Tampere University

EMMI ANTIKAINEN

Time Series Analytics for Decision Support in Chronic Diseases

Clinical case studies

Tampere University Dissertations 952

Tampere University Dissertations 952

EMMI ANTIKAINEN

Time Series Analytics for Decision Support in Chronic Diseases

Clinical case studies

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Information Technology and Communication Sciences
of Tampere University,
for public discussion in the Lecture room TB109
of the Tietotalo building, Korkeakoulunkatu 1, Tampere,
on 15 March 2024, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Information Technology and Communication Sciences
Finland

Responsible supervisor and Custos	Professor Moncef Gabbouj Tampere University Finland	
Supervisor	Professor Mark van Gils Tampere University Finland	
Pre-examiners	Professor Luca Mainardi Polytechnic University of Milan Italy	Associate Professor Ioanna Chouvarda Aristotle University of Thessaloniki Greece
Opponent	Professor Tapio Seppänen University of Oulu Finland	

Cover design: Roihu Inc.

ISBN 978-952-03-3285-3 (print)
ISBN 978-952-03-3286-0 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-3286-0

Carbon dioxide emissions from printing Tampere University dissertations have been compensated.

PunaMusta Oy – Yliopistopaino
Joensuu 2024

PREFACE

The presented research was conducted at the VTT Technical Research Centre of Finland Ltd. A personal grant from Ella and Georg Ehrnrooth foundation to finalize this thesis is gratefully acknowledged.

First and foremost, I thank my supervisors professor Moncef Gabbouj and professor Mark van Gils. I am deeply grateful for the way Moncef welcomed me into his mentorship and team: with kindness and excitement to work together. He always had his door open for me to discuss any plans, problems, or life in general. He continuously inspired me with his creative yet systematic take on technologies. I greatly appreciate the way Mark was always ready to listen and help with any obstacles and regularly broadened my perspective with clever insights. With Mark's guidance, I learned to assess my work from diverse viewpoints and search for the best solution considering all stakeholders. I could not imagine a better pair of supervisors for this work.

I thank my co-authors for the invaluable work and support they offered in order to bring the best value to our work together. I am particularly thankful to Dr Meenakshi Chatterjee for the pleasure of working together so intensively despite the vast distance and time difference between us, for the countless lessons in accurate and firm communication, and for being such a great role model for women in science, myself included. I am also very thankful for the enlightening discussions with professor Jussi Hernesniemi on the clinical perspectives regarding cardiac patient care and with professor Walter Maetzler on the clinical perspectives pertaining to neurological diseases. I further extend my gratitude to all the project consortium members who made the enormous data collections happen in the IDEA-FAST project (https://idea-fast.eu/the-idea-fast-investigators/) and the MADDEC project.

I thank my colleagues during my years at VTT for all the educational discussions, team spirit, and providing the fertile ground and nurturing environment for professional growth from a student to a researcher. Special thanks to Dr Teemu Ah-

maniemi, Dr Juha M. Kortelainen, and Dr Oleg Antropov who played important roles supporting the presented publications. I am also grateful for the practical tips and peer support from the VTT Young Professionals in Tampere and the colleagues who were working on theses of their own. Furthermore, I would like to express my thanks to my new colleagues at Oura Health Oy, especially the science department and the Tampere office, for the amazingly brilliant and enjoyable working environment, the support you have shown towards my thesis project, and the warm and genuine atmosphere in which I have had the pleasure of taking a new step in my career building algorithms for daily life. I also thank the SAMI research group at TUNI for making me feel welcome and part of the group. Thanks for the course projects and interesting discussions.

I am grateful to my friends and family for all the encouraging words and making sure I also had life outside work and the thesis project. Finally, I want to express my heartfelt gratitude to Atte who has been my closest peer support, endured my methods of working, cheered me on through the difficulties, and celebrated the successes with me. I also thank my four-legged companions, Deni and Nano, for taking care of my work-exercise balance and for the emotional support during my studies.

Pirkkala, December 2023

Emmi Antikainen

ABSTRACT

Chronic diseases burden patients with unending symptoms and functional decline, which limit the activities of daily living and decrease working ability. They increase the risk of injuries, comorbidities, and death. The disabilities from chronic diseases are a major contributor to disease burden globally. With the aging and increasingly obese population, chronic diseases are becoming increasingly common.

Time series analytics offer means to investigate the evolution of chronic diseases over time. The analysis of time dependent patterns can facilitate diverse applications for clinical decision support. Modern analytical methods have grown extremely powerful with the accelerated development of computational resources, being able to mine vast amounts of data and enabling the discovery of all the more complex patterns. Moreover, modern sensor technologies and electronic health record systems have boosted the continuous buildup of high quality health data, monitoring physiological events at hospitals and throughout everyday life. This thesis presents four studies that delve into chronic disease related algorithms across various application time spans, ranging from overnight to several months. The thesis centers around cardiorespiratory measurements collected from healthy and chronic disease patients, measured at hospitals, in free-living settings, and in a controlled laboratory environment.

The studies cover contact-free overnight vital sign monitoring for sleep apnoea detection, wearable sensor based continuous monitoring for fatigue and sleep assessment in neurodegenerative and immune-mediated inflammatory diseases, and six-month mortality risk prediction from electronic health records in cardiac patients. The work applies traditional model driven signal processing as well as the more recently emerged data driven deep learning methods, such as transformer neural networks. This thesis presents pragmatic insights on building time series based decision support tools for chronic disease management, and addresses the requirements and limitations related to time series analytics and the underlying data collection across

the above-specified time spans. Robust algorithms for contact-free vital sign monitoring are presented and evaluated in broad physiological conditions, the feasibility of continuous monitoring in outpatients and the diverse measurement associations with health related quality of life are analyzed, and the benefits of applying deep learning on health records but also their disadvantages in clinical use are presented. The results imply the importance of high frequency data in applications with short time spans, data collection context tracking in continuous monitoring, and data quality and coverage across all application time spans. The algorithms proposed in this thesis are validated with data collected from human volunteers, including chronic disease patients from the selected disease groups.

TIIVISTELMÄ

Krooniset sairaudet kuormittavat niistä kärsiviä potilaita tauotta, heikentäen niin työkykyä kuin yleistä toimintakykyä, sekä arkisuoriutumista. Ne kasvattavat loukkaantumisriskiä, oheissairauksien esiintyvyyttä, sekä kuolleisuutta. Toimintakyvyn heikentyminen kroonisten sairauksien vuoksi aiheuttaa merkittävän osan globaalia tautitaakkaa. Väestön ikääntyessä ja ylipainon yleistyessä kroonisista sairauksista on tulossa entistä yleisempiä.

Aikasarja-analytiikka tarjoaa keinoja kroonisten sairauksien kehittymisen tarkasteluun. Aika-riippuvaisia ilmiöitä analysoimalla voidaan mahdollistaa kliinisen päätöksenteon tuen sovellutuksia monipuolisesti eri käyttökohteissa. Modernit analyyttiset menetelmät ovat kehittyneet huomattavan tehokkaiksi laskentatehon yleisen kehityksen myötä, mahdollistaen suurien datamäärien louhimisen ja entistäkin monimutkaisempien yhteyksien ja toistuvien kaavojen paljastamisen. Samaan aikaan modernit sensoriteknologiat ja sähköiset potilastietojärjestelmät ovat edistäneet hyvälaatuisen terveysdatan jatkuvaa kertymistä tietovarastoihin niin arkielämästä kuin sairaalamittauksistakin. Tässä väitöstyössä esitellään neljä tutkimusta, jotka syventyvät kroonisiin sairauksiin liittyviin algoritmeihin eri aikaskaalojen käyttösovelluksissa aina yön yli kestävistä mittauksista useiden kuukausien seurantajaksoihin. Väitöskirjassa keskitytään sydän- ja hengityselinten toimintojen mittauksiin. Mittausdataa kerättiin sekä terveiltä koehenkilöiltä että kroonisesti sairailta potilailta eri ympäristöissä: sairaalassa, kontrolloimattomissa olosuhteissa arkielämässä, sekä kontrolloidussa laboratorioympäristössä.

Esitetyt tutkimukset kattavat elintoimintojen monitoroinnin tutkateknologialla yön yli erityisesti uniapnean seurantasovelluksiin, uupumuksen ja uniongelmien arvioimisen puettavien älylaitteiden välityksellä neurodegeneratiivisten sairauksien sekä tulehduksellisten suolisto- ja reumasairauksien yhteydessä, sekä puolen vuoden kuolleisuusriskin ennustamisen potilastietojärjestelmän tiedoista sydän- ja verisuonisairailla. Väitöstyössä sovelletaan sekä perinteisiä mallipohjaisen signaalinkäsittelyn

menetelmiä, että uusimpia datalähtöisiä syviä neuroverkko-pohjaisia menetelmiä. Aikasarjoihin perustuvien päätöksenteon tuen työkalujen kehittämistä tarkastellaan käytännönläheisesti kroonisten sairauksien hoitoon keskittyen. Työ käsittelee niin aikasarja-analytiikan menetelmiin kuin datan keräämisen liittyviä vaatimuksia ja rajoituksia eri aikaskaalojen käyttösovelluksissa. Väitöstyössä esitellään algoritmeja kontaktittomaan monitorointiin ja validoidaan ne kattavasti erilaisissa fysiologisissa tiloissa, arvioidaan puettavien sensoreiden soveltuvuutta avohoitopotilaiden monitorointiin sekä niistä saatavien mittausten assosiaatioita terveyteen liittyvän elämänlaadun arviointiin, ja tutkitaan syvien neuroverkkojen hyötyjä ja heikkouksia potilastietokantojen käsittelyssä kliinisiä sovelluksia ajatellen. Tutkimustulosten perusteella korkealla näytetaajuudella kerätty mittausdata vaikuttaa sitä tärkeämmältä, mitä lyhyemmän tähtäimen käyttösovellus on kyseessä. Lisäksi datankeruun kontekstin seuraamisesta havaittiin olevan hyötyä jatkuvissa mittauksissa, ja datan laadun ja kattavuuden tärkeys havaittiin kaikissa tutkituissa sovellutuksissa.

CONTENTS

ABBREVIATIONS

1D CNN	1-dimensional convolutional neural network
6MWT	Six-minute walk test
ACF	Autocorrelation function
ADL	Activities of daily living
AI	Artificial intelligence
ANCOVA	Analysis of covariance
ANN	Artificial neural network
ANS	Autonomic nervous system
AUC	Area under the receiver operating characteristics curve
AV node	Atrioventricular node
BCG	Ballistocardiography
BERT	Bidirectional encoder representations from transformers
BMI	Body-mass index
bpm	Beats per minute
CCU	Coronary care unit
CNN	Convolutional neural network
CPAP	Continuous positive airway pressure
CRS	Cardiorespiratory system
CV	Coefficient of variation
CVD	Cardiovascular disease
DALY	Disability-adjusted life-years

DFT	Discrete Fourier transform
DL	Deep learning
DSP	Digital signal processing
ECG	Electrocardiography
EEG	Electroencephalogram
EHR	Electronic health record
ENS	Enteric nervous system
FDA	Food and Drug Administration
FFT	Fast Fourier transform
FMCW	Frequency modulated continuous wave
FN	False negative
FP	False positive
GRACE	Global registry of acute coronary events
HD	Huntington's disease
HR	Heart rate
HRQoL	Health-related quality of life
HRR	Heart rate recovery
HRV	Heart rate variability
IBD	Inflammatory bowel disease
IBI	Interbeat interval
ICU	Intensive care unit
IMID	Immune-mediated inflammatory disease
KS test	Kolmogorov-Smirnov test
KSS	Karolinska sleepiness scale
L5	Least active five hours of a day
LF/HF ratio	Low frequency (0.04-0.15 Hz) to high frequency (0.15-0.40 Hz) ratio
LSTM	Long short-term memory

MADDEC	Mass data in detection and prevention of serious adverse events in cardiovascular disease
MAE	Mean absolute error
mCSI	Modified cardiac sympathetic index
ML	Machine learning
N	Negative
NDD	Neurodegenerative disease
NN	Normal-to-normal peak intervals
P	Positive
PCI	Percutaneous coronary intervention
PD	Parkinson's disease
pH	Hydrogen concentration
PN	Predicted negative
PNS	Parasympathetic nervous system
PP	Predicted positive
PPG	Photoplethysmography
PRO	Patient reported outcome
PSD	Power spectral density
PSG	Polysomnography
PSS	Primary Sjögren's disease
RA	Rheumatoid arthritis
RIP	Respiratory inductance plethysmography
RMSE	Root mean squared error
RMSSD	Root mean square of successive differences
RNN	Recurrent neural network
rpm	Respirations per minute
SA node	Sinoatrial node
SD	Standard deviation

SDSD	Standard deviation of consecutive differences in adjacent normal-to-normal intervals
SLE	Systemic lupus erythematosus
SMA	VTT Stress monitor application
SNS	Sympathetic nervous system
SpO2	Blood oxygen saturation
TAVI	Transcatheter aortic valve implantation
TN	True negative
TP	True positive
VLF	Very low frequencies (0.003-0.04 Hz)
VO_2max	Maximal oxygen consumption

ORIGINAL PUBLICATIONS

Publication I **E. Turppa***, J. M. Kortelainen, O. Antropov, and T. Kiuru, "Vital sign monitoring using FMCW radar in various sleeping scenarios," *Sensors*, vol. 20, no. 22:6505, 2020. DOI: 10.3390/s20226505.

