjusted mean vector magnitude (Equation 6.4). Thresholds are divided into thresholds for the left wrist, right wrist and non-dominant wrist, according to the authors' choice.

The authors determined the mean arm angle ($angle_{mean}$) of a subject via the force in the y -direction of the accelerometer as shown in Equation 6.3 and Figure 6.1. If $angle_{mean}$ is smaller than 15° below the horizontal axis and the $SV M_{gs}$ during an epoch suggests moderate-to-vigorous PA, a participant is deemed to be standing. In all other cases they are classified as either sitting or lying.

$$angle_{mean} = \arcsin(y_{mean}) \quad (6.3)$$

Rowlands *et al.* evaluated the ability of their method to detect body position in a total of 48 subjects in laboratory, free-living or hospital environments and validated it against the thigh-worn activPal inclinometer. The method achieved an intra-individual

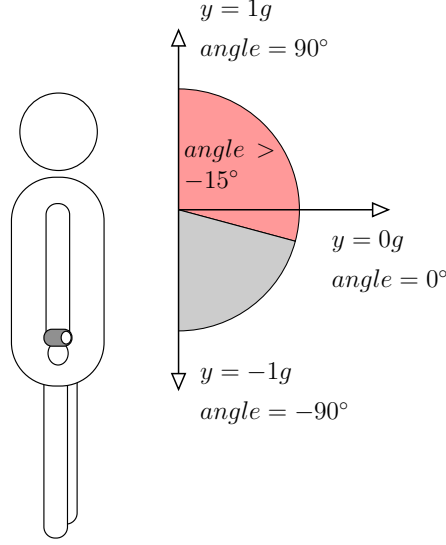


Figure 6.1: Visualisation of the body posture information used as part of Rowland *et al.*'s sedentary sphere. A measured y -axis acceleration of $-1g$ translates into an angle of -90° , a measured y -axis acceleration of $1g$ into an angle of 90° . The red shaded area corresponds to angles of -15° to 90° (where 0° corresponds to the horizontal), which are classified as sedentary or lying positions in Rowland's method. The grey shaded area corresponds to angles from -90° to -15° , which are classified as standing positions if they occur in combination with moderate-to-intense PA determined by mean acceleration thresholds.

agreement of 85%; however posture classification was erroneous for certain activities, such as reading whilst standing and cooking whilst standing.

Another intensity-scoring algorithm for the GENEActiv was derived by Hildebrand *et al.* [141]. The authors asked 30 subjects to wear the GENEActiv accelerometer on their non-dominant wrist and conduct activities of varying intensities. The individual's energy expenditure was monitored via a portable metabolic gas analyser. Hildebrand *et al.* used gravity-adjusted mean vector magnitude (VM_{gs}) (Equation 6.4) of acceleration as the input marker for their classifier.

$$VM_{gs} = \frac{1}{N_{meas}} \sum_{n=1}^{N_{meas}} |\sqrt{x_n^2 + y_n^2 + z_n^2} - g|, \quad (6.4)$$

where N_{meas} is the number of measurements per epoch, x , y , and z refer to the ac-

celerometer axes and g is the gravitational force of the earth, $1g$. Their linear classifier describing the relationship between VM_{gs} and oxygen consumption is given in Equation 6.5.

$$VO_2 = 0.0323mg + 7.49 * VM_{gs} \quad (6.5)$$

In addition to the linear classifier, the authors also defined thresholds for light, moderate and vigorous PA. These are listed in Table 6.2.

Zhang *et al.* [142] moved away from using just a single marker to detect different intensities of activities and instead proposed a range of frequency and time-domain feature extracted using both the Fast Fourier Transform (FFT) and the Discrete Wavelet Transform (DWT) as inputs to different machine learning classifiers. The FFT and the DWT are both methods to extract the frequency content from a time-series signal. In comparison to the FFT, the DWT is able to capture both frequency and time information by using varying lengths of windows across which it extracts the frequency information. Zhang *et al.* recruited 60 subjects and asked them to wear the GENEActiv wrist-band whilst conducting activities of different intensities. It was found that a decision tree performed best when classifying the different activities with an overall accuracy of 96% or above depending on wearing location.

The methods described by Esliger *et al.*, Rowlands *et al.* and Zhang *et al.* were developed for the GENEActiv device. In contrast, many algorithms for other devices are based on proprietary count values. One exception is the work conducted by Bonomi *et al.* [143]. They developed an acceleration-based classification algorithm for the detection of 20 different activities including, amongst others, sitting, lying and walking activities using the TracmorD device (Philips New Wellness Solutions, Lifestyle Incubator, Netherlands). The device is a wrist-worn tri-axial accelerometer with a battery life of three weeks and storage capabilities for up to 22 days of recording. Bonomi *et*

al. trained decision trees with varying lengths of input epochs and found that optimal performance (accuracy of 93%) was achieved with epoch sizes of 6.4s or 12.8s.

Similarly, Ellis *et al.* [144] developed another machine-learning approach to classify household, stairs, walking and running activities using raw accelerometer data from a wrist-worn Actigraph GT3x+ monitor (ActiGraph, LLC., Pensacola, Florida, U.S.A.). The Actigraph GT3x is device which contains a triaxial accelerometer and can be set to operate at various sampling frequencies and epoch lengths. The ground truth (in this scenario defined as the true energy consumption) used to train the random forest classifier was a portable indirect calorimeter, a device that measures oxygen uptake. Ellis *et al.* developed and tested the algorithm on 40 individuals, achieving an accuracy of 87.5%.

In addition to providing signals to methods to identify different types of activities and energy expenditure, accelerometers have also been used to detect sleep status. The algorithm used for processing the raw accelerometer data into minute-by-minute sleep labels was the method developed by Borazio *et al.* [116] as introduced in the previous chapter.

6.3 Methodology

6.3.1 Study Population

Data for this analysis were drawn from the UK Biobank resource. The UK Biobank is a registered charity with the aim of improving prevention, diagnosis and treatment of life-threatening diseases [68]. Between 2006 and 2010, the charity recruited approximately 500,000 subjects aged 40-69 years from across the United Kingdom and collected a range of self-reported data, physiological measurements and biological samples.

Among the data collected were baseline demographic and socioeconomic information (including age, sex, ethnicity, alcohol status, smoking status, Townsend index (an area of residence based measure for social deprivation [145])) as well as self-reported PA data. Baseline information was reassessed at the first repeat assessment in 2012-2013 (subset of 20,000 participants) and at the imaging visits conducted 2014-present (subset of 5,000 participants).

Accelerometer data were collected for a subset of approximately 100,000 participants between May 2013 and 2015 [69]. UK Biobank participants were contacted via email and provided with an Axivity AX3 accelerometer, if they agreed to take part in the activity data collection. Participants were told to wear the three-axial accelerometer continuously on their dominant wrist for a period of seven days. The accelerometer recorded data at a frequency of 100 Hz over a dynamic range of $\pm 8g$ (gravities).

Long-term health-related information was linked to the UK Biobank from sources including the Office for National Statistics (Death and emigration information) and the Department of Health's Hospital Episode Statistics (HES). The latter contained hospital-admission data (up until March 2015) for English residents that were either NHS-funded or NHS-treated. Figure 6.2 shows a timeline summarising the different stages of the UK Biobank data collection and the data used as part of this analysis.

6.3.2 Data Preparation – Self-reported Physical Activity Data

At baseline assessment, patients filled in a touchscreen-based questionnaire providing information about their PA levels. Patients were asked to state how many days they were engaged in more than ten minutes of each, walking, moderate and vigorous activity in a typical week. Participants were then questioned for how many minutes they were engaged in each of the activities on a typical day. In the description of physical activities *carrying light loads*, *cycling at normal pace* were given as examples of moderate activities,

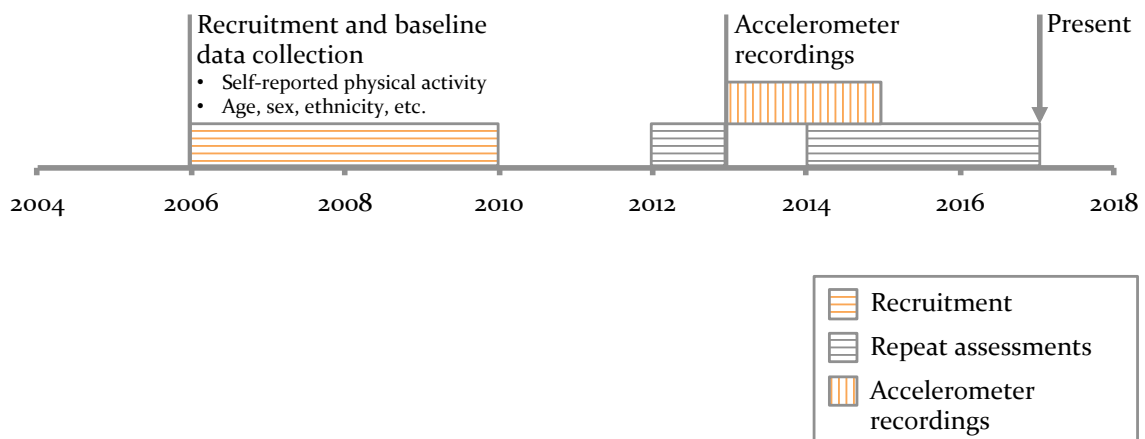


Figure 6.2: *Timeline of UK Biobank data collection showing time of baseline data collection, collection of repeat assessments (run on sub-groups of the original participant pool) (orange diamonds) and time of accelerometer data collection. Hospital statistics information is linked yearly using participants' HES records and is represented with a black dashed line.*

and activities that make you sweat or breathe hard such as fast cycling, aerobics, heavy lifting were given as examples of vigorous activities.

Self-reported PA data were processed using the method described by Bradbury *et al.* [146] based on the International Physical Activity Questionnaire guidelines [147]. Self-reported average daily PA minutes of <10 minutes were recorded as 0, values of $\geq 1,260$ minutes per week (equivalent to an average of 3 hours a day) were truncated at 1,260 minutes.

Walking, moderate and vigorous activities were scored at 2.3, 3.0 and 7.0 excess metabolic equivalents of tasks (METs) respectively [147]. In order to extract participants' average METs per week, the average daily time spent in each of the activities was multiplied by the average number of days doing the exercise and the respective MET scores.

6.3.3 Data Preparation – Accelerometer Data

The raw accelerometer data were calibrated [123] and wear-time periods were identified using the UK Biobank preprocessing methods described by Doherty *et al.* [69]. In order to minimise the computational cost of the UK Biobank analysis, all subsequent methods were computed using an epoch length of 30s. This is in accordance with the epoch length used for the published UK Biobank methods to date [69]. Figure 6.3 provides a visual overview of the accelerometer-marker extraction process. All markers are also listed in Table 6.3.

Summary statistics using the mean vector magnitude of acceleration (VM_{gs}) were extracted and averaged over the total number of wear-epochs of a participant. Summary statistics included the mean (x_1), standard deviation (x_2), median (x_3), 25th (x_4) and 75th percentile (x_5) of VM_{gs} extracted over 30s epochs. In addition to the 24-hour markers, the statistics were also calculated separately for wake times ($x_{W1} - x_{W5}$) and sleep times ($x_{S1} - x_{S5}$). The onset and end of sleep was thereby defined using the PDF-based sleep-boundary detection algorithm described in the previous chapter.

Derivative markers, such as the amount of time spent in different intensities of PA (extracted using the approach of Rowlands *et al.* and Hildebrand *et al.*), were also extracted over the 24-hour period and averaged over the total number of wear days of a participant. Using the algorithm-derived sleep-boundary times, the markers were also extracted separately for wake times. Markers describing the sleep duration and quality of sleep were extracted for sleep times. The latter was done using the algorithm developed by Borazio *et al.*, which assigns minute-by-minute sleep labels based on the standard deviation of the accelerometer signal along the z -axis (see section 5.2.1 for more detail).

Methods with epoch-length-dependent thresholds were adjusted to the new epoch

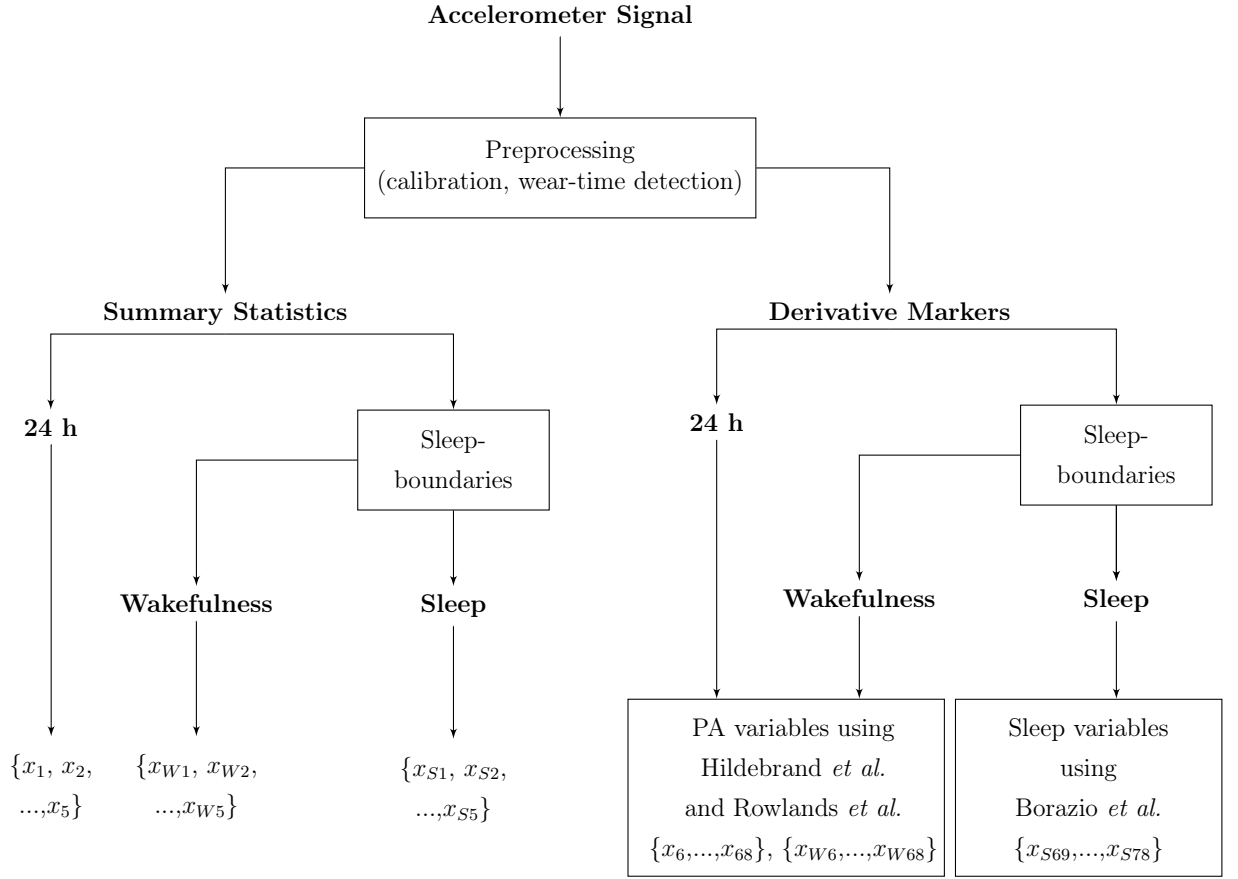


Figure 6.3: Flow diagram of the accelerometer feature extraction. Variables x_1 to x_{78} are defined in Table 6.3. All derivative markers were derived for 24-hour periods and then averaged over the number of wear-days of each participant.

length of 30s. Equation 6.6 shows the calculation used to transform the length of the smoothing filter in the sleep-boundary detection (N_{PDF}), equation 6.7 the formula used to adjust Borazio's idleness length (N_{bor}) (the duration for which sleep episodes are required to occur continuously to be classified as sleep).

$$N_{PDF} = N_{PDF} * \frac{30}{60} \quad (6.6)$$

$$N_{bor} = N_{bor} * \frac{30}{1} \quad (6.7)$$

Rowland's intensity thresholds were adjusted as shown in Equation 6.8.

$$thresh_{100Hz,30s} = thresh_{100Hz,15s} * \frac{30}{15} \quad (6.8)$$

Where appropriate, the average number of times a specific activity occurred during wake/ sleep as well as the mean duration for which it occurred were also included as markers. For this purpose a vector of length N_{rec} (where N_{rec} is the total number of epochs in the recording and n corresponds to an individual epoch) was created. Each element of this vector was set to one for every epoch in which the type of activity occurred, as shown for the example of light PA (based on Rowland's method) in Equation 6.9.

$$light_n = \begin{cases} 1 & \text{if } Rowlands_n = light \\ 0 & \text{otherwise} \end{cases} \quad (6.9)$$

The vector (e.g. $light_n$) was smoothed using a uniformly-weighted moving-average filter of five minutes in length in order to remove the effect of short-term changes in activity. The mean duration and mean number of times the activity occurred during a 24-hour period, wake period and sleep period were then stored as markers. Table 6.3 provides an overview of all accelerometer-derived markers included in the second part of the analysis described in this chapter.

Variable type	Variable name			Description
	24 hour	Wake	Sleep	
Statistic	x_1	x_{W1}	x_{S1}	Mean VM_{gs} averaged over recording period.
Statistic	x_2	x_{W2}	x_{S2}	Standard deviation of VM_{gs} averaged over recording period.
Statistic	x_3	x_{W3}	x_{S3}	Median of VM_{gs} averaged over recording period.
Statistic	x_4	x_{W4}	x_{S4}	25th percentile of VM_{gs} averaged over recording period.
Statistic	x_5	x_{W5}	x_{S5}	75th percentile of VM_{gs} averaged over recording period.
Derivative	x_6	x_{W6}		Daily MET score using the approach of Hildebrand <i>et al.</i>
Derivative	x_7	x_{W7}		Daily percentage of time spent in light PA (Hildebrand <i>et al.</i>)
Derivative	x_8	x_{W8}		Mean duration of light PA bouts (Hildebrand <i>et al.</i>)
Derivative	x_9	x_{W9}		Mean duration of light PA bouts (≥ 5 minutes) (Hildebrand <i>et al.</i>)
Derivative	x_{10}	x_{W10}		Mean duration of light PA bouts ≥ 10 minutes) (Hildebrand <i>et al.</i>)
Derivative	x_{11}	x_{W11}		Number of light PA bouts (Hildebrand <i>et al.</i>)
Derivative	x_{12}	x_{W12}		Number of light PA bouts (≥ 5 minutes) (Hildebrand <i>et al.</i>)
Derivative	x_{13}	x_{W13}		Number of light PA bouts (≥ 10 minutes) (Hildebrand <i>et al.</i>)
Derivative	x_{14}	x_{W14}		Daily percentage of time spent in moderate PA (Hildebrand <i>et al.</i>)
	...	...		...
Derivative	x_{20}	x_{W20}		Number of moderate PA bouts (≥ 10 minutes) (Hildebrand <i>et al.</i>)
Derivative	x_{21}	x_{W21}		Daily percentage of time spent in vigorous PA (Hildebrand <i>et al.</i>)
	...	...		...
Derivative	x_{27}	x_{W27}		Number of vigorous PA bouts (≥ 10 minutes) (Hildebrand <i>et al.</i>)
Derivative	x_{28}	x_{W28}		Daily percentage of time spent in sleeping PA (Rowlands <i>et al.</i>)
	...	...		...
Derivative	x_{33}	x_{W33}		Number of sleeping PA bouts (≥ 10 minutes) (Rowlands <i>et al.</i>)
Derivative	x_{34}	x_{W34}		Daily percentage of time spent in light PA (Rowlands <i>et al.</i>)
	...	...		...
Derivative	x_{40}	x_{W40}		Number of light PA bouts (≥ 10 minutes) (Rowlands <i>et al.</i>)
Derivative	x_{41}	x_{W41}		Daily percentage of time spent in moderate PA (Rowlands <i>et al.</i>)
	...	...		...
Derivative	x_{47}	x_{W47}		Number of moderate PA bouts (≥ 10 minutes) (Rowlands <i>et al.</i>)
Derivative	x_{48}	x_{W48}		Daily percentage of time spent in vigorous PA bouts (Rowlands <i>et al.</i>)
	...	...		...
Derivative	x_{54}	x_{W54}		Number of vigorous PA bouts (≥ 10 minutes) (Rowlands <i>et al.</i>)
Derivative	x_{55}	x_{W55}		Daily percentage of time spent in sedentary/lying PA (Rowlands <i>et al.</i>)
	...	...		...
Derivative	x_{61}	x_{W61}		Number of sedentary/lying PA bouts (≥ 10 minutes) (Rowlands <i>et al.</i>)
Derivative	x_{62}	x_{W62}		Daily percentage of time spent in standing PA (Rowlands <i>et al.</i>)
	...	...		...
Derivative	x_{68}	x_{W68}		Number of standing PA bouts (≥ 10 minutes) (Rowlands <i>et al.</i>)
Derivative			x_{S69}	Sleep duration according to sleep-boundary times
Derivative			x_{S70}	Sleep duration according to Borazio <i>et al.</i>
Derivative			x_{S71}	Sleep efficiency $\frac{bor_sd}{sb_sd}$
Derivative			x_{S72}	Daily percentage of time spent awake at night (Borazio <i>et al.</i>)
			...	...
Derivative			x_{S78}	Number of wake bouts during sleep (≥ 10 minutes) (Borazio <i>et al.</i>)

Table 6.3: List of summary statistics and descriptive markers included in the UK Biobank analysis. The dotted lines represent the different types of duration and frequency markers as shown for the example of light PA using the approach of Hildebrand *et al.* (x_8 to x_{14}).

6.3.4 Data Preparation – Hospital Admission Data

Hospital admission data were extracted from HES data. HF-related entries were defined as entries with ICD10 codes of 'I500', 'I501' or 'I509' or ICD9 codes of '4280' or '4281' according to the international classification of disease (ICD) code definitions. The time since first HF diagnosis (tfirst) was defined as time since first HES-recorded hospitalisation with a primary or secondary cause of HF.

6.3.5 Statistical Analysis

The aim of this analysis was to i) evaluate the ability of accelerometer-derived markers to differentiate between UK Biobank participants with and without HF, as well as to ii) study their added value in combination with evidence-based markers and in comparison against self-reported PA information. For this purpose two analyses were run: the first analysis (from this point onward referred to as '*baseline analysis*') utilised the UK Biobank data collected at baseline to evaluate the classification performance of established HF markers by themselves and in combination with self-reported PA data. Established HF markers were chosen based on the MAGGIC (Meta-Analysis Global Group in Chronic Heart Failure) score [148], a multivariate score developed for HF risk prediction. The score includes age, sex, ethnicity, alcohol status, smoking status, previous diagnosis of COPD and diabetes as markers in its prediction. Other markers such as creatinine levels and ejection fractions are also included in the MAGGIC score, however these were not available for this analysis.

The second analysis (from this point onward referred to as '*accelerometer analysis*') used the subset of UK Biobank participants with available accelerometer recordings to evaluate the classification performance of accelerometer-derived markers by themselves and in combination with established HF markers. Established markers were collected at baseline assessment with exception of disease status information and age, which were extracted for the time of accelerometer recording. Although there is an overlap between

participants in the first and second analysis, the time-difference of three to nine years between the collection of the baseline characteristics and accelerometer-recordings results in differences in the individuals with HF included in the analyses, as well as expected differences in the performance of established HF markers across the two datasets.

Baseline characteristics were presented as medians (inter-quartile range (IQR)) for continuous and as percentage for categorical markers, grouped by sex and HF status at time of baseline recording and accelerometer recording. In order to provide an initial measure of the difference in the values of markers between the HF and no-HF groups, two statistical tests were run. Continuous markers were tested using the Wilcoxon rank sum test, a statistical test suitable for non-normal distributions (such as the age distribution in this analysis). Markers displayed in categorical form were tested using the χ^2 test (see section 3.6 for more detail). *P*-values of 0.001 or smaller were considered to correspond to a significant difference between the HF and no-HF groups for this analysis.

For both analyses, HF participants were matched with age and sex-matched controls. For each HF participant, a control with the same sex and minimum difference in age was found. Controls were found without replacement, i.e. once a participant without HF had been identified as a match, they could not be matched to a second HF patient. If more than one individual without HF matched a HF patient equally well, one of them was selected by chance.

In order to evaluate the predictive value of accelerometer-derived markers, UK Biobank participants and their matched controls were randomly divided into a training and a test set following a 70:30 split (see Figure 6.4). The training set was used for the training of the classifier. The test set remained unseen until the final evaluation of the trained classifier. The process from training to evaluating the classifier was repeated 20 times with different 70:30 partitions of the dataset to provide a valid estimate of the performance of the markers.

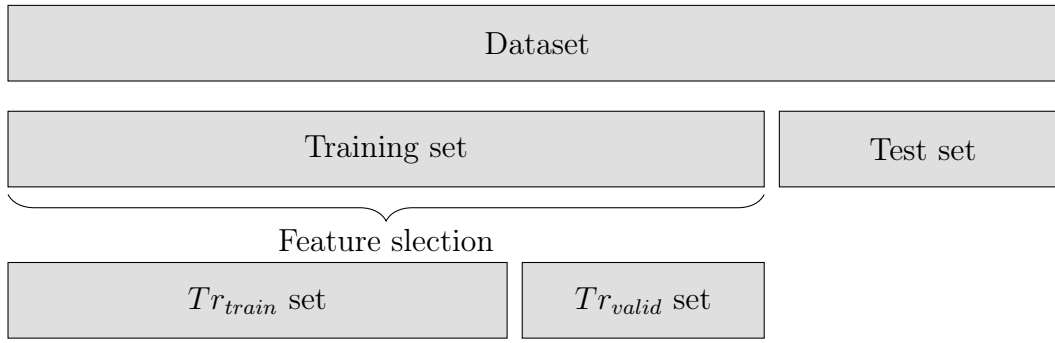


Figure 6.4: *Flow diagram summarising the training, validation and test sets. Ten-fold cross-validation was used to identify the best-performing subset of accelerometer-derived markers for heart failure classification. The derived classifier was then tested on the previously unseen testing set and compared to the evidence-based heart failure markers. This process was repeated 20 times.*

Including a large number of potentially correlated markers in a classifier can introduce noise and therefore worsen the performance of a classifier. Dimensionality reduction can be employed to reduce the number of markers included in a classifier. This can be achieved through feature selection or feature extraction. The prior approach identifies the subset of input markers with the individual or combined greatest predictive power. Feature extraction on the other hand transforms input features onto an optimised feature space. Whilst this method potentially results in better performance, it comes at the cost of interpretability. As the purpose of this analysis is to identify accelerometer-derived markers for eventual use in a clinical home monitoring system, interpretability is favoured.

Feature ranking is a commonly used method to optimise the features included in a classifier. Features can be ranked based on their individual performance (e.g. correlation, predictive value etc.) or the amount of added information they contain (mutual information ranking). The subset of N_{best} best-performing features can then be selected, where N_{best} is chosen using classifier-efficiency or performance restrictions [149].

For the purposes of this analysis, sequential forward selection (SFS) was used to identify the best-performing subset of accelerometer features. The method was applied during the training stage of the algorithm. For this purpose, the randomly selected training set was separated into a training (Tr_{train}) and a validation (Tr_{valid}) subset following a 70:30 split. The individual performance of each marker trained on the training set (Tr_{train}) was subsequently evaluated on the validation set (Tr_{valid}) and the area under the curve (AUC) of the receiver operating curve (ROC) recorded. The process of splitting the training set into subsets and recording the performance of each marker was repeated 10-fold in order to obtain reliable measures of performance. After the cross-validation, the marker with the best mean AUC on the validation set was selected as first marker to be included in the classifier. Using the same approach of performance evaluation, subsequent markers were added until any additional marker resulted in a decrease of performance on the validation set (Tr_{valid}), suggesting an overfitting on the training set (Tr_{train}). The selected features were then used to re-train the final classifier on the entire training set and evaluate it on the previously unseen test set.

Each classifier was evaluated based on its classification performance. The primary outcome measure used for the evaluation was the average test set AUC achieved by the classifier across the 20 random partitions of training and test sets. Accuracy, sensitivity (proportion of correctly-identified HF patients) and specificity (proportion of correctly-identified no-HF patients) were also recorded and confusion matrices created for each classifier.

6.4 Results

Out of a total of 502,602 UK Biobank participants, 390,047 provided complete self-reported PA information at baseline. This included 1,605 participants with previously diagnosed HF. Baseline characteristics of the UK Biobank population stratified by HF status are presented on the left hand side in Table 6.4.

Accelerometer-based PA data with good calibration and wear-performance were collected for 96,705 UK Biobank participants between 2013 and 2015, 596 of whom had previously been diagnosed with HF at the time of recording. Characteristics of the participants with available accelerometer data are shown in the right side of Table 6.4 and are based on data collected during baseline assessment with the exception of age and previous diagnosis of COPD or diabetes, which were based on the first day of accelerometer recording.

For both the self-reported and accelerometer-based populations, patients with HF were on average older than no-HF participants and more likely to be male. At baseline, participants with diagnosed HF in both study populations were less likely to be underweight and more likely to be overweight or obese than no-HF participants. HF participants were more likely to have given up alcohol and to class themselves as previous or current smokers in both analyses. There was a larger percentage of HF patients in the highest social deprivation class compared to no-HF participants. Nearly a third of HF patients had been previously diagnosed with diabetes at baseline in comparison to 5% of no-HF participants. In the UK Biobank sample with accelerometer recordings, 21% of HF patients had previously been diagnosed with diabetes versus only 2% of the participants without a HF diagnosis.