Publication II **E. Antikainen**, H. Njoum, J. Kudelka, D. Branco, R. Z. U. Rehman, V. Macrae, K. Davies, H. Hildesheim, K. Emmert, R. Reilmann, C. J. van der Woude, W. Maetzler, W.-F. Ng, P. O'Donnell, G. Van Gassen, F. Baribaud, I. Pandis, N. V. Manyakov, M. van Gils, T. Ahmaniemi, and M. Chatterjee, "Assessing fatigue and sleep in chronic diseases using physiological signals from wearables: A pilot study," *Frontiers in Physiology*, vol. 13, no. 968185, 2022. DOI: 10.3389/fphys.2022.968185.

Publication III **E. Antikainen**, R. Z. U. Rehman, T. Ahmaniemi, and M. Chatterjee, "Predicting daytime sleepiness from electrocardiography based respiratory rate using deep learning," in *2022 Computing in Cardiology (CinC)*, 2022. DOI: 10.22489/CinC.2022.100.

Publication IV **E. Antikainen**, J. Linnosmaa, A. Umer, N. Oksala, M. Eskola, M. van Gils, J. Hernesniemi, and M. Gabbouj, "Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records," *Scientific Reports*, vol. 13, no. 3517, 2023. DOI: 10.1038/s41598-023-30657-1.

*Author's earlier surname was Turppa.

Author's contribution

Publication I — The author of the thesis participated in the study conceptualisation and analysis planning. She conducted the measurements and the formal data analysis, produced the visualisations and led the manuscript writing.

Publication II — The author of the thesis led the data analysis and manuscript writing, and produced the visualisations. She planned the analysis, implemented data pre-processing, and interpreted the results together with co-authors.

Publication III — The author of the thesis planned the data pre-processing and analysis, led the analysis and manuscript writing, and produced the visualisations. She implemented data pre-processing and interpreted the results together with co-authors.

Publication IV — The author of the thesis led the data pre-processing, analysis, and manuscript writing. She produced the visualisations. She designed the study and interpreted the results together with co-authors.

Valuable comments and editing were provided by the co-authors.

1 INTRODUCTION

Modern sensor technologies have become highly popular for health tracking among the general public. They have become more user-friendly and comfortable to use and have lighter materials and longer battery-lives. The global market for wearable technology is predicted to increase from USD 61.3 million in 2022 to USD 186.1 million by 2030, mostly thanks to health tracking consumer devices [1]. These devices generate a large variety of longitudinal health data with unprecedented potential for creating personalized health solutions ranging from predictive and preventive applications to new gold standards for clinical assessments. The realisation of this paradigm shift in healthcare depends on the clinical acceptance and uptake of the new technologies and the associated algorithms. The sensors and algorithms need to be validated for clinical use, also including any machine learning (ML), artificial intelligence (AI), or other signal processing techniques they may contain. Moreover, they will need to generate easily interpretable insights which can be further integrated to the electronic health record (EHR) systems to enable seamless everyday clinical use. Hospitals around the world have widely adopted EHRs to manage the data masses but currently lack methods to fully exploit the new knowledge embedded in the longitudinal and mixed type data coming from the modern sensors.

One of the most promising application areas for longitudinal data applications lies in chronic disease management. Non-communicable diseases (including chronic diseases) account for over 70 % of yearly deaths globally [2]. The disabilities caused by chronic diseases and injuries already cause over half of the entire disease burden in 11 countries [3]. Minimizing the effects of chronic diseases can improve the patient's health-related quality of life (HRQoL), improve patient outcomes through interventions and prevention, and reduce healthcare costs. Historical data, anomalies and repeating patterns in recorded data or other information can contain far-reaching and crucial information for chronic disease management. For example, in sleep apnoea, repetitive cessation of breathing during sleep can cause daytime sleepiness and

decline in cognitive function, ultimately increasing the risk of more serious adverse health events such as falls or even death [4]. Early detection of disease onset and disease state monitoring can efficiently improve the patient's HRQoL [5].

The evolution of health data in time is extremely important for various healthcare applications. Detecting complicated patterns and complex dependencies has become easier than ever before thanks to the recent advances in computational power. In practice, clinical data analysis applications range over all possible time spans and hence also have different requirements on both input and output. In-patient applications onsite at healthcare facilities may emphasize the importance of real-time detection of patient deterioration, whereas outpatient monitoring may be more focused on the remote assessment of HRQoL. Future development in the field could benefit from wider understanding of the realities of time series applications in patient monitoring. New approaches are needed to extract health insights from different time scales and to allow the clinicians to fully harness their patients' historical health data to improve patient outcomes.

1.1 Scope and objectives

This thesis aims to promote time series analytics for decision support in chronic diseases through three clinical case studies. The thesis aims to offer pragmatic insights on creating time series based tools by focusing on clinical applications that aim to support chronic disease management through monitoring cardiovascular and respiratory functions and health events over the course of one day, weeks, or months. The applications range from near real-time health event detection to monitoring changes in the health-related quality of life to predicting the risk of death. Specifically, the applications cover sleep apnoea monitoring overnight, fatigue and sleep problem monitoring in chronically ill over weeks, and cardiac patient mortality prediction over six months. The scope of the thesis is summarized in Figure 1.1.

This thesis presents time series based algorithms for different time scales in the context of selected clinical use cases. The thesis studies model driven algorithms, which rely on *a priori* knowledge and conventional digital signal processing, and data driven algorithms, which adapt to example data. The performance of the algorithms is evaluated with data from real human subjects, and the suitability of especially deep learning methods for applications over days, weeks or months is analyzed, while

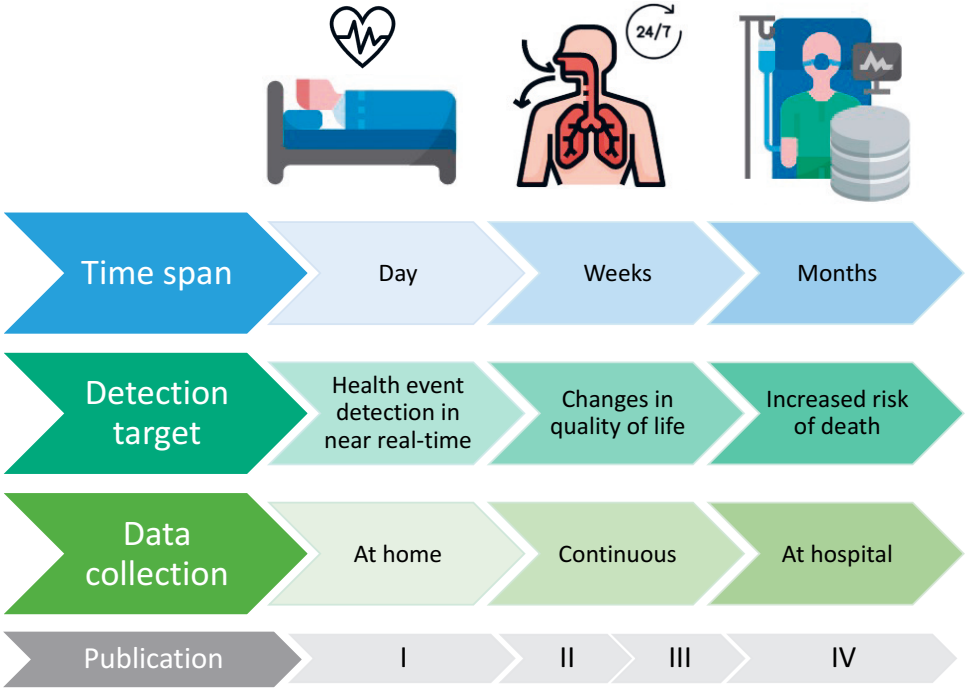

Figure 1.1 Scope of the thesis. The thesis focuses on time series analytics applications in clinical use cases ranging over different time spans of interest with various detection targets and varying data collection settings. The scope of each individual publication is indicated on the lowest row. Icons by Just Icon and adapted from icons by WiStudio and Paomedia, used under CC BY 3.0.

assessing the requirements for data collection methods, quality, and availability for different time spans of interest. This thesis focuses the following research questions.

1. What are the requirements for, and limitations of model driven and data driven time series analytics across different time spans for chronic disease monitoring? How do model driven and data driven methods perform and compare across time spans?
2. What requirements and limitations relate to the clinical use and uptake of time series analytics applications?
3. Which data collection methods are feasible with chronic disease patients?
4. What data collection requirements relate to different time spans of interest?

1.2 Thesis outline

The thesis starts with the Chronic diseases and physiology section, which outlines the physiological functions of the cardiorespiratory system, measured throughout the thesis, and the selected chronic diseases inspected in this thesis. The Methodology and prior work section describes the means of data collection and the analytical techniques applied in the publications, as well as the prior work in the related time series applications. The Study data section describes the data of each publication. The Results section summarizes the results of each publication, which are further discoursed in the Discussion. Finally, Conclusion summarizes the main findings of this work.

2 CHRONIC DISEASES AND PHYSIOLOGY

Human physiology comprises a complex network of processes where everything is interdependent. All subsystems in the human body, the physiological systems, interact and co-operate to maintain conditions that keep us alive and, on the other hand, adjust the bodily functions dynamically to adapt to internal or external changes [6]. These systems include, e.g. the cardiovascular system, respiratory system, nervous system, digestive system, endocrine system, and the list goes on. Quantifying physiological changes with respect to time may provide information that can be used to prevent or predict certain health events.

This thesis studies applications related to chronic diseases and the cardiorespiratory system (CRS), aiming to detect physiological events and exploit physiological time-series to improve the quality of life and health outcomes of the chronically ill. Section 2.1 outlines the anatomy, regulation, as well as normal and abnormal functions of the cardiorespiratory system, whereas section 2.2 describes the chronic diseases included in this thesis. Understanding the basic underlying physiology is fundamental for signal processing and machine learning technologies that aim to transform the collected data into actionable information.

2.1 Cardiorespiratory system

The CRS contains the cardiovascular and respiratory systems. The heart, vessels, lungs, and other organs work together to provide cells with blood full of important substances such as oxygen, hormones, and amino acids, and simultaneously, a means to remove metabolic wastes from the body. As blood also carries heat around the body, the CRS puts all main vital signs into effect: heart rate, respiratory rate, body temperature, blood oxygen, and blood pressure. These vital signs are the physiological measures primarily monitored at healthcare facilities and strong indicators of patient deterioration [7].

The chronic diseases described in section 2.2 pertain to the cardiorespiratory physiology, either directly at the root of the disease or indirectly through more complex causal relationships. Sleep apnoea is directly related to the respiratory system. Neurodegenerative diseases (NDDs) have been associated with autonomic dysfunction, even presymptomatically [8], [9]. NDDs and immune-mediated inflammatory diseases (IMIDs) have also been connected to decreased post-exercise heart rate recovery (HRR) in controlled settings [10]–[15]. Cardiovascular diseases (CVDs) directly relate to malfunction in the cardiovascular system whether it is, for instance, atrial fibrillation due to overwhelming impulses originating outside the sinoatrial (SA) node (regulatory malfunction), or coronary artery disease caused by atherosclerotic plaque in the coronary arteries, causing reduced blood flow in to the heart [16].

The main characteristics and functions of cardiovascular and respiratory subsystems are detailed in subsections 2.1.1 and 2.1.2, respectively. Importantly, the CRS operates dynamically, reacting to changes both within and outside the body. The CRS functions are regulated by the nervous system, as described in 2.1.3.

2.1.1 Cardiovascular system

The blood enters the heart through the right atrium. As illustrated in Figure 2.1, it passes via the tricuspid valve while flowing into the right ventricle, which pumps it on through the pulmonary valve into the pulmonary circulation in the lungs, where the blood regains oxygen while giving up carbon dioxide and other gases [17]. The blood then re-enters the heart through the left atrium and passes the mitral valve upon entering the left ventricle [17]. Finally, the blood leaves the heart through the aortic valve, moving on to the aorta, arteries, and arterioles, flowing into capillaries to distribute the oxygen and other substances to the cells [17]. The cycle restarts when deoxygenated blood passes through the venules and veins back to the heart [17]. The valves open and close based on the pressure differences created by the blood itself between the atria and the ventricles. Nevertheless, the mechanical contraction of the heart is required to create the pressure differences and maintain the blood circulation [17].