6.4.1 Baseline Analysis

Age and sex matched controls for UK Biobank HF patients were found by grouping participants with complete information (no missing values in the markers of interest) by sex and identifying the no-HF participants with minimum difference in age. If more than one participant with the same difference in age was found, one of them was selected by chance. The age distributions of participants in the baseline analysis before and after matching are shown in Figure 6.5. The median (interquartile range) ages were 63 (59-67)

	Self-reported PA			Accelerometer-derived PA		
	No HF	HF	P-value	No HF	HF	P-value
N	388,469	1,605		96,109	596	
Age (years)	57 (50-63)	63 (59-66)	<0.001	63 (56-69)	69 (65-72)	<0.001
Female	202,621 (52%)	324 (20%)	<0.001	54,294 (56%)	153 (26%)	<0.001
Male	185,848 (48%)	1,281 (80%)		41,815 (43%)	443 (74%)	
Ethnic background			0.750			0.557
White	367,556 (95%)	1,526 (96%)		92,803 (96%)	579 (98%)	
Mixed or non-white	18,917 (5%)	67 (4%)		2,969 (3%)	15 (2%)	
Missing	<1%	<1%		<1%	<1%	
BMI (kg/m ²)			<0.001			<0.001
<24.9	127,499 (33%)	259 (17%)		36,817 (38%)	120 (20%)	
≥ 25 and <30	1,165,916 (43%)	619 (39%)		39,489 (41%)	247 (42%)	
≥ 30 and <35	64,899 (17%)	467 (30%)		13,737 (14%)	144 (25%)	
≥ 35	24,213 (6%)	226 (14%)		4,810 (5%)	77 (13%)	
Missing	2%	2%		1%	1%	
Systolic BP (mmHg)	134 (120-147)	138 (125-151)	<0.001	137 (125-151)	140 (127-155)	
Missing	6%	9%		6%	7%	
Alcohol status			<0.001			<0.001
Never	14,891 (4%)	94 (6%)		2,790 (2%)	27 (5%)	
Previous	12,987 (3%)	156 (10%)		2,620 (2%)	48 (8%)	
Current	359,524 (93%)	1,349 (84%)		90,618 (94%)	521 (87%)	
Missing	<1%	<1%		<1%	0	
Smoking status			<0.001			<0.001
Never	212,498 (55%)	565 (35%)		54,840 (57%)	237 (40%)	
Previous	134,928 (35%)	842 (53%)		34,478 (35%)	307 (52%)	
Current	39,201 (10%)	189 (12%)		6,635 (6%)	50 (8%)	
Missing	<1%	<1%		<1%	<1%	
Socioeconomic status			0.058			0.863
Q1 (least deprived)	77,640 (20%)	213 (13%)		21,160 (22%)	110 (18%)	
Q2	77,615 (20%)	274 (17%)		20,362 (21%)	130 (22%)	
Q3	77,712 (20%)	288 (18%)		19,707 (20%)	121 (20%)	
Q4	77,583 (20%)	329 (21%)		18,952 (19%)	117 (20%)	
Q5	77,424 (20%)	498 (31%)		15,818 (16%)	118 (20%)	
Missing	<1%	<1%		<1%	2%	
Previous COPD	2,630 (1%)	109 (7%)	<0.001	283 (<1%)	26 (4%)	<0.001
Previous Diabetes	17,858 (5%)	459 (29%)	<0.001	2,739 (2%)	570 (21%)	<0.001

Table 6.4: Baseline information for UK Biobank participants. Quantile borders for the Townsend score Q1 - Q5 were $[-3.97, -2.43, -1.43, 1.55, 1.125]$ based on the participants with self-reported PA data. PA = physical activity. BP = Blood pressure. BMI = Body mass index. COPD = Chronic obstructive pulmonary disease. Age reported as median (25th, 75th percentile).

for males and 63 (58-66) for females. The corresponding median ages in the matched controls were 63 (59-67) and 63 (58-66) respectively, suggesting that the matching was successful.

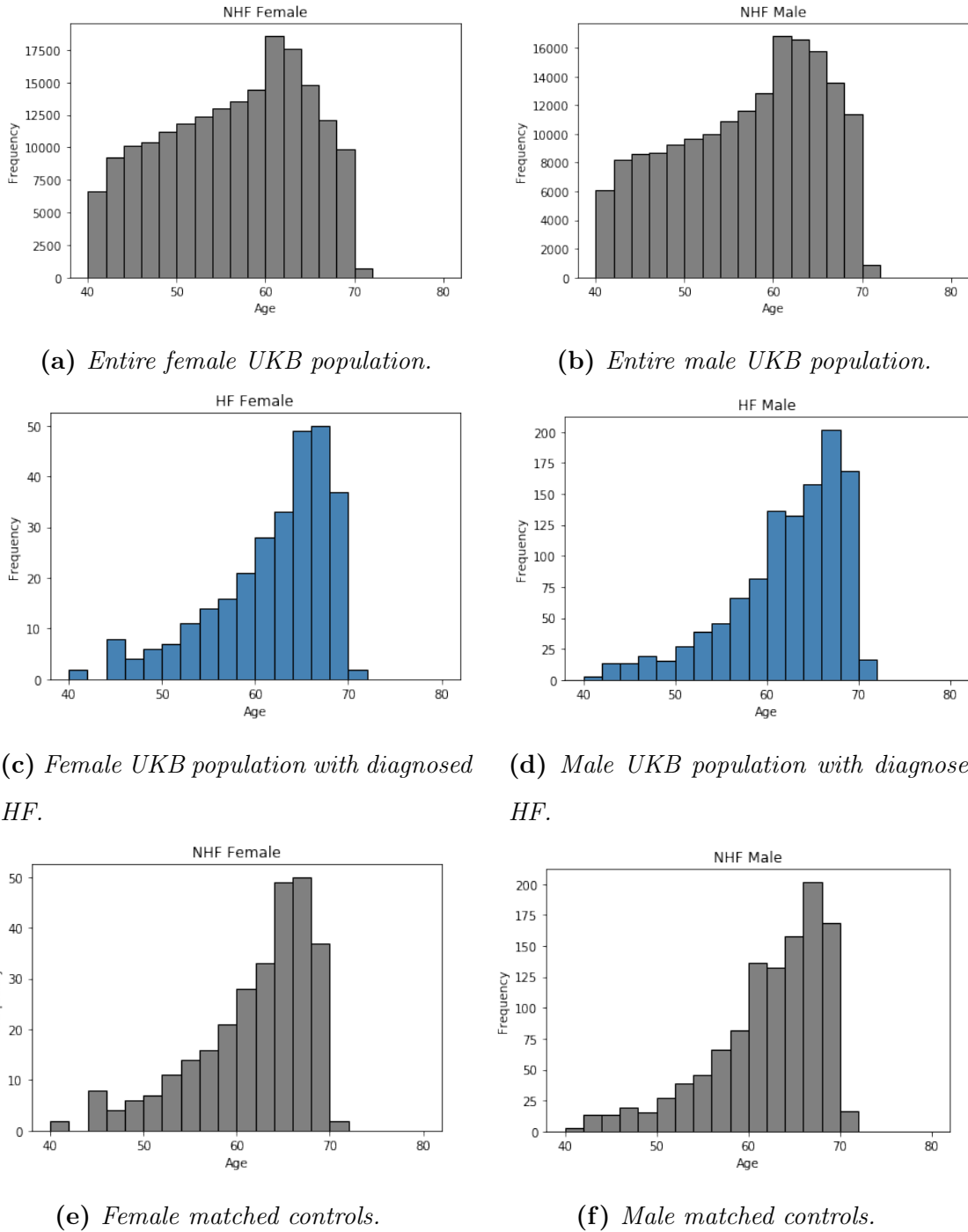


Figure 6.5: Histograms of age distributions by sex before and after control-matching in the baseline analysis.

The final *baseline analysis* was performed on HF participants with complete data and their matched controls (combined number of 2,848). The feature selection process for the *baseline analysis* classifier including only established markers is shown in Figure 6.6. The figure shows the average performance of the classifier across the 20 random partitions for both the Tr_{train} and the Tr_{valid} set on the vertical axis and the number of consecutively added markers on the horizontal axis. After the inclusion of four markers, the discrepancy between the mean AUCs of the two sets began to increase, suggesting that the classifiers started to overfit to the training sets at that point. For the final evaluation of each classifier on the test set of each of the 20 random partitions, the combination of input markers that achieved the best AUC on the training set was selected of each partition.

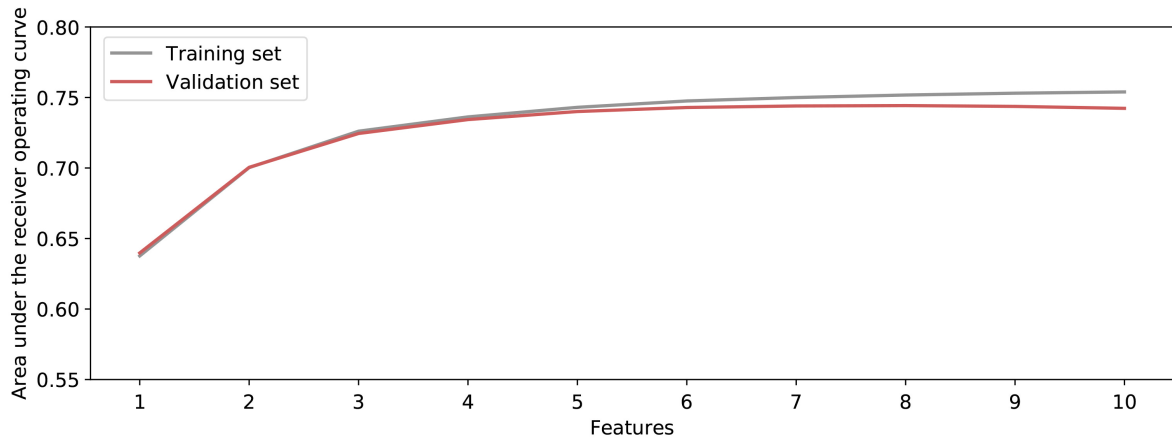


Figure 6.6: Performance of the baseline analysis classifier using only established markers as inputs. Performance was quantified as the mean area under the receiver operating curve (AUC) across 20 random partitions of training and test sets. The horizontal axis shows the number of cumulatively included markers at each stage of the feature selection process, starting with only a single feature on the left hand side of the figure.

Figure 6.7 shows the number of times each of the established markers was selected at different points of the feature selection process across the 20 random partitions. BMI followed by systolic blood pressure were selected most frequently as the first marker in the classifier, suggesting that these obtained the highest AUC score when taken individually.

Diabetes status was also among the markers selected early during the feature selection process.

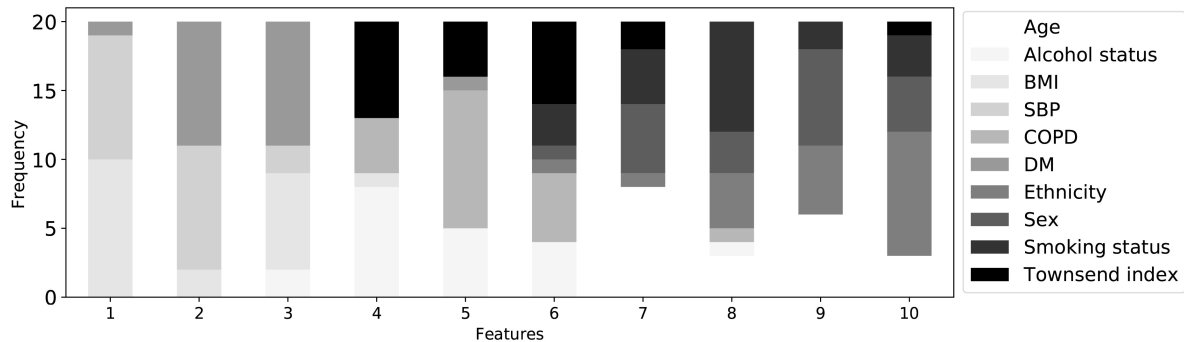


Figure 6.7: *Frequency and rank of markers selected for baseline analysis. BMI = body mass index. SBP = systolic blood pressure. COPD = chronic obstructive pulmonary disease. DM = diabetes mellitus. The marker most frequently selected as initial marker was BMI, followed by SBP and DM.*

The final performance of the *baseline analysis* classifier by itself and in combination with the self-reported PA information (in form of mean daily excess METs) on the previously unseen test sets is shown in Figures 6.8 and 6.9. The *baseline analysis* classifier using only established HF markers itself achieved an average AUC of 0.72. Adding the self-reported PA information did not improve the performance of the classifier (average AUC = 0.72). The PDFs in Figure 6.9 show that for both versions of the classifier there was a significant overlap between the HF and no-HF distributions, however some differentiation was possible.

Performance measures of the classifiers are summarised in Table 6.5. Both classifiers achieved accuracies of 67% and were better at correctly identifying participants without HF (specificities of 0.76 and 0.76) than at correctly identifying those with HF (sensitivities of 0.58 and 0.58), as is confirmed by the confusion matrices (see Tables 6.6 and 6.7).

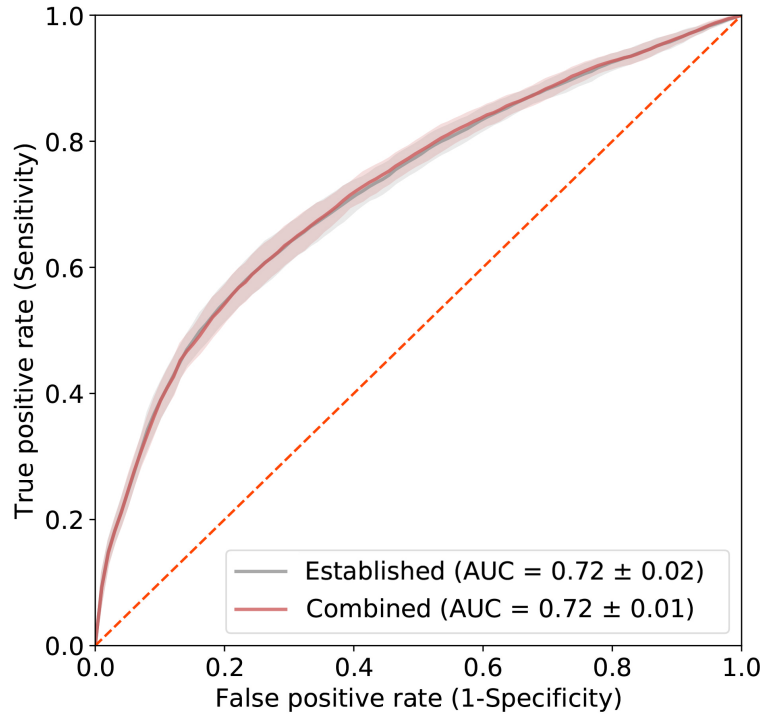
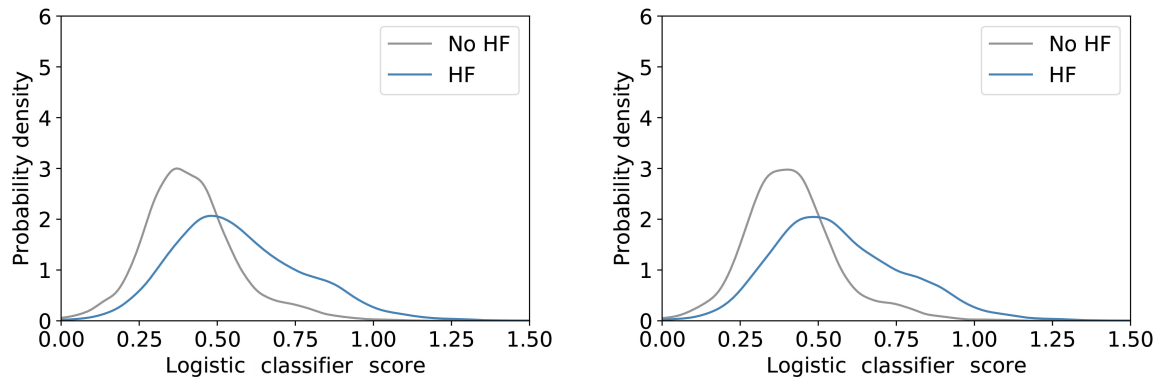


Figure 6.8: Receiver operating curves (ROC) showing the performance of the baseline analysis classifier using only established HF markers and including self-reported PA information. Sensitivities and specificities were calculated as the mean across ten iterations of randomly selected test sets. The shaded areas correspond to the mean $\pm$ one standard deviation for each set of markers.

	Accuracy	Sensitivity	Specificity
Established markers	0.67 (0.02)	0.58 (0.02)	0.76 (0.03)
Combined with self-reported PA	0.67 (0.02)	0.58 (0.02)	0.76 (0.03)

Table 6.5: Mean (standard deviation) of performance of the baseline analysis classifiers across the 20 random partitions of training and test sets. Sensitivity refers to the rate of correctly-identified HF patients, specificity to the rate of correctly-identified participants without HF.



(a) Classifier including only established markers. (b) Combined classifier with self-reported PA marker.

Figure 6.9: Probability density functions (PDFs) of logistic classification scores assigned to the test set. Solid line represents the mean values across 20 random partitions of training and test sets. The shaded area correspond to the standard deviations across the partitions.

		Ground truth		Total
		Positive (HF)	Negative (No HF)	
Classification	Positive (HF)	8266 29%	3317 12%	11583
	Negative (No HF)	5987 21%	10752 38%	16739
Total		14244	14069	N

Table 6.6: Confusion matrix for the classifications performance of the classifier based on established HF-markers evaluated on the test set (across 20 random partitions of training and test sets). Numbers represent the number of 30s-epochs in each category. Percentages represent percentage of 30s-epochs in each category versus overall number of 30s-epochs.

		Ground truth		Total
		Positive (HF)	Negative (No HF)	
Classification	Positive (HF)	8271 29%	3330 12%	11601
	Negative (No HF)	5973 21%	10739 38%	16712
Total		14244	14069	N

Table 6.7: Confusion matrix for combined activity and established marker classifications in the test set averaged across 20 random partitions of training and test sets. Numbers represent the number of 30s-epochs in each category. Percentages represent percentage of 30s-epochs in each category versus overall number of 30s-epochs.

6.4.2 Accelerometer Analysis

A total of 1,048 participants were included in the *accelerometer analysis*. The age distributions of the included HF participants and their matched controls at the time of accelerometer recording are shown in Figure 6.10. The median ages (IQR) in the HF group were 69 (66-72) for males and 68 (64-72) for females. The matched control groups had corresponding medians of 69 (66-72) and 68 (64-72) respectively, suggesting the age and sex-matching was successful.

Figures 6.11 and 6.11a illustrate the training processes of the classifiers including established HF markers (top) and accelerometer-derived features (bottom). In the first classifier, BMI achieved the strongest performance by itself across the 20 partitions of datasets, followed by diabetes status (at the time of accelerometer recording) and smoking status (at baseline) (see Figure 6.12a). It can be seen that the training of the classifier in this new sub-population of the UK Biobank resulted in a different selection of markers than in the *baseline analysis*. This may be because of the difference of up to nine years between the collection of some of the markers included in the classifier and the accelerometer recording, at the time for which the HF status was extracted.

The training process of the *accelerometer-based classifier* is shown in Figure 6.11b with the corresponding visualisation of the feature selection process in Figures 6.12b and 6.12c. It can be seen that there was a wide variability with respect to which markers were selected across the 20 random partitions of training and test sets, potentially reflecting the large correlation between many of the accelerometer-derived markers; however, light PA during wake time as defined by Hildebrand *et al.* (x_{13}) was selected most frequently as the initial feature (n=7 overall). Other frequently included markers were the duration of time spent in sedentary PA during wake hours as defined by Rowlands *et al.* (x_{W28} , $n = 10$), the duration of time spent in light PA for consecutive ten minutes or longer (x_{W7} , $n = 8$) and the number of sleep interruptions during the night that were of length

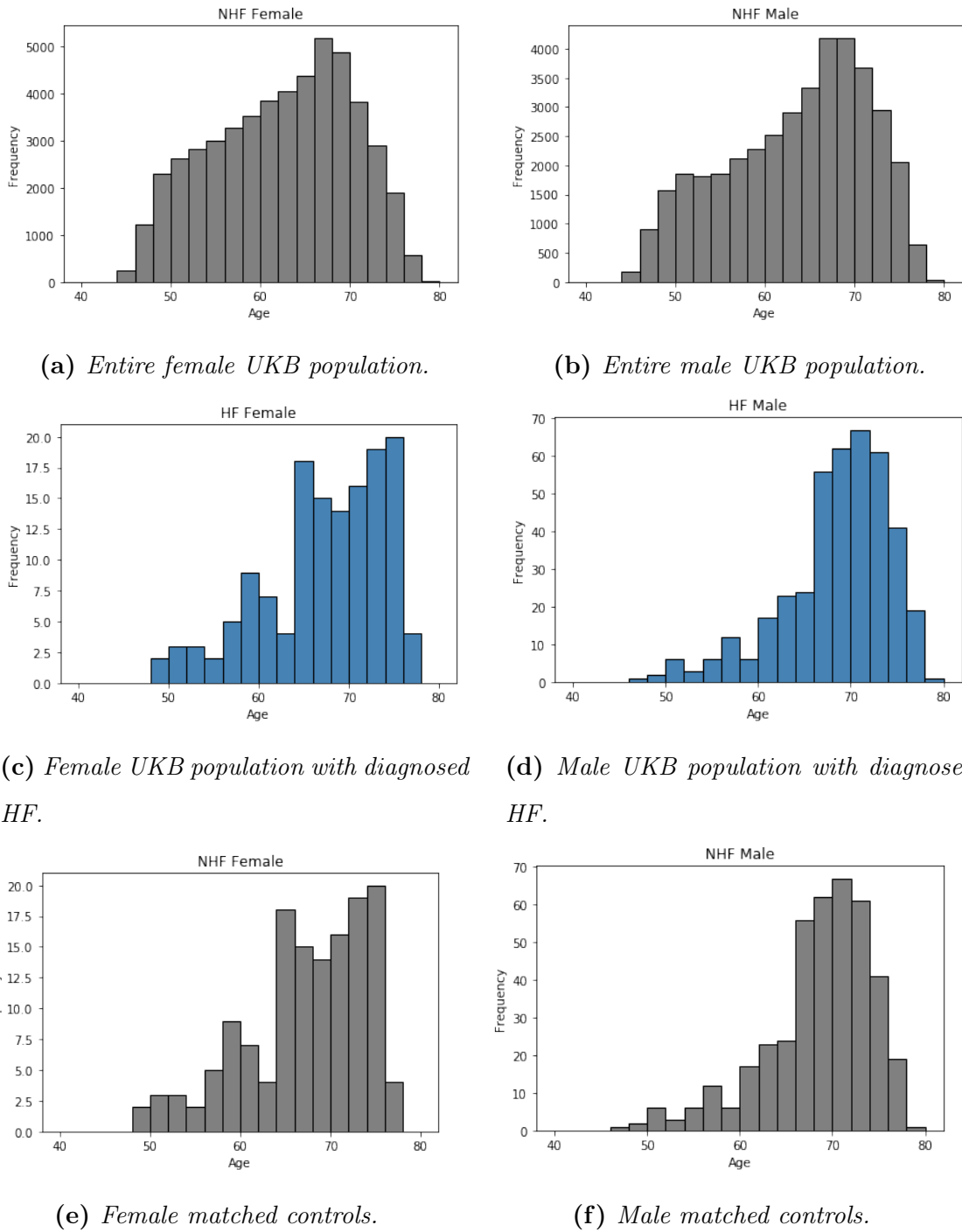
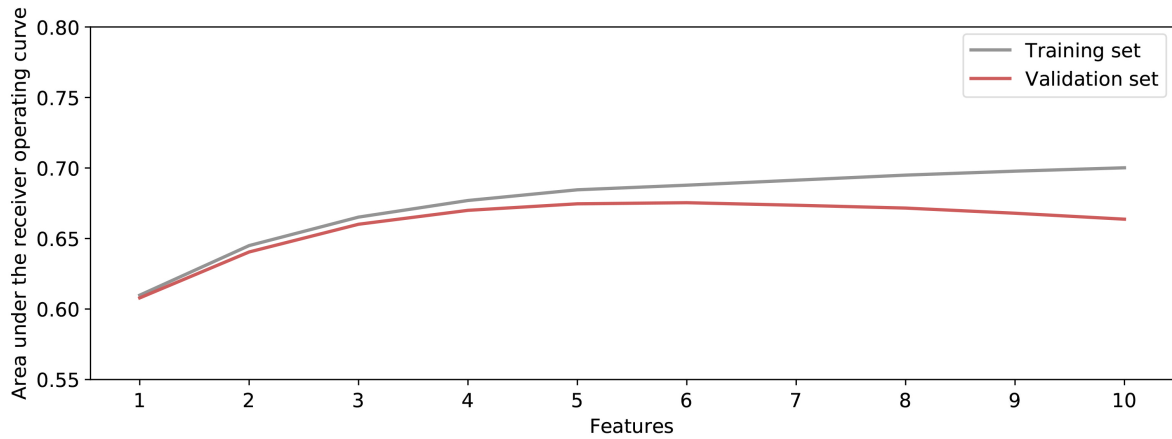
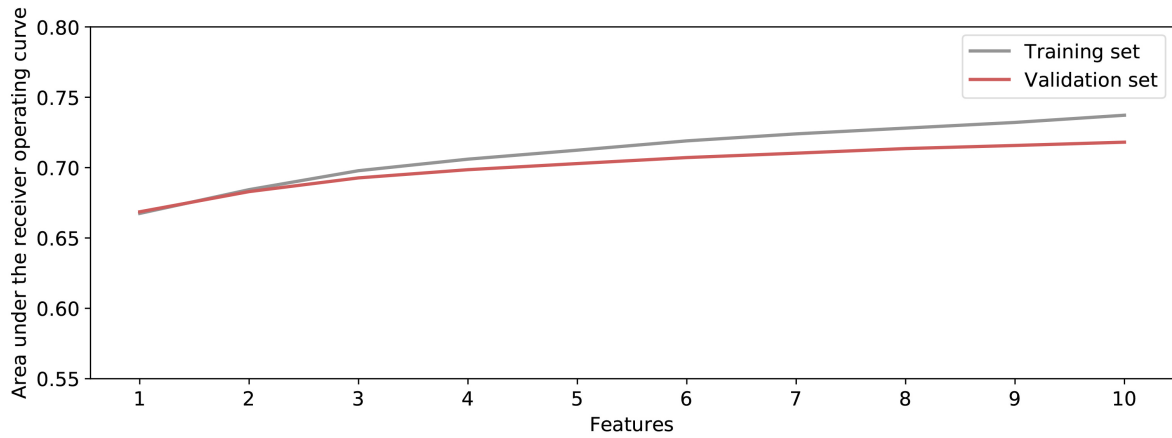


Figure 6.10: Histograms of age distributions by sex before and after control-matching in the accelerometer analysis.

five minutes or longer ($x_{S77,n} = 7$). In general, markers describing the duration and frequency of light and sedentary PA were more often selected than other markers.



(a) Classifier including only evidence-based markers.



(b) Classifier including only accelerometer-derived markers.

Figure 6.11: Performance of the accelerometer analysis classifier. Performance was measured by means of the area under the receiver operating curve (AUC). The horizontal axis shows the number of cumulatively included markers at each stage of the feature selection process, starting with only a single feature on the left hand side of the Figure.

The final performance of the individual and combined classifiers on the previously unseen test sets are shown in Figure 6.13. Despite the time-gap between the collection of some of the established HF markers and the accelerometer recording, the evidence-based markers were able to differentiate between UK Biobank participants with and without HF moderately well and achieved an AUC of 0.64. The accelerometer-derived markers by themselves achieved an AUC of 0.63. This showed that there is meaningful information regarding a subject's HF status contained in data collected from wrist-worn

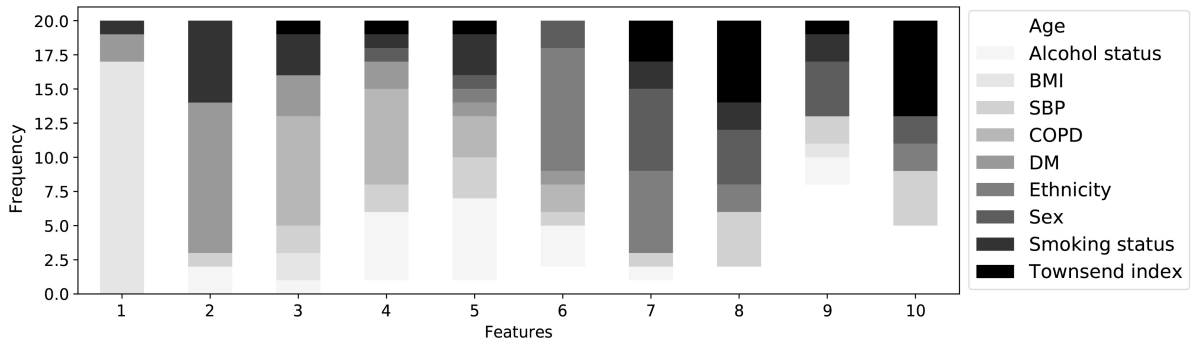
accelerometers. The combined classifier performed best ($AUC = 0.67$), suggesting that the accelerometer-derived markers provided complementary information with respect to that contained within the established HF markers.

The PDFs shown in Figure 6.14 provide further insight into the performance of the different HF classifiers. The PDF of the classifier using established markers presented in Figure 6.14a shows that the classification algorithm identified a subset of HF patients, corresponding to the second peak with a score of about 0.8; however a larger proportion of HF patients had scores lower than 0.5 and so were difficult to differentiate from the no-HF participants. The addition of the accelerometer-derived markers resulted in a larger proportion of HF patients being shifted beyond the 0.5 threshold of the classifier.

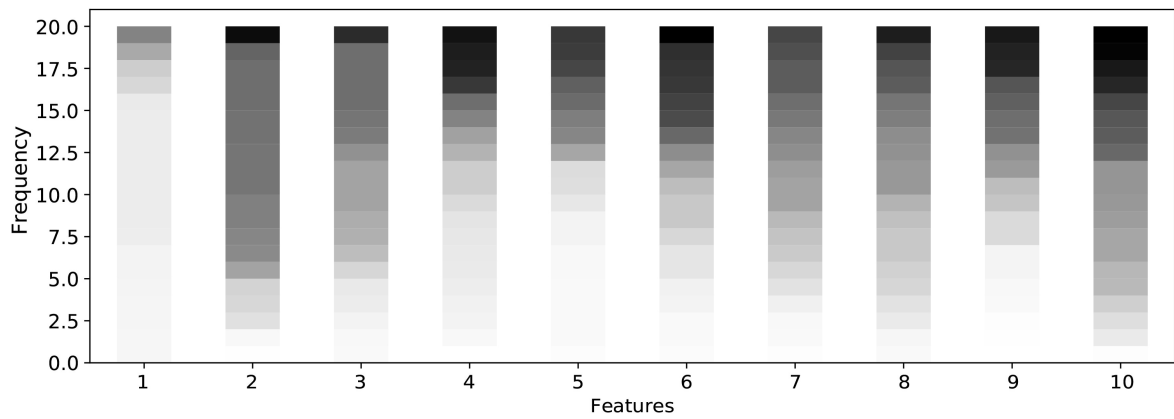
The performance measures for the three classifiers listed in Table 6.8 show that the classifier using combined markers outperformed the classifier using established markers with respect to their overall accuracy and sensitivity of identifying HF patients. The final accuracy (measured with a classification threshold of 0.5) of the combined classifier was 63%, with a sensitivity of 52% and a specificity of 71%. The overall relatively low sensitivity of the classifiers strengthens the impression that whilst the classifier is able to differentiate between a sub-group of HF patients and no-HF participants, it has difficulties identifying a proportion of HF patients correctly. The confusion matrices presented in Figures 6.9 to 6.11 provide further insight into the classification behaviour of the different classifiers.