The SA node (also sinus node) initiates the contraction and sets the heart rate (sinus rhythm) [6]. The specialized cells in the sinus node spontaneously depolarize, disturbing the cell's membrane potential and causing an electrical impulse that spreads to the atria and the atrioventricular node (AV node) [17]. The AV node cru-

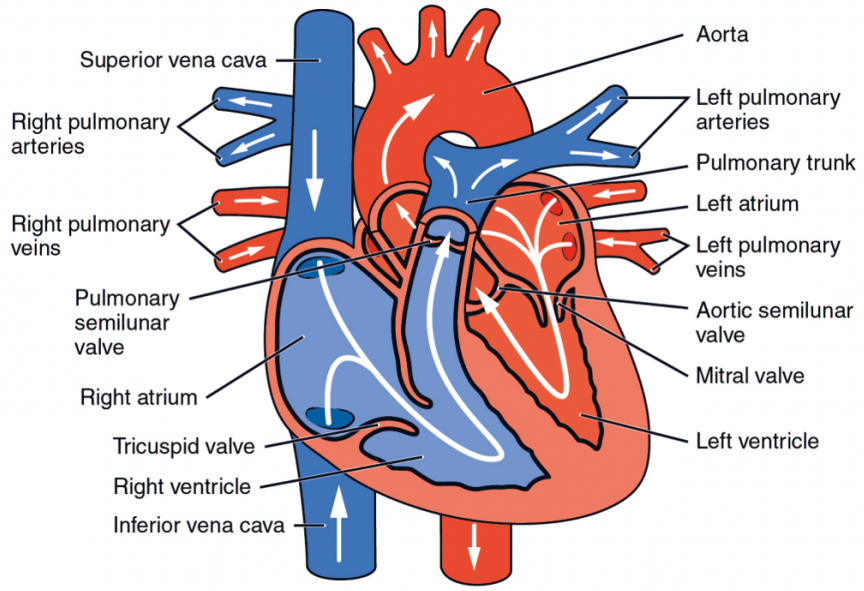

Figure 2.1 The anatomy of the heart. Blood flow is indicated with arrows. Figure used under the CC BY 4.0 license, adapted from [6].

cially delays the impulse to allow the blood flow from the atria to the ventricles [17]. From the AV node, the impulse travels to the bundle of His (a group of specialized conductive muscle cells), which brings it to ventricles where the impulse ultimately reaches the Purkinje fibres and hence the myocardial cells of the ventricles and starts the ventricular contraction [17]. The cardiac conduction system enables the impulse to travel throughout the heart more rapidly than it could travel via the myocardial tissue [17]. The spontaneously depolarizing cells in the SA node create the normal naturally paced sinus rhythm [17]. However, other similar cells exist also elsewhere in the heart and may initiate the contraction should the SA node fail to do so [17].

The above-described cardiac cycle is divided into two phases: diastole and systole. Diastole is the phase when the ventricles relax and fill with blood, whereas the contraction phase is known as the systole. Blood pressure is measured during both phases independently. Blood pressure is further interlinked with the heart rate and stroke volume. Blood pressure, heart rate and storke volume together form the cardiac output, which is regulated by several different mechanisms to ensure sufficient blood flow and pressure in vital organs and perfusion to tissues. These mechanisms include (1) autoregulation within the organs (vasodilation or vasoconstriction based

on metabolic need or blood pressure), (2) neural regulation through the autonomic nervous system (ANS), and (3) endocrine regulation [17]. Neural regulation includes the baroreceptor reflex (baroreflex), chemoreceptor reflex, as well as more sophisticated ANS activity [18]. The baroreflex reacts to blood pressure changes in a fraction of a second and helps the body adjust, e.g., to postural changes, while the chemoreceptor reflex reacts to changes of hydrogen concentration (pH) and affects the respiratory rate accordingly [16], [18]. Finally, endocrine regulation employs different hormones, such as adrenaline and norepinephrine in stress reactions, to adjust the cardiac output [18].

2.1.2 Respiratory system

The respiratory system comprises upper and lower airways and some muscles that facilitate breathing, such as the diaphragm [17]. The upper airways consist of the

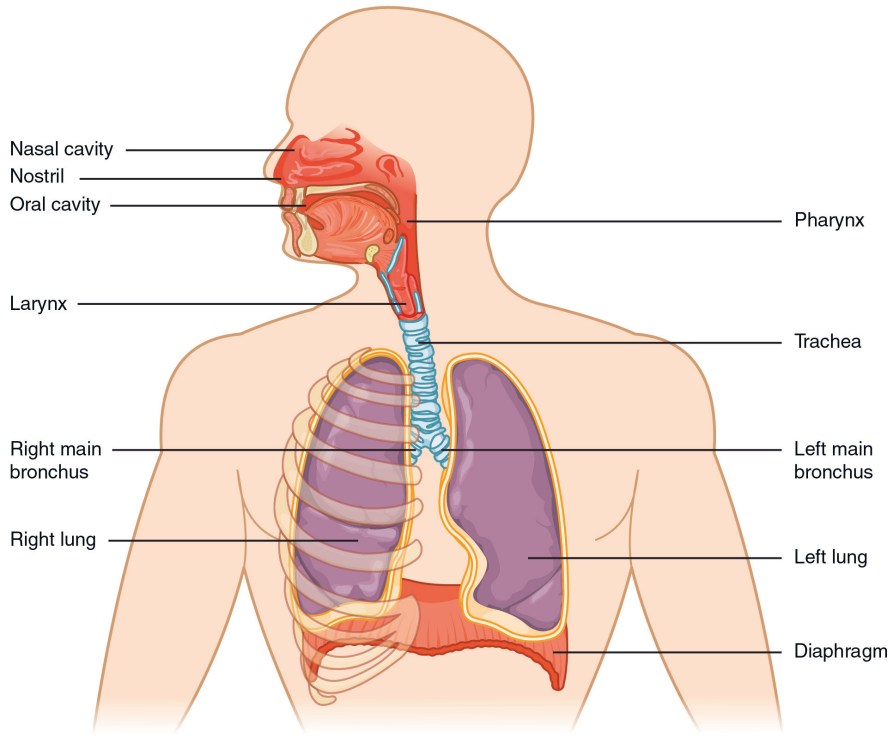

Figure 2.2 The anatomy of the major structures of the respiratory system. Figure from [6], used under the CC BY 4.0 license.

nose and mouth and the respective cavities, paranasal sinuses, and pharynx [17]. The lower airways contain the trachea, bronchi, bronchioles, and the lungs [17]. The larynx is located between the pharynx and the trachea, and is sometimes considered a part of the upper, sometimes the lower airways. The components of the respiratory system are illustrated in Figure 2.2. The air is first inhaled through the airways to oxygenate blood within the pulmonary circulation, and exhaled to remove the waste gases [6]. The oxygen primarily adheres onto the hemoglobin molecules in the blood (over 98%) but in small amounts also dilutes into the plasma [17]. The exchange of gases takes place in the hundreds of millions of alveoli in the lungs. The alveoli do not allow air to mix with blood but facilitate the gas exchange by creating an extremely small distance between the two (on average around 0.6–0.8µm, approximately the diameter of a red blood cell) [17]. This enables diffusion, which is driven by the partial pressure difference of a gas between the inhaled air and the blood.

In contrast to heart rate, respiratory rate can also be controlled voluntarily to some extent, until the chemical stimulation from CO_2 builds up and restarts respiration automatically [17]. However, the brain starts to develop permanent damage after five to eight minutes without oxygen, underscoring the vital role of the respiratory system [16]. Normally, respiration is regulated by the nervous system (impulses originating from the medulla oblongata) and via chemoreceptors that monitor the chemical composition of the blood (oxygen, CO_2, pH) [16].

2.1.3 Neurophysiology

The autonomic nervous system (ANS) adapts the body to any external or internal perturbation by controlling involuntary physiological functions, such as cardiac output, respiration, and blood flow [20]. It is a major part of the peripheral nervous system and extends to nearly all tissues but consciously controlled skeletal muscles. The ANS is responsible for maintaining the integrity of cells, tissues, and organs (i.e., homeostasis) in the dynamically changing settings the human body may experience, be it common everyday life or sudden unexpected events. It responds to, e.g., exercise, stressful situations, different body positions, and illness [20]. Also, the body temperature, immune system, and inflammatory processes are regulated by the ANS. Disruptions in ANS functions can lead to or stem from a plethora of diseases including cardiorespiratory conditions such as hypertension, stroke, sleep disorders, and Parkinson's disease [20]. The ANS consists of three branches: the

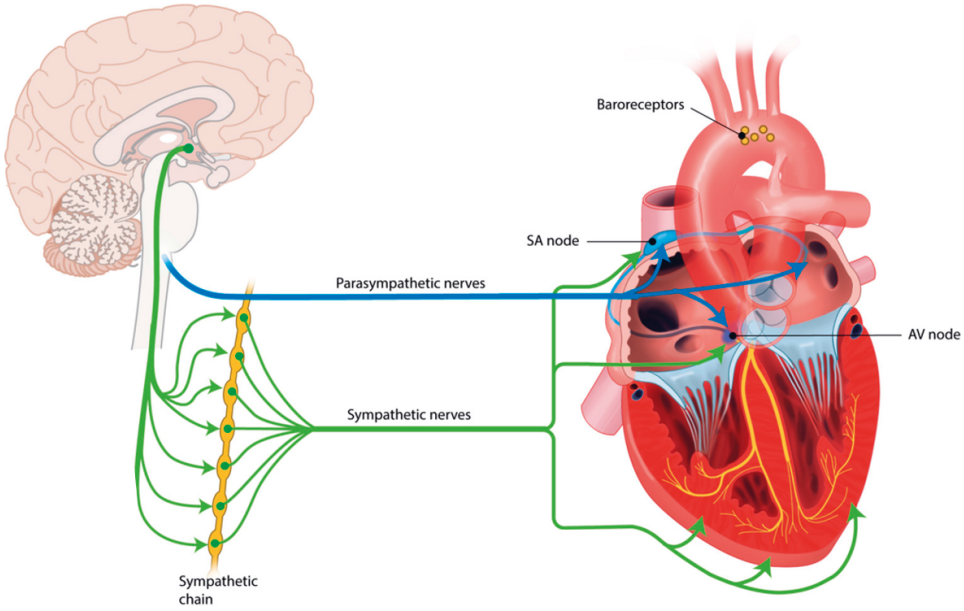

Figure 2.3 The interaction of the autonomic nervous system and the heart. Blood flow is indicated with arrows. Figure used under the CC BY 4.0 license, adapted from [19].

sympathetic, parasympathetic, and enteric nervous systems [20]. The sympathetic nervous system (SNS) can be thought to mainly drive the fight-or-flight response and the parasympathetic nervous system (PNS) the rest-and-digest response. The enteric nervous system (ENS) is a relatively independent branch regulating the gastrointestinal track. The reader should refer to the vast literature for more detailed information [16]. The cardiorespiratory functions are primarily influenced by the interplay between SNS and PNS that affect the rate and force of the contraction. The interaction between the ANS and the heart is illustrated in Figure 2.3.

Activity in the SNS increases heart rate and respiratory rate, while decreasing heart rate variability (HRV), and causes, e.g., vasoconstriction on the skin decreasing peripheral skin temperature [16]. Sympathetic activity is often linked with stress responses; it is not only necessary to induce the fight-or-flight response in case of emergencies but also to improve performance in non-threatening situations.

The dominance of the PNS relates to relaxation. It decreases the heart rate and respiratory rate, allowing increased heart rate variability [16]. The vagus nerve is the primary parasympathetic nerve and delivers information across the body to the central nervous system [16]. Several studies have implied that the proper functioning

and the activation of the vagus nerve decreases cardiac risks [21]–[23]. PNS activity can be promoted by deep inhalation and several commercial health monitoring devices incorporate guided breathing exercises.

Typically, target organs are affected by both sympathetic and parasympathetic nerves, including the heart and respiratory system [6]. As explained above, in these organs the SNS and PNS have the opposite effect with respect to each other. While the activity of one branch does not entirely exclude that of the other, the ratio of the activity in the two branches is referred to as the sympathovagal balance. Many chronic diseases are linked with autonomic dysfunction, contributing to decline in HRQoL [24].