	Accuracy	Sensitivity	Specificity
Established markers	0.61 (0.02)	0.42 (0.08)	0.79 (0.08)
Accelerometer-derived markers	0.59 (0.01)	0.56 (0.06)	0.62 (0.06)
Combined classifier	0.63 (0.02)	0.54 (0.06)	0.71 (0.05)

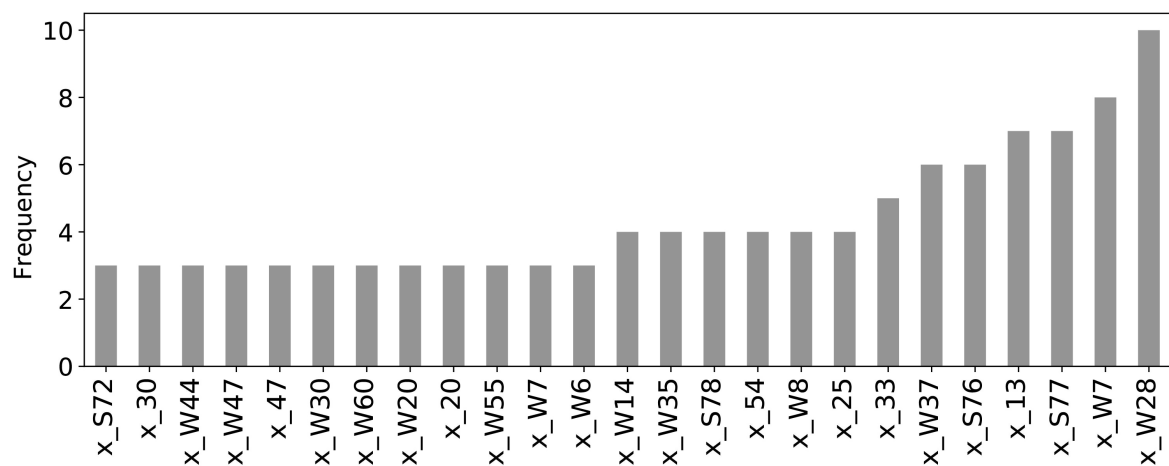
Table 6.8: Mean (standard deviation) of performance of the accelerometer analysis classifiers across the 20 random partitions of training and test sets. Sensitivity refers to the rate of correctly-identified HF patients, specificity to the rate of correctly-identified participants without HF.



(a) *Established markers. BMI = body mass index. SBP = systolic blood pressure. COPD = chronic obstructive pulmonary disease. DM = diabetes mellitus. The markers most frequently selected as initial marker was BMI, followed by DM status and smoking status.*



(b) *Accelerometer-derived markers. A wide range of markers were selected at each iteration. The figure below provides more detail on which markers were selected most frequently.*



(c) *Frequently-selected accelerometer-derived markers. The markers most frequently selected included light PA during wake time, the duration of time spent in sedentary PA during wake hours, the duration of time spent in light PA for consecutive ten minutes or longer and the number of sleep interruptions during the night that were of length five minutes or longer.*

Figure 6.12: *Frequency and rank of markers selected for accelerometer analysis.*

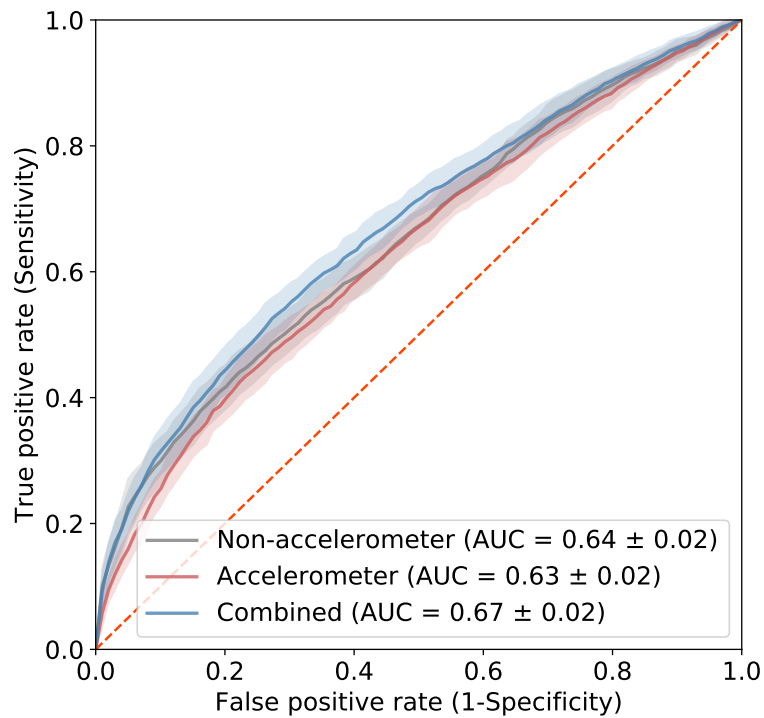


Figure 6.13: Receiver operating curves (ROC) showing the performance of the established markers classifier by itself (grey solid line), the accelerometer analysis classifier using only established markers (red solid line) and the combined classifier (blue solid line). The ROCs show the average performance of the classifiers on the test sets, over 20 random partitions of training and test sets. The shaded areas correspond to the mean $\pm$ one standard deviation for each set of markers.

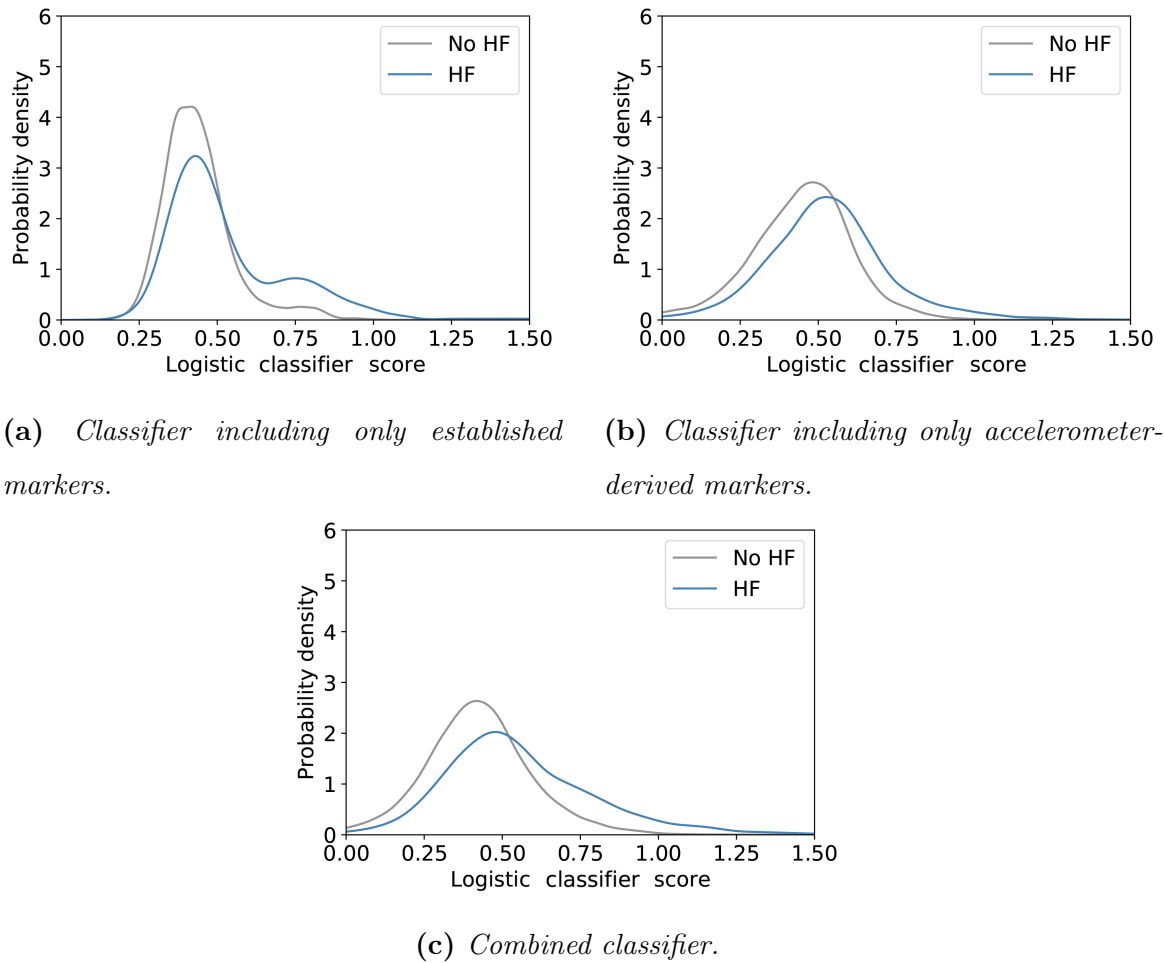


Figure 6.14: Probability density functions (PDFs) of logistic regression scores assigned to the test set. The PDFs include all test set datapoints across the 20 random partitions of training and test sets.

		Ground truth		Total
		Positive (HF)	Negative (No HF)	
Classification	Positive (HF)	2,144 21%	1,133 11%	3,277
	Negative (No HF)	2,976 29%	4,199 40%	7,175
Total		5,120	5,332	N

Table 6.9: Confusion matrix for the accelerometer analysis outcomes using only established markers across 20 random partitions of test sets.

		Ground truth		Total
		Positive (HF)	Negative (No HF)	
Classification	Positive (HF)	2,868 27%	2,011 19%	4,879
	Negative (No HF)	2,373 22%	3,321 32%	5,594
Total		5,120	5,332	N

Table 6.10: Confusion matrix for the accelerometer analysis outcomes using only accelerometer-derived markers across 20 random partitions of test sets.

		Ground truth		Total
		Positive (HF)	Negative (No HF)	
Classification	Positive (HF)	2,747 26%	1,551 15%	4,298
	Negative (No HF)	2,373 23%	3,781 36%	6,154
Total		5,120	5,332	N

Table 6.11: Confusion matrix for the accelerometer analysis outcomes using the combined markers across 20 random partitions of test sets.

6.5 Discussion

The aim of this chapter was to evaluate whether accelerometer-derived markers by themselves or in combination with established HF markers are able to differentiate between UK Biobank participants with and without HF, to study whether they do this better than self-reported PA information and to identify the accelerometer markers that are the most discriminatory and may therefore be suitable for implementation in future home monitoring solutions. The analysis, which was split into two parts because of the nature of the data available for this research, showed that the addition of self-reported PA data did not improve the classification performance of established PA markers. Accelerometer-derived markers on the other hand were able to identify HF patients moderately well and provided added information in addition to established markers. Accelerometer-derived derivative markers outperformed accelerometer-derived statistic summary markers in their ability to differentiate between subjects with and without HF. Markers describing the frequency and duration of different PA intensities provided the most meaningful information on HF status. In addition to this, meaningful information seemed to be contained in the number of sleep interruptions experienced by individuals.

Limitations of this research include the fact that accelerometer data provided by the UK Biobank may be affected by selection bias resulting in a healthier than normal HF population represented in the analysis described in this chapter. In addition to this, there may be a survival bias of participants included in this analysis, with patients being better-off simply because of the fact that they were still alive at the time of data collection. Both biases would lead to a sample of HF patients that is less differentiable from subjects without HF than the typical HF population, which could provide some explanation for the relatively large rate of false negatives in all classifiers.

In addition to this, the analyses of the accelerometer cohort may have been affected by the fact that baseline characteristics collected up to nine years ahead of the accelerometer

recording were included in the classifier. As the baseline analysis of the same markers showed better performance, it would be expected that the overall performance of the combined accelerometer and established marker classifier would improve with a more recent set of input markers.

6.6 Conclusion

Accelerometer-derived markers provide discriminating information, which allows differentiation between individuals with and without HF, information, which is additional with respect to established HF markers. Markers that perform best at identifying HF patients are derivative markers relating to the time spent in different intensities of PA, as well as the number of sleep-interruptions experienced by individuals. These markers may be suitable for implementation in future home monitoring systems.

Conclusion and Future Work

7.1 Repeated-Measure Risk Stratification using Traditional Heart Failure Markers

This thesis set out to explore markers for HF repeated-measure risk stratification. It started with an evaluation of measurements traditionally recorded in HF monitoring studies, namely weight, blood pressure, heart rate and NYHA symptom scores, and investigated their ability to differentiate between next days with and without HF-related adverse events, when using classifier-learned rather than accepted population-derived thresholds.

The evaluation of multi-variate risk prediction methods found that a logistic classifier trained on markers of three-day change in measurements outperformed the generic population-based thresholds used in the past, without giving major improvements. These findings are in agreement with previous findings from the literature, which have shown that individual weight, blood pressure, heart rate and NYHA changes do not provide reliable markers for repeated-measure risk stratification. As a result, additional markers are required.

The analysis of traditional HF markers was limited by the size of the SHF2 study

as well as the small number of HF-related adverse events that occurred during the study: 202 patients were recruited and only 38 patients experienced HF-related adverse events (26 of them with sufficient information to be included in the analysis). An even smaller number of patients experienced HF decompensation, a worsening of HF symptoms which would be expected to be the easiest phenomenon to identify using the recorded measurements. Future work on HF repeated-measure risk stratification should therefore focus on subgroups of patients for whom adverse events or certain types of adverse events are most likely to occur.

In addition to the population size, missing data affected the ability to extract meaningful information from the SHF2 data. Patients may miss measurements because of forgetfulness or non-compliance. Future research should look into the factors affecting missing values and how patients can be encouraged to develop the correct self-monitoring and self-management behaviours. One of the causes for missing values may be patients' inability to take measurements because of disease deterioration. The analysis described in this thesis did not include missing values as a marker to predict adverse outcomes. This is a shortcoming of the presented analysis. Future research should explore the information contained in missing values and include them as a potential marker for patient deterioration. Also, patient baseline characteristics, such as demographic information and mortality risk as assessed by scores such as the MAGGIC score should be considered as input variables for repeated-measure risk prediction models.

7.2 Heart Failure Classification using Wrist-worn Accelerometer Data

After the evaluation of traditional HF markers, the analysis described in this thesis explored the predictive value of new HF markers, in particular PA and sleep markers derived from wrist-worn accelerometers. The limited previous research on PA patterns

of HF patients had focused mainly on common derivative markers, such as step counts and total PA. The research described in this thesis went beyond this to look at which PA characteristics contain most information on HF status. For this purpose, data from the UK Biobank were used, which captured self-reported PA and seven-day accelerometer recordings for more than 600 individuals with HF and a much greater number of individuals without HF. The analysis of the UK Biobank data showed that data from a wrist-worn accelerometer can differentiate between individuals with and without HF and that such data contains additional information to that of established single-measure markers. It furthermore found that derivative markers describing the frequency and intensity of PA performed best at differentiating between the two groups.

Whilst the SHF2 data provided access to a large amount of data on a relatively small number of HF patients, the UK Biobank offered access to a larger number of HF patients, however at the cost of less detailed information on disease status and severity. As a result, the analysis of UK Biobank HF patients was restricted to an analysis of the ability of markers to differentiate between individuals with and without HF. Future work should focus on the collection of continuous or repeated accelerometer data for HF patients, and the trend analysis of these data in longitudinal studies. This may provide a greater insight into changes that occur prior to adverse events and thereby offer a tool for repeated-measure risk stratification.

As part of the SHF2 research conducted for this thesis, a small (n=13) workshop on HF patient interest in wearable technologies was conducted. This workshop showed that even older HF patients are aware of PA and sleep trackers and that they are willing to wear such devices for the purpose of health and wellbeing. Whilst the real test for wearable technologies will be longer-term compliance (as well as the added value of the collected data), it appears feasible that they form part of future home monitoring technologies for HF patients.

7.3 Automated Sleep-boundary Detection using Wrist-worn Accelerometers

Finally, the work conducted as part of this thesis included the development and evaluation of data-driven sleep-boundary detection methods for wrist-worn accelerometers. This work was inspired by the fact that current polysomnography-validated measures of sleep status from wrist-worn accelerometers cannot be used in fully automated analysis as they rely on self-reported sleep-onset and -end (sleep-boundary) information [66]. This information is typically collected through means of a sleep diary [66] or through a button on the accelerometer which participants can press as they go to bed and wake up [150]. As a result, it is not feasible to collect self-reported sleep-boundary times for all participants in large scale studies, such as the UK Biobank [69].

The sleep-boundary detection methods developed as part of this thesis performed well at identifying minute-by-minute sleep and wake times of individuals as defined by self-reported sleep-onset and wake-up times. In order to evaluate the methods' actual ability to perform minute-wise sleep scoring, additional validation against PSG would be required. The mean absolute error between self-reported and algorithm-derived sleep-boundary times ranged from 35.5 - 40.7 minutes. This is comparable to the agreement between different self-reported sources (as found through the comparison of different self-reported measures in the CAPTURE24 study), however further research on the levels of accuracy required for a heart failure population should be conducted and should inform potential future improvements of the methods (e.g. using additional sensors). The best performance was achieved by the random-forest-based method, which used a large array of input variables to differentiate between sleep and wake times. These were subsequently filtered to provide sleep-onset and wake-up times. The performance of the random forest method was closely followed by the performance of the much simpler PDF-based thresholding method. This method used a single, easy to extract input variable

(the standard deviation of the vector magnitude across the three accelerometer axes) and may therefore be more suitable for use in future accelerometer research.

All sleep-boundary detection methods were trained using self-reported sleep and wake information as their ground truth. Self-reported sleep times have been shown to be less accurate than objectively-collected sleep-onset and end-times [150] and may be seen to provide a weaker ground truth than polysomnography assessments. Collecting free-living rather than laboratory data does however provide a more realistic setting for sleep analysis, where recordings may be affected by periods of rest prior to sleep and after wake-up that may otherwise be missed. In addition to this, allowing participants to stay in their natural sleeping environment reduces the impact of data collection on their sleeping behaviour [151]. Nevertheless, future work on the sleep-boundary detection method should include a validation of this method using polysomnography data.

7.4 SUPPORT-HF3: Proposal for the Next Stage of the Study

The intervention evaluated in the SUPPORT-HF2 study relied on guideline-based treatment decisions guided by optimal treatment plans and daily assessments of a patient's risk of disease deterioration. Patients' risk of deterioration was assessed using population-based thresholds, which resulted in a large number of false alerts, particularly when a patient's normal measurements fell outside the normal population-based values. This in turn forced medical staff to rely on their clinical judgement rather than the data-driven alerts themselves. Similar findings of alerting systems with population-based thresholds leading to large numbers of false alerts and subsequent dismissal of the alerts by clinical staff were found in a recent retrospective study by Murphy *et al.* [109].

Based on this knowledge, future home monitoring studies could involve more specific

deterioration alerts enabling a reduction in the number of false alerts and subsequently an increase in both the objectiveness and consistency of treatment decisions. This could be achieved through more suitable markers of patient deterioration, e.g. the standard deviation of three days changes evaluated in this thesis or the repeated down-sampling markers proposed by Koulaouzidis *et al.* Another approach that may provide more specificity is the alerting system described by Velardo *et al.* [152], which established the normality of a measurement based on a patient's previous measurements and triggered an alert whenever a value lay outside the 95% percentile of the distribution of previous measurements. In either scenario, it will be important for alerts to occur with sufficient lead time to allow for appropriate intervention.

The inclusion of new, objective markers may furthermore improve the performance of deterioration alerts. These could include the occurrence of missing values and long-term risk scores, such as the MAGGIC score. In addition to this, as shown through the assessment of the UK Biobank data in this thesis, PA and sleep data collected from wrist-worn accelerometers contain valuable information on HF status and should be further evaluated in the next SUPPORT-HF study. Ideally, this would involve the continuous collection of raw accelerometer data to gain a better understanding of time-series changes ahead of deterioration. Should this not be possible, weekly day-time assessments of the amount of time spent in different intensities of activities may provide a useful insight into patient PA trends over time. Sleep quality and fragmentation of sleep can be assessed using accelerometers, however other technologies such as mattress-embedded sensors could provide potential alternatives depending on feasibility (e.g. battery power and storage capability of devices) as well as patient preference.

Whilst improved HF markers will likely have some impact on the effectiveness of HF monitoring systems, previous evidence suggests that predicting adverse events ('crisis prediction') will remain a challenging task [153]. The more dynamic approach of 'health maintenance' has therefore been proposed as a potential alternative to 'crisis prediction'.

This approach aims to retain patients in a safe state by adjusting treatments on a regular, potentially even daily, basis. Based on the normality of collected measurements, patients could either take on the responsibility of adjusting their treatments themselves or obtain treatment instructions from their health professional. Cleland *et al.* [153] state that 'health maintenance' may even increase patient engagement with tele-monitoring systems due to its more dynamic nature in comparison to traditional HF monitoring approaches. This in turn may help lead to better and more consistent performances of HF home monitoring systems.

Previous reviews of home monitoring systems for HF patients have shown varying performances, which may at least partly be explained by differences study population and the intensity/ involvement of interventions. Keeping patients engaged in their home monitoring system may increase their involvement with their care and consequently lead to better long-term outcomes.

The combined monitoring system, including personalised alerts/ treatment adjustment instructions should be appropriately evaluated [154]. In its initial phases, the focus of the evaluation should lie on ease of use and the technical feasibility of the system. In the later phases, emphasis should lie on the clinical and economic benefit of the system. At this point, system evaluation should be done in form of a randomised controlled study. Similarly to the SUPPORT-HF2 study, patients in the control group could be provided with the SUPPORT-HF system without the personalised alerts and feedback, whilst patients in the intervention group could receive access to the system in addition to alerts and feedback. A longer monitoring duration of at least twelve months would be recommended to increase the likelihood of patients experiencing disease deterioration during the study duration. If the alerts and feedback were found to be effective in reducing patients' likelihood of adverse events before the end of the monitoring periods, it could be considered to move all patients to the intervention arm for ethical reasons.

References

- [1] N. Conrad, A. Judge, J. Tran, H. Mohseni, D. Hedgecote, A. Perez Crespillo, M. Allison, H. Hemingway, J. Cleland, J. McMurray, and K. Rahimi, “Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals,” *The Lancet*, 2017.
- [2] C. Taylor, R. Ryan, L. Nichols, N. Gale, F. Hobbs, and T. Marshall, “Survival following a diagnosis of heart failure in primary care,” *Family Practice*, vol. 34, pp. 161–168, Apr. 2017.
- [3] NICE, “CG108 Chronic heart failure: Management of chronic heart failure in adults in primary and secondary care,” 2010.
- [4] A. Louis, T. Turner, M. Gretton, A. Baksh, and J. Cleland, “A systematic review of telemonitoring for the management of heart failure,” *European Journal of Heart Failure*, vol. 5, pp. 583–590, Oct. 2003.
- [5] A. Desai and M. Konstam, “Home Monitoring Heart Failure Care Does Not Improve Patient Outcomes Looking Beyond Telephone-Based Disease Management,” *Circulation*, vol. 125, pp. 828–836, Feb. 2012.
- [6] E. Benjamin, M. Blaha, S. Chiuve, M. Cushman, S. Das, R. Deo, S. d. Ferranti, J. Floyd, and O. Fornage, Myriam et al., “Heart Disease and Stroke Statistics—2017 Update: A Report From the American Heart Association,” *Circulation*, Jan. 2017.
- [7] S. Stewart, A. Jenkins, S. Buchan, A. McGuire, S. Capewell, and J. McMurray, “The current cost of heart failure to the National Health Service in the UK,” *European Journal of Heart Failure*, vol. 4, pp. 361–371, June 2002.
- [8] K. Smolina, F. L. Wright, M. Rayner, and M. J. Goldacre, “Determinants of the decline in mortality from acute myocardial infarction in England between 2002 and 2010: linked national database study,” *BMJ*, vol. 344, Jan. 2012.
- [9] P. Ponikowski, A. Voors, S. Anker, H. Bueno, J. Cleland, and A. e. Coats, “2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure,” *European Journal of Heart Failure*, vol. 18, pp. 891–975, Aug. 2016.
- [10] J. Cleland, H. Dargie, S. Hardman, T. McDonagh, and P. Mitchell, “National Heart Failure Audit,” tech. rep., National Institute for Cardiovascular Outcomes Research (ICOR), 2013.
- [11] M. Burch, “Heart failure in the young,” *Heart*, vol. 88, pp. 198–202, Aug. 2002.

- [12] C. Wong, N. Hawkins, M. Petrie, P. Jhund, R. Gardner, C. Ariti, and et al., “Heart failure in younger patients: the Meta-analysis Global Group in Chronic Heart Failure (MAGGIC),” *European Heart Journal*, vol. 35, pp. 2714–2721, Oct. 2014.
- [13] Medical University of South Carolina, “Cardiology subspecialties, heart failure reduced ejection fraction figure.”
- [14] T. McDonagh, R. Gardner, A. Clark, and H. Dargie, *Oxford Textbook of Heart Failure*. Oxford University Press Oxford, July 2011.
- [15] The Consensus Trial Study Group, “Effects of Enalapril on Mortality in Severe Congestive Heart Failure,” *New England Journal of Medicine*, vol. 316, pp. 1429–1435, June 1987.
- [16] S. Chatterjee, G. Biondi-Zoccai, A. Abbate, F. D’Ascenzo, D. Castagno, B. V. Tassell, D. Mukherjee, and E. Lichstein, “Benefits of β blockers in patients with heart failure and reduced ejection fraction: network meta-analysis,” *BMJ*, vol. 346, p. f55, Jan. 2013.
- [17] G. Oghlakian, I. Sipahi, and J. Fang, “Treatment of Heart Failure With Preserved Ejection Fraction: Have We Been Pursuing the Wrong Paradigm?,” *Mayo Clinic Proceedings*, vol. 86, pp. 531–539, June 2011.
- [18] A. Maggioni, U. Dahlström, G. Filippatos, O. Chioncel, M. C. Leiro, and J. e. Drozd, “EURObservational Research Programme: regional differences and 1-year follow-up results of the Heart Failure Pilot Survey (ESC-HF Pilot),” *European Journal of Heart Failure*, vol. 15, pp. 808–817, July 2013.
- [19] F. Pons, J. Lupón, A. Urrutia, B. González, E. Crespo, and C. e. Díez, “Mortality and Cause of Death in Patients With Heart Failure: Findings at a Specialist Multidisciplinary Heart Failure Unit,” *The Lancet*, vol. 63, pp. 303–314, Jan. 2010.
- [20] K. Rahimi, D. Bennett, N. Conrad, T. Williams, J. Basu, J. Dwight, M. Woodward, A. Patel, J. McMurray, and S. MacMahon, “Risk Prediction in Patients with Heart Failure,” *JACC: Heart Failure*, vol. 2, pp. 440–446, Apr. 2014.
- [21] P. Peterson, J. Rumsfeld, L. Liang, N. Albert, and A. e. Hernandez, “A Validated Risk Score for In-Hospital Mortality in Patients With Heart Failure From the American Heart Association Get With the Guidelines Program,” *Circulation: Cardiovascular Quality and Outcomes*, vol. 3, pp. 25–32, Jan. 2010.
- [22] S. Pocock, C. Ariti, J. McMurray, A. Maggioni, L. Køber, and I. e. Squire, “Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies,” *European Heart Journal*, vol. 34, pp. 1404–1413, May 2013.
- [23] U. Sartipy, U. Dahlstroehm, M. Edner, and L. Lund, “Predicting survival in heart failure: validation of the maggic heart failure risk score in 51,043 patients from the swedish heart failure registry,” *European Journal of Heart Failure*, vol. 16, 2013.
- [24] A. Bui, T. Horwich, and G. Fonarow, “Epidemiology and risk profile of heart failure,” *Nature Reviews Cardiology*, vol. 8, pp. 30–41, Jan. 2011.

- [25] M. Lainscak, L. Blue, A. Clark, U. Dahlström, K. Dickstein, and I. e. Ekman, "Self-care management of heart failure: practical recommendations from the Patient Care Committee of the Heart Failure Association of the European Society of Cardiology," *European Journal of Heart Failure*, vol. 13, pp. 115–126, Feb. 2011.
- [26] J. Zhang, K. Goode, P. Cuddihy, and J. o. Cleland, "Predicting hospitalization due to worsening heart failure using daily weight measurement: analysis of the Trans-European Network-Home-Care Management System (TEN-HMS) study," *European Journal of Heart Failure*, vol. 11, pp. 420–427, Apr. 2009.
- [27] J. Lewin, M. Ledwidge, C. O'Loughlin, C. McNally, and K. McDonald, "Clinical deterioration in established heart failure: What is the value of BNP and weight gain in aiding diagnosis?," *The European Journal of Heart Failure*, vol. 7, pp. 953–957, Oct. 2005.
- [28] M. Lan, L. Samy, N. Alshurafa, M.-K. Suh, H. Ghasemzadeh, A. Macabasco-O'Connell, and M. Sarrafzadeh, "WANDA: An End-to-End Remote Health Monitoring and Analytics System for Heart Failure Patients," *Preventive Medicine*, 2012.
- [29] N. Alnosayan, E. Lee, A. Alluhaidan, S. Chatterjee, L. Houston-Feenstra, M. Kagoda, and W. Dysinger, "MyHeart: An intelligent mHealth home monitoring system supporting heart failure self-care," in *2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pp. 311–316, Oct. 2014.
- [30] S. Inglis, R. Clark, F. McAlister, S. Stewart, and J. Cleland, "Which components of heart failure programmes are effective? A systematic review and meta-analysis of the outcomes of structured telephone support or telemonitoring as the primary component of chronic heart failure management in 8323 patients: Abridged Cochrane Review," *European Journal of Heart Failure*, vol. 13, pp. 1028–1040, Sept. 2011.
- [31] The Criteria Committee of the New York Heart Association, "Nomenclature and Criteria for Diagnosis of Diseases of the Heart and Great Vessels.," pp. 253–256, Boston, Mass: Little, Brown & Co, 9 ed., 1994.
- [32] W. Abraham, S. Compton, G. Haas, B. Foreman, R. Canby, R. Fishel, S. McRae, G. Toledo, S. Sarkar, D. Hettrick, and FAST Study Investigators, "Intrathoracic impedance vs daily weight monitoring for predicting worsening heart failure events: results of the Fluid Accumulation Status Trial (FAST)," *Congestive Heart Failure (Greenwich, Conn.)*, vol. 17, pp. 51–55, Apr. 2011.
- [33] P. Adamson, A. Smith, W. Abraham, K. Kleckner, and R. e. Stadler, "Continuous Autonomic Assessment in Patients With Symptomatic Heart Failure Prognostic Value of Heart Rate Variability Measured by an Implanted Cardiac Resynchronization Device," *Circulation*, vol. 110, pp. 2389–2394, Oct. 2004.
- [34] W. Abraham, P. Adamson, R. Bourge, M. Aaron, M. Costanzo, and L. e. Stevenson, "Wireless pulmonary artery haemodynamic monitoring in chronic heart failure: a randomised controlled trial," *The Lancet*, vol. 377, pp. 658–666, Feb. 2011.