2.2 Chronic diseases

Chronic diseases typically develop when risk factors actualize as gradual changes in physiology, eventually causing pathological changes in tissues. Chronic diseases inflict a continuous burden of symptoms on to the patient, typically impairing their health-related quality of life (HRQoL) and reducing working ability [25], [26]. In many cases, accelerated functional decline weakens the patient's ability to manage activities of daily living (ADL), such as dressing, eating, walking, and bathing [27]. Chronic diseases are persistent conditions that may require frequent medical attention, such as control visits to monitor the disease progression.

This work focuses on sleep apnoea, Parkinson's disease (PD), Huntington's disease (HD), inflammatory bowel disease (IBD), primary Sjögren's disease (PSS), rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), and CVDs. These diseases and the related symptoms can be devastating to ADLs and the quality of life. For instance, PD, IBD, RA, and sleep apnoea are all examples of chronic diseases linked with physical and mental fatigue and daytime sleepiness, which are known to cause functional impairment and hence deteriorate the HRQoL [28]–[30]. Due to population aging and increasing obesity, the prevalence of chronic diseases is expected to rise [2], [31]. Chronic diseases may generally increase the risk of death, either directly or via comorbidities or increasing the risk of incidents, such as falls.

2.2.1 Sleep apnoea

Sleep apnoea causes breathing cessations during sleep. It can be due to physical obstruction from other tissues in the airways (obstructive sleep apnoea), neural malfunction (central sleep apnoea), or both (complex sleep apnoea) [32]. In obstructive sleep apnoea, a partial obstruction causes hypopnoea, abnormally shallow breathing, and the severity of sleep apnoea is typically measured via the number of apnoeas and hypopnoeas per hour (apnoea-hypopnoea index) [33]. Individual cessations may last 10 seconds or longer, repeating 300–500 times per night [16]. The symptoms may include headache, mood swings, problems focusing, and daytime sleepiness [33]. Additionally, the patient may snore and experience repeated interruptions to sleep and restless sleep [33]. Moreover, sleep apnoea is considered a risk factor for developing cardiovascular diseases [34].

Sleep apnoea is typically diagnosed with an overnight polysomnography (PSG), which includes for instance electroencephalogram (EEG), electromyogram (measuring muscle activations), and nasal cannula for sleep and respiratory monitoring [33]. While sleep apnoea cannot be strictly cured, it can be efficiently treated, for instance, by treating physical obstructions to breathing (e.g. via losing weight or tonsillectomy), reducing alcohol consumption and smoking, or using a continuous positive airway pressure (CPAP) mask while sleeping.

2.2.2 Neurodegenerative diseases

Neurodegenerative diseases such as Parkinson's disease and Huntington's disease cause brain cells to gradually deteriorate and die. They yield gradually progressive symptoms that eventually destroy the patient's ability to control their own body [35]. There is currently no known cure for these diseases and the best results are achieved through early diagnosis and disease state monitoring, which enable timely interventions and treatment [36]. The disease progression and symptoms may be controlled via medication, deep brain stimulation surgery, and supportive therapies (such as physical or speech therapy) [35].

PD symptoms include bradykinesia (slowness of movement), tremor, muscle stiffness, cognitive decline, and problems with balance and sleep [35], [37]. PD causes dysfunction of the autonomous nervous system [37]. It induces annual costs exceeding $50 billion in the United States alone [38]. In 2019, neurological disorders in-

cluding PD and dementias affected 2,660 million individuals resulting in 97.7 million years of life lost or lived with disability (disability-adjusted life-years, DALYs) [3].

HD is often hereditary, and the symptoms may include chorea (uncontrolled movements), dystonia (repetitive movements, abnormal postures), problems with balance and movements, as well as cognitive and behavioural changes. As the disease progresses, the medical costs increase, too [39].

2.2.3 Immune-mediated inflammatory diseases

Immune-mediated inflammatory diseases such as inflammatory bowel disease, primary Sjögren's syndrome, rheumatoid arthritis, and systemic lupus erythematosus relate to dysregulated immune responses.

In inflammatory bowel disease the gastrointestinal tract becomes chronically inflamed. The IBD is specified as either Crohn's disease or ulcerative colitis, depending on the location of the inflamed tissue and the affected tissue layers [40]. The damaged tissue may cause abdominal pain, weight loss, bloody stool, diarrhea, and fatigue [40], [41]. These symptoms can restrict the patient's life severely and decrease working ability [40]. IBD symptoms may be treated with medication, ostomy (a surgically created exit point for secretions), and/or by surgically removing parts of the gastrointestinal tract [40].

Primary Sjögren's syndrome is an autoimmune disease where the body's own immune system damages glands in the eyes and mouth [42]. In addition to extremely dry eyes and mouth, the symptoms may include cough, problems eating and talking, tooth decay, dry skin, muscle and joint pain, and fatigue [42], [43]. The symptoms may be alleviated through medication (e.g. eye drops, drugs that increase saliva production, pain medication) or physically blocking tear ducts either with punctal plugs or through surgery [42], [43].

Rheumatoid arthritis is another autoimmune disease; it causes inflammation in the lining of the joints, which can gradually lead to bone erosion and joint disfigurement [44]. The symptoms may also include joint stiffness, weight loss, fever, and fatigue [44]. Medication in an early disease state may slow down disease progression and prevent joint deformities [45]. Additionally, the symptoms may be treated via pain medication, physical therapy, or surgery [45].

Systemic lupus erythematosus is a widespread autoimmune disease that may induce inflammation several organs including for instance the heart, lungs, kidneys,

skin, joints, and the nervous system [46]. The symptoms may include similar symptoms as RA, in addition to hair loss, malar rash (extending across the face), scaly rash, sensitivity sun light exposure, abdominal pain, and headaches [47]. SLE may develop comorbid diseases such as osteoporosis, diabetes, and cardiovascular diseases [46]. The treatment aims to reduce and control the symptoms and may consist of medication (pain medication, hydroxychloroquine) and treatment and prevention of the comorbid diseases [48].

2.2.4 Cardiovascular diseases

Cardiovascular diseases comprise the deadliest group of diseases globally, causing the death of over 18 million individuals every year [2], [3], [49]. CVD incidence increased by 24.6% over the previous decade (2010-2019), to an alarming 55.5 million cases a year [3]. Most CVDs can be preventable and treatable but require preventive and predictive methods, accessible to all [50].

CVDs encompass for example coronary artery disease, peripheral arterial disease, cardiomyopathy, cardiac dysrhythmias, rheumatic heart disease, among others. The symptoms vary from disease to disease but may include pain, dizziness, running out of breath, palpitations, and fatigue [51]. In some cases, the first identified symptom may be a heart attack. CVDs may be treated with medications, a pacemaker, or otherwise surgically, for example with a transcatheter aortic valve implantation or percutaneous coronary intervention (angioplasty).

3 METHODOLOGY AND PRIOR WORK

This chapter describes the data collection methods and analytics used in this thesis. Section 3.1 outlines the applied physiological measurements and the data collection methods. Sections 3.2 and 3.3 detail the signal processing and machine learning methods employed in this thesis, and section 3.4 describes the algorithm performance evaluation metrics. Section 3.5 summarizes the methods used in each substudy. Finally, section 3.6 presents a literature review summarizing prior work.

3.1 Health sensing and health records

Clinical care generates large volumes of data. Inpatients staying at hospital wards, including but not limited to emergency and intensive care patients, may be continuously monitored to detect sudden shifts in their condition. The medical state of outpatients, on the other hand, may be very sparsely monitored with occasional measurements, and questionnaires which capture patient reported outcomes (PROs), such as daytime sleepiness, reported on a predefined scale. Currently, clinical patient data is widely recorded in EHRs, which may at best record a comprehensive description of all healthcare services that pertain to a specific patient; hospital visits, diagnoses, procedures, measurements, medications, and other activities.

Healthcare facilities routinely use 12-lead electrocardiography (ECG), photoplethysmography (PPG) from the fingertip, thermometers, and cuff-based blood pressure monitors to monitor the patient's vital signs. Patients may also be sent home with ambulatory devices such as a Holter monitor or PSG equipment to collect monitoring data over a full 24 hour period or overnight, respectively. Modern clinical patient monitoring devices can track heart rate, blood oxygen saturation (SpO2), temperature, and respiration (exhaled CO2) but may require finger clips, electrodes, nasal cannula, or other devices attached to the patient.

Smart wearable devices such as smart watches and rings can be used to monitor

physiological signals continuously and are already widely used to quantify physical workouts and sleep. Commercial wearable devices often rely on PPG, requiring a contact with the skin, risking eczema in some users. Unobtrusive sensing technologies such as radars and ballistocardiography (BCG) on the other hand do not require skin contact but restrict monitoring to conditions where the subject is mainly still, e.g., in bed or driving a car. These technologies can facilitate the measurement of physiological signals including heart rate, heart rate variability, and respiratory rate on daily basis, and even continuously.

The safety, quality, and performance of all medical devices are ensured through regulations and standards. Medical devices need to conform to requirements set by the local regulatory authority, such as the Food and Drug Administration (FDA) in the United States or the European Commission in Europe. It is noted that these regulations apply to commercial devices intended for medical use, excluding many common commercial wearables.

This section first describes the main physiological measures used to quantify cardiorespiratory functions in subsection 3.1.1 and secondly, the related different means of data collection used in this thesis in subsections 3.1.2, 3.1.3 and 3.1.4.

3.1.1 Physiological measurements

Heart rate (HR) is a typical physiological measurement recorded by consumer devices as well as clinical devices. It can promptly indicate acute changes in the health, but it is also commonly used to support physical training. Heart rate monitoring technology has been suggested for maximal oxygen consumption (VO_2max) prediction and therefore monitoring long-term changes in cardiorespiratory fitness in free-living settings [52], [53]. Furthermore, heart rate and heart rate variability based approaches play a major role in sleep and recovery assessment [54]–[57]. Heart rate measurement in non-clinical devices may be based on, at least, ECG, PPG, BCG, image-based or radar-based technologies.

The time between individual heart beats is referred to as R-to-R interval when the R peaks are obtained via ECG, or interbeat interval (IBI) otherwise. HRV measures the variation between consecutive intervals. HRV measurements can be divided into long-term ($\geq$24h), short-term (about 5 min) and ultra-short-term (1 to <5 min) measurements [58]. Long-term and short-term measurements are frequently studied as they are associated with many health outcomes [59], [60]. However, the length

of the measurement period determines which physiological processes the measurements actually reflect and, hence, long-term and short-term measurements cannot be used interchangeably [58]. Moreover, HRV can be described by several different parameters, typically divided into time domain, frequency domain, non-linear, and other groups of parameters depending on the underlying assumptions for modeling [61], [62]. The ratio of low frequency to high frequency content (LF/HF) is currently considered to best reflect the sympathovagal balance and is used by some devices to detect sympathetic dominance, although this interpretation has also been challenged [58]. A more frequently available metric is the root mean square of successive differences (RMSSD) between beats. As a time domain feature, it is more easily obtainable from short-term or ultra-short term measurements. RMSSD can be satisfactorily obtained even from low sampling frequency ECG, at 50 Hz [63].

Respiration creates large periodic movements in the body that occur much less frequently than heart beats. Respiration can therefore be relatively easy to measure but it is also susceptible to artefacts. In Study I, a respiratory inductance plethysmography (RIP) belt was used to collect reference data. It's inductance changes with the circumference around the thorax as air flows in and out of the lungs, although physical contact with the surroundings, like rubbing against a bed mattress, may cause artefacts. The respiratory cycle also naturally affects the heart rate, causing its increase during inhalation and decrease during exhalation. This makes respiration rate highly available as a derived measure from a variety of heart monitors. Increased average respiration rate at rest can be a sign of an illness or a stressful event. Respiratory sensing has been studied for localization of trapped victims (e.g. due to earthquakes), apnoea detection, sleep stage classification, emotion recognition, and athletics [64].

Skin temperature normally exhibits patterns related to the circadian rhythm and the hormonal cycle in women. Skin temperature has been associated with acute stress reactions and it has been applied in, e.g., sleep-wake classification, fever detection, and menstruation prediction [65]. As compared to e.g. HR, skin temperature shows slower variation [66]. However, the measurement point for skin temperature determines which physiological phenomena it can capture. The temperature measured from the chest, near the heart, is not expected to vary as rapidly as from fingers and correlates better with the core temperature. In hospitals, core temperature tends to be the primary focus of interest, although more difficult to measure.

3.1.2 Frequency modulated continuous wave radar

A frequency modulated continuous wave (FMCW) radar operates on a predefined carrier frequency, transmitting a continuous millimeter wave signal at a continuously changing frequency, sweeping across a frequency band [67]. When the transmitted signal collides with a surface (here, the skin), it is partly reflected back to the radar. The frequency of the received signal corresponds to the original transmit frequency and hence indicates the original transmit time. The distance to the detected surface can be calculated from the beat frequency; the absolute value of the difference between the transmit and received signal frequencies, shown in Figure 3.1. The set of observed distances, i.e., the complex range profile, can be extracted from the set of beat signals of a single frequency sweep by applying the Fast Fourier Transform (FFT). The complex FFT results can be further processed to extract the beat signal amplitude and phase for a specific distance, i.e, a range bin [67].

The thesis uses an FMCW developed at VTT Technical Research Centre of Finland Ltd, which operates at a carrier frequency of 24 GHz and has a bandwidth of 250 MHz [67]. It is able to detect micromotions below 1 µm, and it has a range resolution of 60 cm and an adjustable sampling frequency.

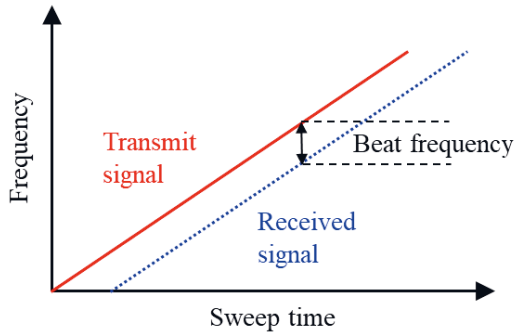

Figure 3.1 The operating principle of frequency modulated continuous wave radar [68]. ©2016 IEEE

3.1.3 Wearable sensors

An increasing amount of wearable and other non-intrusive devices for activity and health tracking started to flow to the markets around 2010 [69]. The broad uptake

of wearables among the general public came with the concept of "quantified self", a new trend of tracking one's own health. The high availability of wearable devices enabled individuals with means to monitor their health and well-being, simultaneously increasing awareness of the physiological phenomena in their own body. The global market for just wearable sensors, which reached USD 61.3 billion in 2022, is forecasted to exceed USD 186 billion by 2030 [1].

Self-tracking devices actively collect invaluable, even round-the-clock data about the user's health. Modern sensor technologies achieve accuracy between 1.1% to 6.7% mean absolute percentage error for heart rate at different activity intensities in controlled settings [70], [71]. They are able to use adequately high sampling frequencies whilst maintaining a long-lasting battery life, covering up to several days.

VitalPatch, as used in this thesis, is a patch-like wearable sensor that measures ECG, heart rate, R-to-R interval, respiratory rate, skin temperature, step count, posture, and more [72]. It is a wireless disposable biosensor with a battery life up to seven days. VitalPatch has a CE (class IIa medical device) and an FDA certification. It records ECG at 125 Hz sampling frequency and ECG-derived HR, R-to-R interval, and respiratory rate at 0.25 Hz frequency. Skin temperature is recorded similarly at 0.25 Hz, and step count and posture at 1 Hz frequency. The encrypted data is transferred through a wireless connection to a cloud-based platform, which also offers a user interface for clinicians. The device can store up to 10 hours of data if the connection is interrupted and upload the data upon re-connection.

VitalPatch has demonstrated a mean absolute error of 0.72 beats per minute (bpm) for HR and 1.89 respirations per minute (rpm) for respiratory rate in healthy subjects, when compared to gold standard heart rate from 3-lead ECG and respiratory rate from nasal cannula [72]. Heart rate was found accurate despite movement and induced hypoxia, while respiratory rate was predominantly underestimated and sensitive to artefacts from low frequency periodic movement [72]. The VitalPatch also estimates core temperature and has demonstrated mean absolute errors of 0.29–0.42 °C as compared to an oral and a swallowed pill-like thermometer [73].

3.1.4 Electronic health records

Electronic health records are the modern standard of clinical data storage. They can record the patients' full history in a digital form, including patient outcomes and care paths. EHRs can contain diverse data including lab results, procedure details,

prescriptions, hospital visits, clinical notes, and so on. They can cover millions of patients over long time periods, thus encompassing major promise for predictive applications. However, the data tend to be incomplete and even erroneous due to human error, hence requiring special attention to handling missing data and poor data quality. Additionally, EHR data are sparse and heterogeneous by nature, i.e, the records can include numerical data, images, and text, all recorded in an episodic manner. This further complicates the application of data-driven technologies. Because EHRs are recorded by healthcare providers, the data may also exhibit temporal patterns reflecting, for instance, changes in the practical care guidelines or effects of a major public health disruption, such as the peak in mortality observed due to COVID-19.

Although EHRs are widely adopted around the world, different healthcare providers may use different standards. Interoperability between different healthcare providers nationally and internationally, without risking data privacy, could enable maximal input data volume and diversity to EHR based algorithms and wider uptake [74]. A larger pool of data is more likely to include more observations of rare conditions and generally more varying demographic and other background factors, and data collection across systems, institutions, and countries may be crucial to obtain sufficient volume and variability for algorithm development [74]. When applying an algorithm in a new EHR system, a different data standard may lead to erroneous analysis results [74].

This thesis uses a combination of three registries: a hospital district EHR, a registry specialized in cardiac patients, and a national mortality registry. All three registries were collected in Finland and have been validated for cardiovascular diseases such as strokes, coronary heart disease and heart failure [75]–[77]. The combined database was automatically collected in a previous retrospective registry study, Mass Data in Detection and Prevention of Serious Adverse Events in Cardiovascular Disease (MADDEC) [78].

3.2 Signal processing

Signal processing is a broad field of engineering, comprising a spectrum of methodology aimed to alter and analyze signals. Digital signal processing (DSP) specifically focuses on digital signals and computational processing methods. This work employs signal processing and machine learning methods for detecting and predicting health

related events across various time spans. The time span of interest for each sub-study affected the selected methodology.

3.2.1 Time domain analysis

Time domain analysis investigates signal properties and behaviour with respect to time. This contains simple statistical metrics, like the mean and standard deviation, median, different percentiles, and the minimum and the maximum values observed over specific intervals, which can be used to summarize the time series. For instance, resting HR equals to the minimum observed HR when the person is at rest/sleeping. In this work, heart rate recovery was defined as the post-exertion decrease in HR (difference between observed maximum and minimum), measured over one minute after at least 6 min of walking.

The five least active hours of a day (L5) can be used to represent rest. In continuous measurements, the L5 window can be located by selecting the window with highest percentage of laying down. The selected L5 was allowed to contain up to 100 recorded steps and was required a minimum of 80 % of laying down.

Time domain analysis also comprises non-statistical analysis, such as peak detection or the analysis of signal coverage. In this work, signal coverage for frequently sampled signals was defined as the ratio between the number of observations and the number of expected observations. For irregular signals, such as IBI, coverage was defined as the sum of IBIs divided by the observed period of time.

3.2.2 Spectral and cepstral analysis

In spectral analysis, the time domain signal is transformed to the frequency domain, e.g. using a Fourier transform, and analyzed therein. FFT is computationally efficient algorithm for discrete Fourier transforms (DFT). For a signal $\bar{x}$ of length N, the DFT is defined by

$$X_k = \Sigma_{n=0}^{N-1} x_n e^{-2\pi ikn/N}, \tag{3.1}$$

where $k = 0, ..., N-1$ and i is the imaginary unit. The inverse DFT is defined by

$$x_n = \frac{1}{N} \Sigma_{n=0}^{N-1} X_k e^{2\pi ikn/N}. \tag{3.2}$$

The DC component X_0 equals to the mean of the signal.

Cepstral analysis is a variant of spectral analysis, where a logarithmic transformation is applied on the frequency domain representation to emphasize small frequencies. This is followed by the inverse Fourier transform which brings the representation to the so called *quefrency* domain (instead of the time domain), and the real part of the inverse transform compose the *cepstrum*.

For a non-stationary process like the heart beat, the power spectral density (PSD) can be used to estimate the frequency content. In Welch's method, the signal is divided into overlapping segments and the PSD is obtained as the average of the modified periodograms calculated for each segment [79].

3.2.3 Autocorrelation

Autocorrelation describes a signal's correlation with lagged versions of itself and can indicate periodic characterictics. The autocorrelation function (ACF) is defined for $k < n$ as

$$R_k = \frac{\Sigma_{i=1}^{n-k}(s_i - \mu)(s_{i+k} - \mu)}{\Sigma_{i=1}^{n}(s_i - \mu)^2}, \tag{3.3}$$

where s_i is a sequence at lag i and μ is the signal mean.

3.2.4 Heart rate variability analysis

In the presented work, heart rate variability analysis primarily inspects time and frequency domain features, and additionally geometrical and non-linear features in Study II. The frequency domain features are based on the PSD estimation (see subsection 3.2.2) The full set of included HRV features is presented in Table 3.1.

It should be noted that the HRV features are based on normal-to-normal peak intervals (NN), which improve the analysis as compared to unprocessed IBI [61], [80]. NN are obtained by removing abnormal beats from the IBI (or R-to-R interval) data and replacing them by the means of linear interpolation. The Malik method was employed to remove abnormal beats, that is, IBI deviating over 20 % from the previous interval were considered abnormal. For ECG, however, a limit of 15 % was applied instead.

Table 3.1 Heart rate variability features.

Type	Feature	Unit	Description
Time	NN mean	ms	Mean of normal-to-normal peak intervals (NN)
	NN CV	ms	Coefficient of variation of NN
	NN SD	ms	Standard deviation of NN
	NN median	ms	Median of NN
	NN range	ms	Difference between maximum and minimum of NN
	RMSSD	ms	Root mean square of consecutive differences in adjacent NN
	CVSD	ms	Coefficient of variation of consecutive differences in adjacent NN
	SDSD	ms	Standard deviation of consecutive differences in adjacent NN
	NN50		Number of interval differences greater than 50 ms
	NN20		Number of interval differences greater than 20 ms
	pNN50	%	Percentage of interval differences greater than 50 ms
	pNN20	%	Percentage of interval differences greater than 20 ms
	HR mean	bpm	Heart rate mean
	HR SD	bpm	Heart rate standard deviation
	HR min	bpm	Heart rate minimum
	HR max	bpm	Heart rate maximum
Frequency	VLF	ms^2	Power spectral density in very low frequencies (0.003–0.04 Hz)
	LF	ms^2	Power spectral density in low frequencies (0.04–0.15 Hz)
	HF	ms^2	Power spectral density in high frequencies (0.15–0.40 Hz)
	Total power	ms^2	Total power spectral density; sum of VLF, LF, and HF
	LF/HF	%	The ratio of LF and HF
	LFnu		LF normalized to the sum of LF and HF
	HFnu		HF normalized to the sum of LF and HF
Geometrical	Triangular index		Number of all NN divided by the maximum of the NN density distribution
Non-linear	CSI		Cardiac sympathetic index
	mCSI		Modified cardiac sympathetic index
	CVI		Cardiac vagal index
	SD1	ms	Poincaré plot, SD1
	SD2	ms	Poincaré plot, SD2
	SD2/SD1	%	SD2 to SD1 ratio

3.2.5 Feature normalization

Some methods and applications may benefit from normalized input features. For example, physiological measurements can show significant inter-individual differences as they are affected by the individual's age, sex, and other variables [81]. Normalization can mitigate the effect from inter-individual variance when investigating physiological data for generalizable patterns. Additionally, many ML methods tend to perform better when input features are scaled to a unified, confined range, as a feature with a broad numerical range might otherwise dominate over features with narrow range when fitting the model.

Typical methods for normalization include, e.g., standardization (z score normalization) and min max scaling. In ML, the normalization parameters are commonly derived from the training data. For physiological data, normalization with respect to the individual's typical measurement values may provide more meaningful features and improve generalization. To normalize data with respect to one's resting parameters, standardization parameters (mean and standard deviation) can be calculated from the L5 data. Hence, L5 normalized data is achieved by taking the difference to the L5 mean and by dividing by the L5 standard deviation.

3.3 Machine learning

Machine learning algorithms learn dependencies from existing evidence by applying optimization and statistical tools. An ML algorithm uses a set of prior observations, the training data, to create a data-driven model. ML methods can be coarsely divided into supervised learning, unsupervised learning, and reinforcement learning. A suitable learning paradigm is selected based on the available training data. Whereas supervised learning methods require a well defined target output for each training example, unsupervised learning methods look for inherent dependencies and reinforcement learning uses rewarding and punishing functions to guide the learning process [82]. The presented work focuses on supervised learning.

The performance of an ML model on previously unobserved data (generalizability) is highly dependent on the training data. To enable a generalizable model, the data should comprise a representative sample that is diverse and sufficiently large, with high quality while still representing the real life use case. The generalizability

of a model can be tested by with-holding a part of the example data as a test set, which remains unseen by the model during the development phase. Upon development, the development data can be divided into training and validation sets to maximize generalizability from the training set to the validation set. Cross-validation methods iteratively change the training and validation set composition, which can help select optimal model hyperparameters [83].