- [35] E. Kraigher-Krainer, A. Lyass, J. Massaro, D. Lee, J. Ho, D. Levy, W. Kannel, and R. Vasan, "Association of physical activity and heart failure with preserved vs. reduced ejection fraction in the elderly: the Framingham Heart Study," *European Journal of Heart Failure*, vol. 15, pp. 742–746, July 2013.
- [36] C. Chessum, J. Norman, B. Pozehl, K. Duncan, and M. Hertzog, "Effects of sleep and depression on physical activity performance in individuals with heart failure," *Cardiopulmonary Physical Therapy Journal*, vol. 25, p. 121, Dec. 2014.
- [37] V. M. Conraads, M. A. Spruit, F. Braunschweig, M. R. Cowie, L. Tavazzi, M. Borggrefe, M. R. S. Hill, S. Jacobs, B. Gerritse, and D. J. van Veldhuisen, "Physical activity measured with implanted devices predicts patient outcome in chronic heart failure," *Circulation. Heart Failure*, vol. 7, pp. 279–287, Mar. 2014.
- [38] M. Dontje, M. van der Wal, R. Stolk, J. Brügemann, T. Jaarsma, P. Wijnvliet, C. van der Schans, and M. de Greef, "Daily Physical Activity in Stable Heart Failure Patients," *Journal of Cardiovascular Nursing*, vol. 29, no. 3, pp. 218–226, 2014.
- [39] F. Braunschweig, P. T. Mortensen, D. Gras, W. Reiser, T. Lawo, H. Mansour, P. Sogaard, B. Stegemann, H.-J. Bruns, and C. Linde, "Monitoring of Physical Activity and Heart Rate Variability in Patients With Chronic Heart Failure Using Cardiac Resynchronization Devices," *The American Journal of Cardiology*, vol. 95, pp. 1104–1107, May 2005.
- [40] S. Tierney, H. Elwers, C. Sange, M. Mamas, M. Rutter, M. Gibson, L. Neyses, and C. Deaton, "What influences physical activity in people with heart failure? A qualitative study," *International Journal of Nursing Studies*, vol. 48, pp. 1234–1243, Oct. 2011.
- [41] M. Jehn, A. Schmidt-Trucksäss, H. Hanssen, T. Schuster, M. Halle, and F. Koehler, "Association of physical activity and prognostic parameters in elderly patients with heart failure," *Journal of Aging and Physical Activity*, vol. 19, pp. 1–15, Jan. 2011.
- [42] S. Tyagi, M. Berger, J. Fox, M. Curley, S. Strath, J. Rubenstein, J. Roth, and M. Widlansky, "Abstract 18222: Pacemaker Quantified Physical Activity Predicts All-Cause Mortality," *Circulation*, vol. 130, p. A18222, Nov. 2014.
- [43] J. Walsh, A. Charlesworth, R. Andrews, M. Hawkins, and A. Cowley, "Relation of daily activity levels in patients with chronic heart failure to long-term prognosis," *American Journal of Cardiology*, vol. 79, no. 10, pp. 1364–1369, 1997.
- [44] K. Izawa, S. Watanabe, K. Oka, K. Hiraki, Y. Morio, Y. Kasahara, N. Takeichi, T. Tsukamoto, N. Osada, and K. Omiya, "Physical activity in relation to exercise capacity in chronic heart failure patients," *International Journal of Cardiology*, vol. 152, pp. 152–153, Oct. 2011.
- [45] E. Rullman and T. Gustafsson, "Physical Activity Assessed with Accelerometer is an Accurate and Independent Predictor of Mortality in Patients with Advanced Heart Failure," *The FASEB Journal*, vol. 29, p. LB598, Apr. 2015.

- [46] E. Ketchum and W. Levy, "Multivariate Risk Scores and Patient Outcomes in Advanced Heart Failure," *Congestive Heart Failure*, vol. 17, pp. 205–212, Sept. 2011.
- [47] D. Aranki, G. Kurillo, P. Yan, D. M. Liebovitz, and R. Bajcsy, "Real-Time Tele-Monitoring of Patients with Chronic Heart-Failure Using a Smartphone: Lessons Learned," *IEEE Transactions on Affective Computing*, vol. 7, pp. 206–219, July 2016.
- [48] M. Jehn, S. Prescher, K. Koehler, S. von Haehling, S. Winkler, and O. e. Deckwart, "Tele-accelerometry as a novel technique for assessing functional status in patients with heart failure: Feasibility, reliability and patient safety," *International Journal of Cardiology*, vol. 168, pp. 4723–4728, Oct. 2013.
- [49] M. Tabata, R. Shimizu, D. Kamekawa, M. Kato, K. Kamiya, A. Akiyama, Y. Kamada, S. Tanaka, C. Noda, and T. Masuda, "Six-minute walk distance is an independent predictor of hospital readmission in patients with chronic heart failure," *International Heart Journal*, vol. 55, no. 4, pp. 331–336, 2014.
- [50] M.-K. Suh, C.-A. Chen, J. Woodbridge, M. Tu, J. Kim, A. Nahapetian, L. Evangelista, and M. Sarrafzadeh, "A Remote Patient Monitoring System for Congestive Heart Failure," *Journal of Medical Systems*, vol. 35, pp. 1165–1179, May 2011.
- [51] E. Haeuber, M. Shaughnessy, L. Forrester, K. Coleman, and R. Macko, "Accelerometer monitoring of home- and community-based ambulatory activity after stroke1," *Archives of Physical Medicine and Rehabilitation*, vol. 85, pp. 1997–2001, Dec. 2004.
- [52] J. Bardram, M. Frost, K. Szanto, and G. Marcu, "The MONARCA Self-Assessment System – A Persuasive Personal Monitoring System for Bipolar Patients," *PervasiveHealth*, 2012.
- [53] S. Vorrink, H. Kort, T. Troosters, and J.-W. Lammers, "Level of daily physical activity in individuals with COPD compared with healthy controls," *Respiratory Research*, vol. 12, no. 1, p. 33, 2011.
- [54] A. Alahmari, B. Kowlessar, A. Patel, A. Mackay, J. Allinson, J. Wedzicha, and G. Donaldson, "Physical activity and exercise capacity in patients with moderate COPD exacerbations," *The European Respiratory Journal*, vol. 48, pp. 340–349, Aug. 2016.
- [55] M. Faurholt-Jepsen, M. Frost, M. Vinberg, E. Christensen, J. Bardram, and L. V. Kessing, "Smartphone data as objective measures of bipolar disorder symptoms," *Psychiatry Research*, vol. 217, pp. 124–127, June 2014.
- [56] S. Javaheri, T. Parker, J. Liming, W. Corbett, H. Nishiyama, L. Wexler, and G. Roselle, "Sleep Apnea in 81 Ambulatory Male Patients With Stable Heart Failure: Types and Their Prevalences, Consequences, and Presentations," *Circulation*, vol. 97, pp. 2154–2159, June 1998.

- [57] T. Bradley and J. Floras, "Sleep Apnea and Heart Failure: Part II: Central Sleep Apnea," *Circulation*, vol. 107, pp. 1822–1826, Apr. 2003.
- [58] S. Javaheri, R. Shukla, H. Zeigler, and L. Wexler, "Central sleep apnea, right ventricular dysfunction, and low diastolic blood pressure are predictors of mortality in systolic heart failure," *Journal of the American College of Cardiology*, vol. 49, pp. 2028–2034, May 2007.
- [59] M. Naughton and G. Lorenzi-Filho, "Sleep in heart failure," *Progress in Cardiovascular Diseases*, vol. 51, pp. 339–349, Feb. 2009.
- [60] M. Bennett, M. Shao, and E. Gorodeski, "Home monitoring of heart failure patients at risk for hospital readmission using a novel under-the-mattress piezoelectric sensor: A preliminary single centre experience," *Journal of Telemedicine and Telecare*, vol. 23, pp. 60–67, Jan. 2017.
- [61] S. Rials, J. Hatlestad, A. Smith, D. Pubbi, D. Slotwiner, and J. Boehmer, "Night-time Elevation Angle in Heart Failure Patients Indicates Orthopnea and Paroxysmal Nocturnal Dyspnea," *Journal of Cardiac Failure*, vol. 23, p. S81, Aug. 2017.
- [62] K. Rahimi, C. Velardo, A. Triantafyllidis, N. Conrad, S. A. Shah, T. Chantler, H. Mohseni, E. Stoppani, F. Moore, C. Paton, C. Edmin, J. Ernst, and L. Tarassenko, "A user-centred home monitoring and self-management system for patients with heart failure: A multi-centre cohort study," *European Heart Journal - Quality of Care and Clinical Outcomes*, July 2015.
- [63]
- [64] C. Babbs, "Oscillometric measurement of systolic and diastolic blood pressures validated in a physiologic mathematical model," *BioMedical Engineering OnLine*, vol. 11, p. 56, Aug. 2012.
- [65] B. de Lindert and E. Van Someren, "Sleep Estimates Using Microelectromechanical Systems (MEMS)," *Sleep*, vol. 36, pp. 781–789, May 2013.
- [66] V. van Hees, S. Sabia, K. Anderson, S. Denton, J. Oliver, M. Catt, J. Abell, M. Kivimaki, M. Trenell, and A. Singh-Manoux, "A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer," *Plos One*, vol. 10, Nov. 2015.
- [67] A. Kosmadopoulos, C. Sargent, D. Darwent, X. Zhou, and G. Roach, "Alternatives to polysomnography (PSG): A validation of wrist actigraphy and a partial-PSG system," *Behavior Research Methods*, vol. 46, pp. 1032–1041, Jan. 2014.
- [68] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins, "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS medicine*, vol. 12, Mar. 2015.

- [69] A. Doherty, D. Jackson, N. Hammerla, T. Plötz, P. Olivier, and M. e. Granat, “Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study,” *Plos One*, vol. 12, Feb. 2017.
- [70] S. E. Crouter, J. R. Churilla, and D. R. B. Jr, “Estimating energy expenditure using accelerometers,” *European Journal of Applied Physiology*, vol. 98, pp. 601–612, Oct. 2006.
- [71] S. David, “Multivariate Density Estimation and Visualization,” *Humboldt-Universität Berlin, Center for Applied Statistics and Economics (CASE)*, vol. 16, 2004.
- [72] D. Jr, S. Lemeshow, and R. Sturdivant, *Applied Logistic Regression*. John Wiley & Sons, Apr. 2013.
- [73] D. Cox, “The Regression Analysis of Binary Sequences,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 20, no. 2, pp. 215–242, 1958.
- [74] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001.
- [75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Oct. 2011.
- [76] R. Killick, K. Haynes, I. Eckley, P. Fearnhead, and J. Lee, “changepoint: Methods for Changepoint Detection,” Oct. 2016.
- [77] R. Killick, K. Haynes, I. Eckley, P. Fearnhead, and J. Lee, “Package ‘changepoint’,” 2016.
- [78] J. Gorthi, C. Hunter, A. Mooss, V. Alla, and D. Hilleman, “Reducing Heart Failure Hospital Readmissions: A Systematic Review of Disease Management Programs,” *Cardiology Research*, vol. 5, pp. 126–138, Nov. 2014.
- [79] D. Scherr, P. Kastner, A. Kollmann, A. Hallas, J. Auer, H. Krappinger, H. Schuchlenz, G. Stark, W. Grander, G. Jakl, G. Schreier, and F. M. Fruhwald, “Effect of Home-Based Telemonitoring Using Mobile Phone Technology on the Outcome of Heart Failure Patients After an Episode of Acute Decompensation: Randomized Controlled Trial,” *Journal of Medical Internet Research*, vol. 11, Aug. 2009.
- [80] S. Chaudhry, Y. Wang, J. Concato, T. Gill, and H. Krumholz, “Patterns of Weight Change Preceding Hospitalization for Heart Failure,” *Circulation*, vol. 116, pp. 1549–1554, Oct. 2007.
- [81] J. Lewin, M. Ledwidge, C. O’Loughlin, C. McNally, and K. McDonald, “Clinical deterioration in established heart failure: What is the value of BNP and weight gain in aiding diagnosis?,” *European Journal of Heart Failure*, vol. 7, pp. 953–957, Oct. 2005.

- [82] A. Roth, I. Kajiloti, I. Elkayam, J. Sander, M. Kehati, and M. Golovner, "Telecardiology for patients with chronic heart failure: the 'SHL' experience in Israel," *International Journal of Cardiology*, vol. 97, pp. 49–55, Oct. 2004.
- [83] J. Brophy, G. Dagenais, F. McSherry, W. Williford, and S. Yusuf, "A multivariate model for predicting mortality in patients with heart failure and systolic dysfunction," *The American Journal of Medicine*, vol. 116, pp. 300–304, Mar. 2004.
- [84] C. O'Connor, D. Whellan, D. Wojdyla, E. Leifer, R. Clare, and S. e. Ellis, "Factors Related to Morbidity and Mortality in Patients With Chronic Heart Failure With Systolic DysfunctionClinical Perspective: The HF-ACTION Predictive Risk Score Model," *Circulation: Heart Failure*, vol. 5, pp. 63–71, Jan. 2012.
- [85] W. Levy, D. Mozaffarian, D. Linker, S. Sutradhar, S. Anker, and A. e. Cropp, "The Seattle Heart Failure Model Prediction of Survival in Heart Failure," *Circulation*, vol. 113, pp. 1424–1433, Mar. 2006.
- [86] H. Wedel, J. McMurray, M. Lindberg, J. Wikstrand, J. Cleland, and J. e. Cornel, "Predictors of fatal and non-fatal outcomes in the Controlled Rosuvastatin Multinational Trial in Heart Failure (CORONA): incremental value of apolipoprotein A-1, high-sensitivity C-reactive peptide and N-terminal pro B-type natriuretic peptide," *European Journal of Heart Failure*, vol. 11, pp. 281–291, Mar. 2009.
- [87] L. Manzano, D. Babalis, M. Roughton, M. Shibata, S. Anker, and S. e. Ghio, "Predictors of clinical outcomes in elderly patients with heart failure," *European Journal of Heart Failure*, vol. 13, pp. 528–536, May 2011.
- [88] S. Pocock, C. Ariti, J. McMurray, A. Maggioni, L. Kober, and I. e. Squire, "Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies," *European Heart Journal*, vol. 34, pp. 1404–1413, May 2013.
- [89] C. O'Connor, W. Abraham, N. Albert, R. Clare, W. Gattis Stough, M. Gheorghide, B. Greenberg, C. Yancy, J. Young, and G. Fonarow, "Predictors of mortality after discharge in patients hospitalized with heart failure: An analysis from the Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF)," *American Heart Journal*, vol. 156, pp. 662–673, Oct. 2008.
- [90] J. Foody, M. Farrell, and H. Krumholz, "Beta-Blocker Therapy in Heart Failure: Scientific Review," *JAMA*, vol. 287, pp. 883–889, Feb. 2002.
- [91] J. Cleland, A. Louis, A. Rigby, U. Janssens, and A. Balk, "Noninvasive Home Telemonitoring for Patients With Heart Failure at High Risk of Recurrent Admission and Death: The Trans-European Network-Home-Care Management System (TEN-HMS) study," *Journal of the American College of Cardiology*, vol. 45, pp. 1654–1664, May 2005.
- [92] P. Parwani and J. Ryan, "Heart Failure Patients With Low Blood Pressure: How Should We Manage Neurohormonal Blocking Drugs?," *Circulation: Heart Failure*, vol. 5, pp. 819–819, Nov. 2012.