3.3.1 Artificial neural networks

Artificial neural networks (ANN) are a subset of ML models that mimic the biological neural cells networks. An ANN comprises an input layer, at least one hidden layer, and an output layer. Each layer may include a number of neurons which are connected to the neurons on the other layers. Each neuron in a hidden layer takes a weighted sum of inputs (originating from the previous layer) and a bias term, and passes it through an activation function that is typically non-linear and enables the ANN to implement complex dependencies. In deep learning, the ANN includes multiple hidden layers. However, ANNs are generally *black boxes*, i.e., how the model arrived to a specific output cannot be explained as the underlying math becomes too complicated to humanely comprehend (more in subsection 3.6.4).

ANNs can be trained using, e.g., backpropagation: a set of samples are first passed through the network (forward pass), then the error to the target output is computed and changes to the weights are propagated backwards to each neuron one-by-one (backward pass). The data is typically passed in smaller batches over several iterations, which adjusts the magnitude of the applied weight updates. An ANN is trained over several epochs. A single epoch means one round of backpropagating over all batches. It should be noted that instead of finding the global minimum, the training process may get stuck in a local minimum. This can be battled through, e.g, the careful selection of activation function, error optimizer, batch size, and learning rate [84]. The training may also yield a model that has overfitted to the training data and generalizes poorly. Overfitting may be reduced, e.g, via early stopping (less training epochs) or by applying dropout (randomly disconnecting neurons) [85].

Specialized ANNs with specialized neurons and layer architectures have been developed to handle different types of inputs. Very deep ANNs are nowadays pretrained on huge amounts of data and later fine-tuned to solve specific problems [86]. This work employs convolutional and transformer neural networks.

3.3.2 Convolutional neural networks

Convolutional neural networks (CNN) are specialized in spatial data, such as images. A convolutional layer convolves the input with sliding kernels and applies the non-linear activation function to each convolution output, producing a set of feature maps. Convolutional layers are typically coupled with pooling layers, which simply shrink the input (without trainable parameters) to improve spatial invariance. For example, maxpooling only keeps the maximum value of each smaller block of the input. The spatial invariance makes CNNs optimal for spatial pattern recognition.

The number of trainable parameters in a convolutional layer depends on the kernel size and the number of kernels. Typically the kernel size is small, enabling deep CNNs. It is argued that in deep architectures, the first convolutional layers learn simple patterns and the deeper layers building on them learn more complex patterns.

CNNs are extensively used in image processing including tasks like the classification of hand-written digits or animal species, object detection in traffic images, and magnetic resonance image segmentation. The most popular architectures, like the AlexNet, VGG or U-Net, typically achieved the state-of-the-art results in their task [87]–[89]. With image input, a CNN is typically two-dimensional with squared $n \times n$ kernels. A one-dimensional CNN (1D CNN) with $1 \times n$ kernels can be applied for spatial pattern recognition in time series. The architecture of the model is illustrated in Figure 3.2.

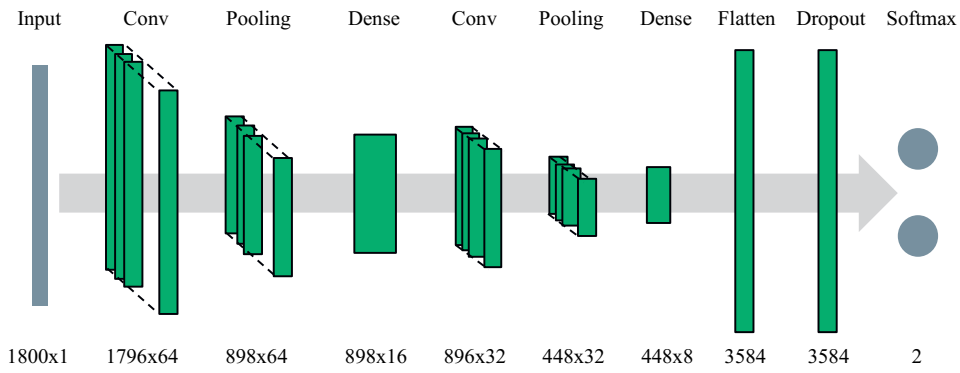

Figure 3.2 The 1D CNN architecture applied in Study III. The model consisted of 10,426 trainable parameters.

3.3.3 Transformer neural networks

Transformer neural networks specialize in sequential input data, such as written natural language. Inspired by the advancements emerging through attention, Vaswani et al. built the network architecture exclusively on attention mechanisms [90]. Transformers learn long-range bidirectional dependencies while outperforming previous approaches in training time [90].

The attention mechanism applies weights to the input sequence elements and runs the resulting sequence through a softmax function, converting the sequence to weights that sum up to one, also known as attention scores. Originally, attention was developed for sequence-to-sequence autoencoder networks, which aimed to transform one sequence to another but performed poorly on long sequences [91]. In self-attention applied by transformers, attention is used to find dependencies between different positions within one sequence [92].

A layer in a transformer neural network consists of two sub-layers: a multi-head attention layer and a feed-forward layer [90]. Multi-head attention is able to apply multiple attention functions in parallel, improving computational efficiency. Because the full sequence is available for learning the attention weights, transformer models learn bi-directional dependencies.

The Transformer inspired other transformer variations. The most popular variant is the Bidirectional Encoder Representations from Transformers (BERT), which only employs the encoder part of the original Transformer [93]. The encoder is trained by masking random elements of the training data and letting the model predict the masked input, and fine-tuned to a specific task [93]. Another variant, XLNet, surpassed the performance of BERT in several tasks [94]. As opposed to BERT, XLNet is an autoregressive model and learns by maximizing the expected likelihood over all permutations of the factorization order [94].

The attention weights of the pre-trained models may be visualized for specific input sequences [94]. This has inspired hope of finding explainability or interpretability to the inner workings of transformers. The presented work utilizes a visualization tool, BertViz, to inspect attention weights of fine-tuned models [95].

3.4 Performance evaluation

Algorithm performance evaluation may consist of diverse metrics. A descriptive set of performance metrics is selected based on the type of the problem (e.g. regression, classification) and captures the possible performance gaps by considering different view-points and characteristics of the underlying data, like class imbalance, participant cohorts, or varying measurement conditions. In regression problems, the algorithm output is a continuous variable and the performance metrics may be statistical or based on the distance from the true value. In classification problems, the algorithm output is discrete. The presented work focuses on binary classification, and thus multi-class performance evaluation is not discussed here.

3.4.1 Regression performance

Mean absolute error (MAE) describes the average deviation from the ground truth value and is defined over N samples as

$$MAE = \frac{1}{N}\Sigma_{i=1}^{N} \left|x_i - y_i\right|, \tag{3.4}$$

where x_i is a single value to evaluate against the corresponding reference value y_i, and $i = 1, ..., N$.

Root mean squared error (RMSE) similarly describes deviation from the ground truth but emphasizes individual large errors. It is especially useful to evaluate when even individual large errors are unacceptable. RMSE is defined as

$$RMSE = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N} \left(x_i - y_i\right)^2}, \tag{3.5}$$

where x_i is a single value to evaluate against the corresponding reference value y_i.

Various correlation metrics also describe the agreement between two continuous variables. The Pearson correlation coefficient r measures linear correlation between two variables and gives a value between -1 (perfect negative correlation) and 1 (perfect positive correlation), where zero corresponds to no correlation. For N samples, it

is defined as the ratio of covariance and the multiplied standard deviations

$$r = \frac{\Sigma_{i=1}^{N} (x_i - \bar{x}) (y_i - \bar{y})}{\sigma_x \sigma_y}, \tag{3.6}$$

where $\bar{x}$ and $\bar{y}$ are the sample means for the two variables, and σ_x and σ_y are the standard deviations of x and y, respectively. Pearson correlation is, however, sensitive to outliers. The correlation coefficient is often inspected together with its p value, which originates from a statistical test with a null hypothesis H_0 that correlation between two variables is zero. The p value describes the probability that the observed sample was drawn from a population with no actual correlation, i.e., the probability that the null hypothesis would be falsely rejected. A predefined significance level α (typically 0.01 or 0.05 depending, for instance, on sample size) defines how small the p value must be before H_0 is rejected. It is additionally noted, that correlation coefficients beyond Pearson r may be selected when the relationship between two variables is not expected to be linear. For instance, Spearman correlation coefficient is non-parametric and measures the monotonicity of the relationship, rather than linearity.

The Pearson correlation assumes the independence of observation errors, which may be violated when multiple observations are obtained from each participant, in a study comprising multiple participants [96]. Repeated measures correlation r_{rm} similarly describes linear correlation in range [-1,1], but takes the repeated nature of the observations into consideration and describes the common within-individual correlation between the two variables [96]. It employs analysis of covariance (ANCOVA) to adjust for between-individual variability and is defined as

$$r_{rm} = \sqrt{\frac{SS_{measure}}{SS_{measure} + SS_{error}}}, \tag{3.7}$$

where the sign depends on that of the common slope across individuals, $SS_{measure}$ is the sum of squared differences between the regression outputs and the sample mean (regression sum of squares) and SS_{error} is the sum of squared differences between individual observations and the corresponding regression outputs (error sum of squares) as obtained from ANCOVA [96].

ANCOVA itself can be used to assess the statistical significance of the difference between the means of two or more groups while adjusting for the effect of covariates.

One-way ANCOVA models the relationship between the jth dependent variable of the ith group and the corresponding covariate observation x_{ij} as

$$y_{ij} = \mu + \tau_i + \beta(x_{ij} - \bar{x}) + \epsilon_{ij} \tag{3.8}$$

where μ is the observed mean, τ_i is the effect of the ith group, β is the regression slope, $\bar{x}$ is the average of covariate observations, and ϵ_{ij} is the error associated with the jth observation of the ith group [96]. The F test and its p value are typically used to assess the significance of difference between groups. If so, pairwise differences can be further examined via ad hoc tests like Tukey's method, which performs all pairwise comparisons while adjusting the p value for multiple comparisons. The effect size from the covariate(s) is often assessed using partial η^2.

The Kolmogorov-Smirnov (KS) test is a non-parametric test, which can be used to test the difference of two empirical samples with unknown underlying distribution.

The Bland-Altman plot vizualizes the difference between two continuous variables, typically two measurement of the same phenomenon, with respect to their average. Using the average instead of either variable alone avoids false indications of a systematic error proportional to the magnitude [97]. Both systematic and random error can be evaluated from the Bland-Altman plot.

3.4.2 Classification performance

Classification performance is often described via the confusion matrix or measures that can be derived from it. The confusion matrix, as depicted in Table 3.2, presents the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classifications, or predictions. Table 3.2 also denotes the number of positive (P) and negative (N) samples, as well as the number of predicted positive

Table 3.2 Confusion matrix

		Prediction output	
		PP	PN
Ground truth	P	TP	FN
	N	FP	TN

(PP) and predicted negative (PN) samples.

Accuracy, sensitivity, and specificity are defined

$$\text{accuracy} = \frac{TP + TN}{P + N} \tag{3.9}$$

$$\text{sensitivity} = \frac{TP}{P} \tag{3.10}$$

$$\text{specificity} = \frac{TN}{N}, \tag{3.11}$$

respectively. Accuracy describes how often the classification or prediction was correct, whereas sensitivity and specificity describe the accuracy for positive and negative classes, respectively. Together, sensitivity and specificity can capture how many actual positive and negative samples were correctly identified by the algorithm.

A binary classifier typically produces a value, which is interpreted as a positive or negative class according to some threshold. Adjusting the threshold can change the model sensitivity and specificity without changing the underlying model *per se*. The model performance at any threshold can be depicted in a plot of sensitivity with respect to 1-specificity, where a completely random model would be represented by a straight line from (0,0) to (1,1). An ideal model, on the other hand, would achieve the (0,1) point with perfect sensitivity and specificity. Area under the receiver operating characteristics curve (AUC) describes the area under the curve in the sensitivity to inverse-specificity plot and is sometimes preferred over accuracy, as it describes performance over a range of thresholds. An AUC of 0.5 describes a random classifier and an AUC of 1.0 would indicate a perfect classifier.

Sensitivity (also known as recall) may also be coupled with precision, defined as

$$\text{precision} = \frac{TP}{PP}. \tag{3.12}$$

Together, sensitivity and precision capture how many of the positive cases were identified but also how many of the produced positive classifications were correct.

3.5 Method summary

Table 3.3 summarizes the data collection and analysis methods in each study.

Table 3.3 Summary of methods

Study	Data collection method	Analysis method	Performance metrics
I	FMCW radar	Autocorrelation, cepstral analysis, HRV analysis	MAE, RMSE Pearson correlation, Bland-Altman plot
II	Wearable ECG sensor	Time domain analysis (statistical analysis, HRR, coverage) HRV analysis	Repeated measures correlation ANCOVA KS
III	Wearable ECG sensor	1D CNN	Accuracy, sensitivity specificity
IV	EHR	Transformers (BERT, XLNet)	AUC, sensitivity precision

3.6 Prior work

Real-time health event monitoring primarily requires high accuracy measurement capabilities, whereas FMCW radars can be sensitive to motion artefacts. Continuous monitoring with wearables faces different challenges in different measurement settings. EHRs can become massive and hard to utilize. This section reviews the literature and the current state-of-the-art results related to each of the case studies.

3.6.1 Real-time health event detection with FMCW radars

Radars in healthcare monitoring have been primarily studied for applications where the measurement equipment can be rather stationary, such as monitoring seated or sleeping individuals at home, offices, or care facilities [98], [99]. They have also been used for locating live earth-quake victims behind collapsed concrete walls via respiratory rate detection, where the detected individual is again stationary [100]. Inpatients staying at healthcare facilities or nursing home residents often stay in bed for long periods of time and could benefit from the contact-free monitoring which radars can offer.

Prior studies have validated heart rate and respiratory rate monitoring with FMCW radars against certified medical devices but restricted their experiments on participants sitting up. Wang et al. included 10 participants in their study and vali-

dated their heart rate and respiratory rate extraction method against a medical patient monitoring device [101]. In their experiments the participants were sitting still, and an 80 GHz radar was positioned 1 m in front, behind, and to the left side of the participant, sequentially. The positioning in front of the participant was additionally repeated with 2 m distance. They reported the best combined result from the 1 min frontal measurement setting with 8.1 % and 6.9 % relative errors for heart rate and respiratory rate, respectively [101]. Similarly, Adib et al. studied participants in seated positions but aimed to monitor 2–3 individuals simultaneously, given a distance of 1–2 m between two individuals [99]. Their study included 14 participants seated in different orientations and at varying distances with respect to a radar mounted on a table. As compared to a medical device, they reported accuracies of 99% and 99.4% for heart rate and respiratory rate, respectively, when the participant was sitting [99].

Arsalan et al. focused on heart rate estimation from seated participants and used a commercial heart rate chest belt sensor as the reference device [102]. They proposed adaptive band-pass filtering on the time-domain displacement signal, based on Kalman filtering to narrow down the frequency band [102]. They used a 60 GHz FMCW radar on 14 individuals. The validation was however restricted to 11 heart beat estimation per individual, as all participants were measured for 20 s and they used a 10 s sliding window (1 s interval). They reported an RMSE of 5.3 bpm for 7 participants holding their breath during the measurement and 7.0 bpm for the remaining 7 breathing naturally [102].

Other studies have investigated FMCW radars for monitoring individuals in bed but lack validation with respect to certified medical devices. Anitori et al. had six participants lie down in four different positions on a bed [103]. They mounted four radars of different carrier frequencies between 2.4 GHz and 24 GHz to the ceiling above and, without seeing significant differences for heart rate and respiratory rate extraction, selected the 9.6 GHz radar for further investigation. They compared the accuracy of the radar technology against an ECG belt while the participant would either breath normally or hold their breath. They reported that their best method, a FFT on the phase signal showed up to 5% error in 55% of the measurements. Lim et al. also measured five participants while lying on a bed and wearing a commercial reference sensor [104]. Interestingly, the participants were lying on a surgical bed and a surgeon was moving beside bed in some of the measurements. They used a 60 GHz

FMCW radar and applied beam-forming to improve signal quality upon interference from a surgeon [104]. They reported 0.75–1.66 bpm and 2.1–4.7 bpm MAE (or over 90% and 93% accuracies) for respiratory rate and heart rate, respectively, in the presence of a surgeon during 100 s measurements [104].

Overall, FMCW radars have achieved encouraging results in vital sign monitoring. Nevertheless, their performance in lying positions has not been evaluated as compared to medical grade reference devices. Moreover, prior studies have not included wide ranges of both respiratory rates and heart rates, and thus the applicability of FMCW radars for sleep apnoea monitoring has remained an open question.

3.6.2 Continuous health monitoring with wearables

In contrast to radars, wearable sensors can be used to monitor an individual any time and anywhere, also during exercise. In exchange, wearables depend on rechargeable batteries and require skin contact. Even so, they have potential to produce high coverage continuous health data. Their feasibility for patient monitoring has been studied both for inpatient and outpatient applications [105], [106].

Weenk et al. studied the feasibility of two wearable devices for continuous patient monitoring on 20 patients at the surgical and internal medicine wards [105]. Both wearables were worn simultaneously for 2–3 days and measured ECG, heart rate, respiratory rate, and skin temperature [105]. They reported positive feedback from both patients and nurses. The nurses found the devices easy to attach whereas the patients reported increased sense of safety due to continuous monitoring and acknowledged the light weight and low interference with ADLs [105]. The wearables additionally reduced interruptions of sleep by allowing monitoring from a distance [105]. However, 20 % of the measurements observed by nurses were incomplete and accidental data deletions, data transfer failures, as well as movement artifacts and loosened skin contact reduced the coverage and quality of the collected data [105]. The study criticized the inaccuracy of respiratory rate measurements. The authors noted that using wearable sensors for in-patient monitoring creates high demand for automatic patient deterioration alarms but the risk of false alarms creating alarm-fatigue in nurses must be considered.

Integrating alarm systems to wearables could offer additional value for inpatient monitoring. Väliaho et al. validated a PPG based wrist-worn medical device against an ECG for atrial fibrillation detection in a study including 173 inpatients [107].

They reported high sensitivity and specificity (over 94 % and 98 %, respectively) for atrial fibrillation in detector time-frames ranging from 10 to 60 minutes. The patients found the wearable sensor more comfortable than the Holter ECG, with a statistically significant difference, and most patients would have preferred the wearable for measurements at home [107]. The study also reported low interference with ADLs and mobility.

Outpatient monitoring imposes a greater challenge as compared to inpatient monitoring due to the less controlled measurement environment. Free-living outpatients, such as home care clients, may continue their typical daily living, which may interfere with the continuous measurement. Interviews with healthcare providers, health information professionals, and providers of commercial remote monitoring solutions have indicated that the top concerns regarding remote patient monitoring were mainly about the lack of integration with electronic health records, with gaps in accessibility to professional interfaces and interpretable reports as well as interoperability between outputs from different sensors [108]. The measurement inaccuracies due to human or technical errors were also concerning to the expert interviewees [108]. In 2020, Soon et al. published a review covering 30 wearable devices in outpatient monitoring settings [106]. They reported a gap in the amount of clinical validation studies and presented that several devices lack peer-reviewed publications.

Despite the added challenge, remote monitoring has also demonstrated new promising opportunities. Sokas et al. created an automated walk-test mimicking a six minute walk test (6MWT), regularly used in controlled environments to assess cardiovascular fitness [109]. The study covered a total of 99 participants, including CVD patients and healthy volunteers. They demonstrated that detecting eligible (unguided) walking periods during activities of daily living with a commercial wrist-worn device can be feasible and exhibits different walking distances between CVD patients and healthy participants. Furthermore, Iqbal et al. reviewed the effect of wearable based alerting systems on clinical outcomes in remote monitoring [110]. They assessed hospitalization, all-cause mortality, length of stay, emergency department, and outpatient visits. They reported reduced length of stay and all-cause mortality thanks to the alerting systems, supporting the important role of early detection (of patient deterioration) for improving patient outcomes. However, the results regarding hospitalization were inconclusive, whereas no improvement was found in terms of emergency department and outpatient visits [110]. The inspected studies included

various alerting systems with some alerting the participant instead of healthcare professionals [110]. Most studies included in the review considered cardiorespiratory measures.

Importantly, continuous monitoring with wearable sensors could offer sensitive objective assessment of HRQoL. Currently, the HRQoL of a patient is primarily evaluated in healthcare via subjective patient reported outcomes, which are susceptible to recall bias. Prior works lack proposals for objective metrics that adequately describe HRQoL through physiological functions in uncontrolled free-living settings. Additionally, the feasibility of continuous monitoring has not been sufficiently addressed for physiological monitoring in chronic patients, especially in outpatients.

3.6.3 Predicting health event risks from EHRs

EHR applications are often simplified to only consume a small selection of data from the EHR, such as well defined diagnostic or other codes, continuous clinical measurements, or free text clinical notes [111]–[116]. Some studies have proposed data representations that can capture several types of inputs, even representations that could be applied to the entire EHR [117]–[121]. Regardless of the challenges, EHR data are seen full of potential for discovering new information about cardiovascular disease mechanisms, drug development, personalized medicine, risk prediction for decision support, among other benefits [122].

In the context of cardiovascular diseases, EHR data have been typically applied for some specific diagnostic cohort. Motwani et al. used a multi-site registry of 10,030 patients with suspected coronary artery disease and trained a boosted (decision tree based) ensemble classifier to predict 5-year all-cause mortality [123]. A total of 745 deaths occurred within the 5 years. They demonstrated that a combination of clinical and coronary computed tomography angiography features obtained an AUC of 0.79 and outperformed several individual clinical scores (AUC values 0.61–0.64). Notably, the results were obtained with 10-fold cross-validation without conducting tests with held-out data. The other individual scores were selected as input features in the model. Later, Hernesniemi et al. used a 9,066 registry from a single site and predicted 6-month mortality in patients with acute coronary syndrome [124]. A total of 660 patients died within 6 months. They evaluated a logistic regression model and an extreme gradient boosting model obtaining AUC values of 0.87 and 0.89, respectively, both exceeding the performance of the traditionally used Global

Registry of Acute Coronary Events (GRACE) score (AUC 0.82). Some of the ML model features overlapped with GRACE score components [124]. Their training data covered the years 2007-2014 and 2017, and the test data covered 2015-2016.

In addition to CVD patient-specific mortality risk prediction in the studies by Motwani et al. and Hernesniemi et al., mortality risk prediction has been studied in other patient groups. Rajkomar et al. combined three deep learning models (long short-term memory or LSTM, attention-based time aware neural network, and boosted time-based decision stumps network) and predicted in-patient mortality within 24h of admission with AUCs of 0.93-0.95 at two sites [119]. They used EHRs from two hospitals in the USA, including a total of 114,004 patients. Mortality studies have also been especially applied for intensive care unit (ICU) patients. Choi et al. studied mortality risk prediction during ICU admission using a graph convolutional transformer model and achieved an AUC of 0.60 [120]. Their data were highly imbalanced and covered multiple sites in the USA over two years. Yang et al., on the other hand, reported an AUC of 0.85 for mortality risk prediction among ICU patients [125]. They used the codes representing diagnoses, treatments, interventions, and diagnosis related groups, as well as hospitalization type and admission and discharge times. They applied an LSTM classifier with attention mechanisms in two levels, for visits and for individual variables. They used EHR data of 7,491 patients that reportedly included well balanced target groups. They argued for the benefits of using EHRs comprising a variety of diseases and conditions but also noted that the proposed model is not suitable for real-time early detection of high-risk patients in the ICU due to the requirement of longitudinal EHR data for input.

In CVD patients, heart failure is another popularly studied prediction target alongside mortality. Choi et al. developed the RETAIN model and demonstrated it for heart failure prediction [111]. Their recurrent neural network (RNN) model used attention mechanisms and was trained on data extracted from an EHR database of over 263,000 patients. They reported an 87% AUC when testing on a held-out subset containing 15 % of the data. On the same year, Choi et al. also predicted the risk of heart failure from EHR data by stacking medical concept vectors, a specifically trained representation for a patient's records [126]. They created a 100-dimensional numerical vector space, where similar concepts would reside close to each other. The concepts were captured through diagnoses, medications, and procedures. They

trained the concept representation on an EHR database of over 265,000 patients (or smaller subsets) and, from the same EHR, selected a subset of 32,787 patients (3,884 with heart failures) to train several ML based prediction models. They reported AUCs up to 0.81, obtained from six-fold cross validation. Later, Chu et al. applied deep adversarial learning on an RNN to predict the occurrence of heart failure readmission, all-cause mortality, or a composite endpoint within a three-month window and achieved AUC values of 0.74, 0.66, and 0.87, respectively [127]. They used 197 features selected by clinical experts and processed all data into categorical form. They included 2,102 patients from an EHR of heart failure patients. They reported better AUCs for shorter prediction windows as opposed to longer (ranging from 3 to 24 months) and longer sequences as opposed to shorter ones.