- [93] G. Fonarow, K. Adams, W. Abraham, C. Yancy, W. Boscardin, ADHERE Scientific Advisory Committee, Study Group, and Investigators, "Risk stratification for in-hospital mortality in acutely decompensated heart failure: Classification and regression tree analysis," *JAMA*, vol. 293, pp. 572–580, Feb. 2005.
- [94] M. Hsieh, T. Auble, and D. Yealy, "Validation of the Acute Heart Failure Index," *Annals of Emergency Medicine*, vol. 51, pp. 37–44, Jan. 2008.
- [95] G. Felker, J. Leimberger, R. Califf, M. Cuffe, B. Massie, K. Adams, M. Gheorghiade, and C. O'Connor, "Risk stratification after hospitalization for decompensated heart failure," *Journal of Cardiac Failure*, vol. 10, pp. 460–466, Dec. 2004.
- [96] B. Huynh, A. Rovner, and M. Rich, "Long-term Survival in Elderly Patients Hospitalized for Heart Failure: 14-Year Follow-up From a Prospective Randomized Trial," *Archives of Internal Medicine*, vol. 166, pp. 1892–1898, Sept. 2006.
- [97] D. Lee, P. Austin, J. Rouleau, P. Liu, D. Naimark, and J. Tu, "Predicting Mortality Among Patients Hospitalized for Heart Failure: Derivation and Validation of a Clinical Model," *JAMA*, vol. 290, pp. 2581–2587, Nov. 2003.
- [98] W. Abraham, G. Fonarow, N. Albert, W. G. Stough, and M. e. Gheorghiade, "Predictors of in-hospital mortality in patients hospitalized for heart failure: insights from the Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF)," *Journal of the American College of Cardiology*, vol. 52, pp. 347–356, July 2008.
- [99] M. Bouvy, E. Heerdink, H. Leufkens, and A. Hoes, "Predicting mortality in patients with heart failure: a pragmatic approach," *Heart*, vol. 89, pp. 605–609, June 2003.
- [100] M. Böhm, K. Swedberg, M. Komajda, J. Borer, I. Ford, A. Dubost-Brama, G. Lerebours, and L. Tavazzi, "Heart rate as a risk factor in chronic heart failure (SHIFT): the association between heart rate and outcomes in a randomised placebo-controlled trial," *The Lancet*, vol. 376, pp. 886–894, Sept. 2010.
- [101] D. Kotecha, M. Flather, D. Altman, J. Holmes, G. Rosano, and J. e. Wikstrand, "Heart Rate and Rhythm and the Benefit of Beta-Blockers in Patients With Heart Failure," *Journal of the American College of Cardiology*, vol. 69, pp. 2885–2896, June 2017.
- [102] A. Bayes-Genis, M. de Antonio, A. Galán, H. Sanz, A. Urrutia, R. Cabanes, L. Cano, B. Gonzalez, C. Diez, T. Pascual, R. Elosua, and J. Lupon, "Combined use of high-sensitivity ST2 and NTproBNP to improve the prediction of death in heart failure," *European Journal of Heart Failure*, vol. 14, pp. 32–38, Jan. 2012.
- [103] M. Senni, P. Parrella, R. De Maria, C. Cottini, M. Böhm, and P. e. Ponikowski, "Predicting heart failure outcome from cardiac and comorbid conditions: The 3c-HF score," *International Journal of Cardiology*, vol. 163, pp. 206–211, Feb. 2013.
- [104] I. Anand, B. Greenberg, R. Fogoros, I. Libbus, and R. Katra, "Design of the Multi-Sensor Monitoring in Congestive Heart Failure (MUSIC) Study: Prospective Trial to Assess the Utility of Continuous Wireless Physiologic Monitoring in Heart Failure," *Journal of Cardiac Failure*, vol. 17, pp. 11–16, Jan. 2011.

- [105] I. Anand, W. H. W. Tang, B. Greenberg, N. Chakravarthy, I. Libbus, and R. Katra, "Design and Performance of a Multisensor Heart Failure Monitoring Algorithm: Results From the Multisensor Monitoring in Congestive Heart Failure (MUSIC) Study," *Journal of Cardiac Failure*, vol. 18, no. 4, pp. 289–295, 2012.
- [106] F. Javed, S. Farrugia, M. Colefax, and K. Schindhelm, "Early Warning of Acute Decompensation in Heart Failure Patients Using a Noncontact Measure of Stability Index," *IEEE transactions on bio-medical engineering*, vol. 63, pp. 438–448, Feb. 2016.
- [107] G. Koulaouzidis, D. Iakovidis, and A. Clark, "Telemonitoring predicts in advance heart failure admissions," *International Journal of Cardiology*, vol. 216, pp. 78–84, Aug. 2016.
- [108] D. Bates, *lme4: Mixed-effects modelling with R*. Springer, 2010.
- [109] M. Murphy, "Telehealth Alerts and Nurse Response," *Telemedicine Journal and E-Health: The Official Journal of the American Telemedicine Association*, Dec. 2017.
- [110] M. Hardinge, H. Rutter, C. Velardo, S. Shah, V. Williams, L. Tarassenko, and A. Farmer, "Using a mobile health application to support self-management in chronic obstructive pulmonary disease: A six-month cohort study eHealth/ telehealth/ mobile health systems," *BMC Medical Informatics and Decision Making*, vol. 15, June 2015.
- [111] Y. Miura, Y. Fukumoto, T. Miura, K. Shimada, M. Asakura, T. Kadokami, S.-i. Ando, S. Miyata, Y. Sakata, H. Daida, M. Matsuzaki, S. Yasuda, M. Kitakaze, and H. Shimokawa, "Impact of Physical Activity on Cardiovascular Events in Patients With Chronic Heart Failure," *Circulation Journal*, vol. 77, no. 12, pp. 2963–2972, 2013.
- [112] G. Parati, C. Lombardi, F. Castagna, P. Mattaliano, P. Filardi, P. Agostoni, and Italian Society of Cardiology (SIC) Working Group on Heart Failure members, "Heart failure and sleep disorders," *Nature Reviews. Cardiology*, vol. 13, pp. 389–403, July 2016.
- [113] M. Arzt, H. Woehrle, O. Oldenburg, A. Graml, A. Suling, E. Erdmann, H. Teschler, K. Wegscheider, and S. Investigators, "Prevalence and Predictors of Sleep-Disordered Breathing in Patients With Stable Chronic Heart Failure: The SchlaHF Registry," *JACC: Heart Failure*, vol. 4, pp. 116–125, Feb. 2016.
- [114] D. Hayes, M. I. Anstead, J. Ho, and B. A. Phillips, "Insomnia and chronic heart failure," *Heart Failure Reviews*, vol. 14, pp. 171–182, Sept. 2009.
- [115] A. Roebuck, V. Monasterio, E. Geder, M. Osipov, J. Behar, A. Malhotra, T. Penzel, and G. D. Clifford, "A review of signals used in sleep analysis," *Physiological Measurement*, vol. 35, no. 1, 2014.

- [116] M. Borazio, E. Berlin, N. Kücüküildiz, P. Scholl, and K. Laerhoven, "Towards Benchmarked Sleep Detection with Wrist-Worn Sensing Units," in *2014 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 125–134, Sept. 2014.
- [117] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin, "Automatic sleep/wake identification from wrist activity," *Sleep*, vol. 15, pp. 461–469, Oct. 1992.
- [118] J. Webster, D. Kripke, S. Messin, D. Mullaney, and G. Wyborney, "An activity-based sleep monitor system for ambulatory use," *Sleep*, vol. 5, no. 4, pp. 389–399, 1982.
- [119] C. Kushida, A. Chang, C. Gadkary, C. Guilleminault, O. Carrillo, and W. Dement, "Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients," *Sleep Medicine*, vol. 2, pp. 389–396, Sept. 2001.
- [120] N. Oakley, "Validation with polysomnography of the Sleepwatch sleep/ wake scoring algorithm used by the Actiwatch activity monitor system," *Technical Report to Mini-Mitter Co., Inc.*, 1997.
- [121] A. Sadeh, K. Sharkey, and M. Carskadon, "Activity-based sleep-wake identification: an empirical test of methodological issues," *Sleep*, vol. 17, pp. 201–207, Apr. 1994.
- [122] C.-C. Yang and Y.-L. Hsu, "A Review of Accelerometry-Based Wearable Motion Detectors for Physical Activity Monitoring," *Sensors*, vol. 10, pp. 7772–7788, Aug. 2010.
- [123] V. van Hees, Z. Fang, J. Langford, F. Assah, A. Mohammad, I. da Silva, M. Trenell, T. White, N. Wareham, and S. Brage, "Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: an evaluation on four continents," *Journal of Applied Physiology*, vol. 117, pp. 738–744, Oct. 2014.
- [124] V. van Hees, F. Renström, A. Wright, A. Gradmark, M. Catt, K. Chen, M. Löf, L. Bluck, J. Pomeroy, N. Wareham, U. Ekelund, S. Brage, and P. Franks, "Estimation of Daily Energy Expenditure in Pregnant and Non-Pregnant Women Using a Wrist-Worn Tri-Axial Accelerometer," *Plos One*, vol. 6, July 2011.
- [125] V. van Hees, L. Gorzelniak, E. Leon, M. Eder, M. Pias, S. Taherian, U. Ekelund, F. Renstrom, P. Franks, A. Horsch, and S. Brage, "Separating Movement and Gravity Components in an Acceleration Signal and Implications for the Assessment of Human Daily Physical Activity," *Plos One*, vol. 8, Apr. 2013.
- [126] K. Ellis, S. Godbole, S. Marshall, G. Lanckriet, J. Staudenmayer, and J. Kerr, "Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms," *Frontiers in Public Health*, vol. 2, Apr. 2014.

- [127] Z. Shan, H. Ma, M. Xie, P. Yan, Y. Guo, W. Bao, Y. Rong, C. L. Jackson, F. B. Hu, and L. Liu, "Sleep duration and risk of type 2 diabetes: a meta-analysis of prospective studies," *Diabetes Care*, vol. 38, pp. 529–537, Mar. 2015.
- [128] M. Jackowska and A. Steptoe, "Sleep and future cardiovascular risk: prospective analysis from the English Longitudinal Study of Ageing," *Sleep Medicine*, vol. 16, pp. 768–774, June 2015.
- [129] F. Cappuccio, D. Cooper, L. D'Elia, P. Strazzullo, and M. Miller, "Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies," *European Heart Journal*, vol. 32, pp. 1484–1492, June 2011.
- [130] K. Wulff, S. Gatti, J. Wettstein, and R. Foster, "Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease," *Nature Reviews. Neuroscience*, vol. 11, pp. 589–599, Aug. 2010.
- [131] G. Plasqui and K. Westerterp, "Physical Activity Assessment With Accelerometers: An Evaluation Against Doubly Labeled Water," *Obesity*, vol. 15, pp. 2371–2379, Oct. 2007.
- [132] C.-C. Yang and Y.-L. Hsu, "A Review of Accelerometry-Based Wearable Motion Detectors for Physical Activity Monitoring," *Sensors (Basel, Switzerland)*, vol. 10, pp. 7772–7788, Aug. 2010.
- [133] A. Rowlands, T. Yates, T. Olds, M. Davies, K. Khunti, and C. Edwardson, "Sedentary Sphere: Wrist-Worn Accelerometer-Brand Independent Posture Classification," *Medicine and Science in Sports and Exercise*, vol. 48, pp. 748–754, Apr. 2016.
- [134] W. Tryon, *Activity Measurement in Psychology and Medicine*. Springer Science & Business Media, Nov. 2013.
- [135] J. Coates and F. Meade, "The energy expenditure and mechanical energy demand in walking," *Ergonomics*, vol. 3, pp. 97–119, 1960.
- [136] T. Wong, J. Webster, H. Montoye, and R. Washburn, "Portable accelerometer device for measuring human energy expenditure," *IEEE transactions on bio-medical engineering*, vol. 28, pp. 467–471, June 1981.
- [137] H. Montoye, R. Washburn, S. Servais, A. Ertl, J. Webster, and F. Nagle, "Estimation of energy expenditure by a portable accelerometer," *Medicine and Science in Sports and Exercise*, vol. 15, no. 5, pp. 403–407, 1983.
- [138] M. J. Mathie, B. G. Celler, D. N. H. Lovell, and A. C. F. Coster, "Classification of basic daily movements using a triaxial accelerometer," *Medical and Biological Engineering and Computing*, vol. 42, pp. 679–687, Sept. 2004.
- [139] K. Kiani, C. Snijders, and E. Gelsema, "Computerized analysis of daily life motor activity for ambulatory monitoring," *Technology and Health Care: Official Journal of the European Society for Engineering and Medicine*, vol. 5, pp. 307–318, Oct. 1997.

- [140] D. Esliger, A. Rowlands, T. Hurst, M. Catt, P. Murray, and R. Eston, "Validation of the GENE Accelerometer:," *Medicine & Science in Sports & Exercise*, vol. 43, pp. 1085–1093, June 2011.
- [141] M. Hildebrand, V. VAN Hees, B. H. Hansen, and U. Ekelund, "Age group comparability of raw accelerometer output from wrist- and hip-worn monitors," *Medicine and Science in Sports and Exercise*, vol. 46, pp. 1816–1824, Sept. 2014.
- [142] S. Zhang, A. Rowlands, P. Murray, and T. Hurst, "Physical Activity Classification Using the GENE Wrist-Worn Accelerometer:," *Medicine & Science in Sports & Exercise*, vol. 44, pp. 742–748, Apr. 2012.
- [143] A. Bonomi, A. H. Goris, B. Yin, and K. Westerterp, "Detection of type, duration, and intensity of physical activity using an accelerometer," *Medicine and Science in Sports and Exercise*, vol. 41, pp. 1770–1777, Sept. 2009.
- [144] K. Ellis, J. Kerr, S. Godbole, G. Lanckriet, D. Wing, and S. Marshall, "A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers," *Physiological Measurement*, vol. 35, pp. 2191–2203, Oct. 2014.
- [145] P. Townsend, P. Phillimore, and A. Beattie, *Health and Deprivation: Inequality and the North*. Croom Helm, Jan. 1988.
- [146] K. Bradbury, W. Guo, B. Cairns, M. Armstrong, and T. Key, "Association between physical activity and body fat percentage, with adjustment for BMI: a large cross-sectional analysis of UK Biobank," *BMJ Open*, vol. 7, Mar. 2017.
- [147] M. Sjostrom, B. Ainsworth, A. Bauman, F. Bull, C. Hamilton-Craig, and J. Sallis, "Guidelines for data processing analysis of the International Physical Activity Questionnaire (IPAQ) - Short and long forms," *IPAQ*, Jan. 2005.
- [148] S. Pocock, C. Ariti, J. McMurray, A. Maggioni, L. Kober, I. Squire, K. Swedberg, J. Dobson, K. Poppe, G. Whalley, and R. Doughty, "Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies," *European Heart Journal*, vol. 34, pp. 1404–1413, May 2013.
- [149] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, pp. 16–28, Jan. 2014.
- [150] D. Lauderdale, K. Knutson, L. Yan, K. Liu, and P. Rathouz, "Sleep duration: how well do self-reports reflect objective measures? The CARDIA Sleep Study," *Epidemiology (Cambridge, Mass.)*, vol. 19, pp. 838–845, Nov. 2008.
- [151] T. Arora, O. Omar, and S. Taheri, "Assessment for the possibility of a first night effect for wrist actigraphy in adolescents," *BMJ Open*, vol. 6, Oct. 2016.
- [152] C. Velardo, S. Shah, O. Gibson, G. Clifford, C. Heneghan, H. Rutter, A. Farmer, and L. Tarassenko, "Digital health system for personalised COPD long-term management," *BMC Medical Informatics and Decision Making*, vol. 17, Feb. 2017.

-
- [153] J. Cleland and R. Antony, “It makes SENSE to take a safer road,” *European Heart Journal*, vol. 32, pp. 2225–2227, Sept. 2011.
- [154] T. Farnia, M.-C. Jaulent, and O. Steichen, “Evaluation Criteria of Noninvasive Telemonitoring for Patients With Heart Failure: Systematic Review,” *Journal of Medical Internet Research*, vol. 20, no. 1, 2018.