EHRs can also be utilized for validation. Kolek et al. used an EHR with over 33,000 patients to validate a previously developed atrial fibrillation risk model [128]. In the inspected 5 years period, 7.3% of the patients developed atrial fibrillation. However, the authors reported that the risk prediction model under-predicted the disease in low risk patients and over-predicted it in high risk patients. Similarly, Rodriguez et al. used EHR data from northern California, USA to validate the Pooled Cohort Equations (PCE, intended for atherosclerotic CVD risk assessment) in different ethnic groups [129]. They observed overestimated risks by 20–60% in the investigated groups.

EHRs provide large data sets, which is a basic requirement for the very powerful deep learning techniques, and prior works have reported successful experiments applying DL methods on EHR data. Especially attention based DL methods have been involved in the state-of-the-art studies, even though EHR data are also a challenging starting point due to their unstructured, multi-modal, error-prone, and episodic contents. However, attention based deep learning has remained underutilized in CVD patient mortality risk prediction, and their capabilities to combine the different types of health records has not been thoroughly addressed.

3.6.4 Algorithm requirements in clinical applications

The acceptance of a novel technology for clinical use can be achieved through extensive validation of the method, investment in user experience among patients, a user-friendly interface for the clinicians, designing it to fit existing workflows, and most importantly, a clinically relevant and actionable output. Considering algorith-

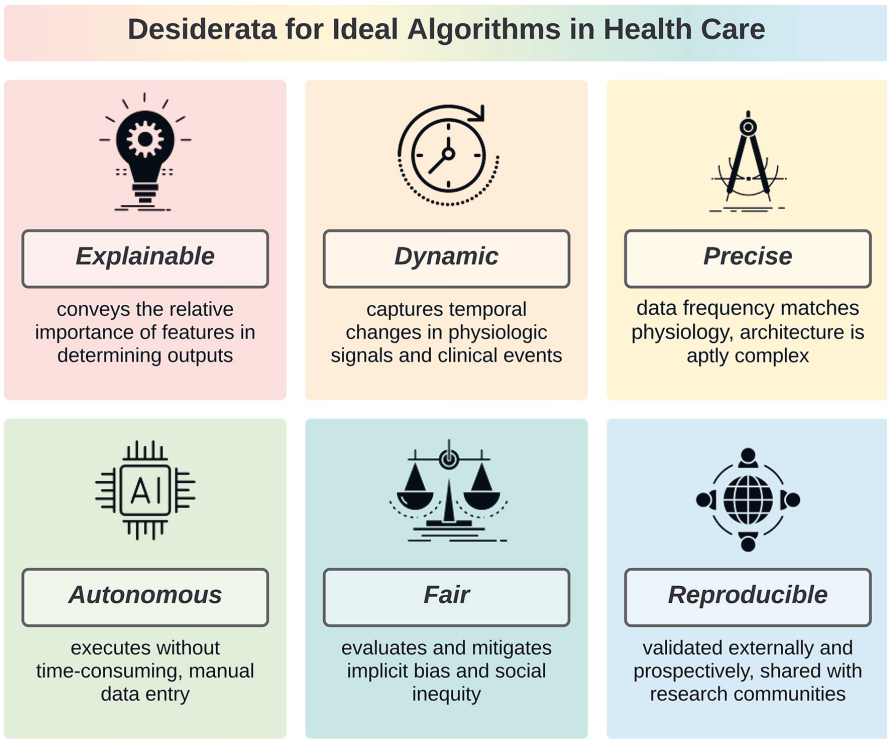

Figure 3.3 The six properties of an ideal algorithm for a healthcare application [134]. Figure from [134], used under the CC BY 4.0 license.

mic requirements for clinical algorithms from the start of the development can help avert many pitfalls that may block algorithm adoption in clinical practice. There are several guidelines to support coherent and sufficient reporting in scientific publications presenting new algorithms with clinical use cases, and guidelines written directly for clinicians to help them evaluate ML applications [130]–[133]. Loftus et al. formulated a framework for ideal healthcare algorithms, which lists six desirable characteristics to maximize the algorithm's benefits to patients, clinicians, and researchers [134]. As depicted in Figure 3.3, they define an ideal algorithm as explainable, dynamic, precise, fair, reproducible, and autonomous [134].

Mathematically complex multilayered solutions, such as deep learning algorithms, often become black boxes where the role of each input in computing the output becomes incomprehensible. A black box algorithm could be based, e.g., on clinicians' actions rather than physiological events, potentially leading to devastating outcomes

in clinical use [134]. Whereas more traditional ML algorithms can be naturally explainable, recent research in the field of ML and AI has put significant effort to shed light inside the black boxes, too [135]. Research has mainly focused on model interpretability and produced methods such as LIME to explain individual prediction outputs and SHAP that can additionally help evaluate feature relationships in the model [136], [137]. However, they have received criticism as the created interpretations can be misdirected by adversarial methods, which naturally reduces their credibility [138]. On the other hand, novel DL mechanisms such as attention have been proposed, building weights into the model itself and offering a means for interpretability [114]. It has been suggested that attention weights may not purely represent importance across the input, but some methods may mitigate this phenomenon [139].

A dynamic algorithm is able to consume new data in real-time in continuously evolving conditions and capture any relevant temporal changes [134]. Dynamic algorithms can ease repetitive and continuous use cases [134]. In chronic disease related applications, algorithms covering different time spans can capture changes of different scales and be utilized in very different applications. The use case as well as the intended implementation setting determine the data that are available and the ways to exploit them to produce the desired output [134]. Different use cases may also set different requirements for algorithm validation. In practice, real-time implementation of algorithms can be a computationally demanding task where issues such as latency, data shifts and keeping algorithms up to date (e.g. via continual learning) need to be considered [86]. Testing and updating the algorithm should be possible also after clinical adoption [133].

The performance of the algorithm and therefore precision is often the focal point in ML research especially when the research aims to produce novel algorithms. For clinical use cases, the performance metrics should convey realistic information about their generalizability in the real use case. The algorithm performance should be evaluated against a reliable reference, preferably the clinical gold standard [133]. The precision of the algorithm is also interlinked with the data collection method. For instance, any signals need to be sampled at frequencies capable of capturing the phenomenon of interest (Nyquist theorem) and too large an input feature set can cause the algorithm to suffer from the curse of dimensionality [134]. The algorithm should be built on sufficient quantities of data to enable generalization to new observations.

Algorithm fairness is closely related to performance evaluation; a fair algorithm benefits different patients equally regardless of any socioeconomic or ethnic factors [134]. The selection of representative samples is central to creating fair algorithms, in both training and validation. For instance, multicenter data capturing sufficient numbers of different demographics may serve this goal [134]. Performance gaps should be analysed to identify any lack of fairness.

Reproducibility of the validation results is crucial for credibility and, thus, acceptability. Recent studies have demonstrated difficulties in this aspect with ML algorithms [140], [141]. Some guidelines have been suggested for standardized reporting of the methods and results [141].

An autonomous algorithm can be a major help to the clinicians, automating and aiding tasks and, importantly, saving time [134]. Making the algorithm easy to use without requiring time consuming additional steps outside the standard workflows can promote uptake. Simultaneously, the clinician should be able to supervise and control the algorithm in order to make an informed decision regarding the output.

Beyond the list by Loftus et al., there are additional requirements related to the practical adoption of clinical algorithms. For example, all algorithm inputs must be available at the intended time of algorithm usage, according to the existing workflows [133]. Additionally, a very specific use case or other built-in restrictions within the algorithm may block its clinical uptake because it quickly becomes impractical to adopt and maintain countless very specific algorithms [133]. The range of the algorithm should be well justified. Algorithms also need interoperability to operate despite changes in the source of data, such as a different device or data storage system, to avoid added costs due to requiring specific data sources. Finally, algorithms will need to conform to regulations, such as the Medical Devices Regulation in the European Union, and follow ethical guidelines, such as the ethics guidelines for trustworthy AI by the European Commission [142]–[145].

4 STUDY DATA

This thesis analyzes three different data sets, with studies II and III using different subsets of one data set. Table 4.1 summarizes the data sets used in each substudy (after applying any exclusion criteria). The data set characteristics are further detailed in sections 4.1 to 4.3.

Table 4.1 Summary of the study data sets

Study	N	Female	Male	Cohorts	Study period	No. of sites	Age range (average)
I	10	2	9	Healthy	32 min	1	25–55 (37)
II*	136	86	50	Healthy, PD, HD, IBD, PSS, RA, SLE	1–21 days	4	21–82 (52)
III*	82	53	29	Healthy, PD, HD, IBD, PSS, RA, SLE	3–12 days	4	21–82 (51)
IV	57,377	7940†	12,548†	CVD	Decades	1	18–105 (65)

4.1 Simulated hypopnoea detection data

Unlike the remaining studies, Study I used a data set collected in laboratory settings. Informed consent was obtained from all participants. Each participant was appointed a time at the laboratory to participate in the study. The participants were measured with the VTT 24 GHz FMCW radar mounted above a bed (at about 2 m distance) using sampling frequencies of 110 Hz and 154 Hz. A CE certified class II medical device (the Embla titanium portable polysomnography system) was used as a reference

*Studies II and III used subsets of the same database

†Sex unknown for a large part of the database

device: the participants wore two ECG electrodes (under the right side collarbone and on the lower left thoracic cage) measuring at 256 Hz sampling frequency, and a RIP belt (on the thorax) operating at 32 Hz sampling frequency [146]. Additionally, a ballistocardiography based sensor sheet by VTT was installed under the bed mattress topper to collect supporting reference data at 110 Hz sampling frequency.

The measurement protocol consisted of three 2 min activities: relaxed respiration, hypopnoea simulation (comprising a 1 min shallow respiration period followed by 1 min of normal respiration), and post-exercise respiration. The first two activities were repeated in supine, right lateral recumbent, prone, and left lateral recumbent positions while the participant was lying on a bed. The hypopnoea simulation was preceded by an additional 2 min supine relaxed respiration period. The post-exercise respiration activity was measured only in the supine position, after two minutes on a treadmill with roughly 15% inclination. The participants had the option to interrupt the treadmill exercise at any time but none did. The same protocol for relaxed respiration in different positions and post-exercise respiration in the supine position was additionally repeated with another sampling frequency. The order of the full protocol with 110 Hz sampling frequency and the reduced protocol with 154 Hz sampling frequency were alternated between participants.

4.2 Fatigue and sleep disturbance study data

Studies II and III exploited different subsets of the same database. The data was collected in the IDEA-FAST project, which aims to identify digital endpoints for fatigue, sleep disturbances, and activities in daily living in NDDs and IMIDs [147]. The study was registered with the German Clinical Trial Registry (DRKS00021693) and received ethical approvals from the ethical committees of the Medical Faculty of Kiel University, Newcastle upon Tyne Hospitals National Health Service (NHS) Foundation Trust/Newcastle University, Erasmus University Medical Centre in Rotterdam, and Georg-Huntington-Institute in Münster, over June to November 2020. The participants were recruited by the aforementioned study sites and enrolled for up to 60 days. Informed consent was obtained from all participants. The participants were either healthy or diagnosed with one of PD, HD, IBD, PSS, RA, or SLE. Inclusion criteria included age of over 18 years, smartphone usage in the past three months, ability to follow written and oral instructions, to walk, sit, and

stand independently, to socialize and communicate, and have over 15 points in the Montreal Cognitive Assessment (MoCA) assessment of cognitive abilities. Exclusion criteria consisted of certain comorbities (such as chronic fatigue syndrome or major sleep disorders), physical traumas with hospitalization in the past 3 months, cancer diagnosis in the past three years, major psychiatric disorders, suicidal attempt in the past 5 years or suicidal ideation in the past half a year, substance/ethanol abuse, and severe visual impairment.

The participants were instructed to wear the VitalPatch biosensor for five consecutive days while conducting their usual daily activities. The participants were also asked to report selected patient reported outcomes (PROs) via an electronic questionnaire four times a day on a smart phone using the VTT Stress Monitor Application (SMA) [148]. The SMA prompted a questionnaire at 9:00, 13:00, 17:00 and 21:00 local time each day, allowing a response within 3 hours (2.5 h for the final daily questionnaire). If the questionnaire had been opened but not submitted, and it was no longer in active view, the app would prompt a reminder every 15 minutes. The questionnaire included 7-point Likert items for physical and mental fatigue, anxiousness, depression, pain, sleep quality ("How was your sleep?"), and physical and mental activities of the day, a 10-point Karolinska Sleepiness Scale (KSS) for sleepiness, and additional questions touching sleep times. The exact set of questions varied with the time of the day. The applications and devices were explained in detail to the participants. The study team provided the participants with informational material, telephone support, and optional home visits to support device usage.

VitalPatch was used for the following continuous physiological measurements: HR, R-to-R interval, respiratory rate, and skin temperature. The data were sorted in time and duplicate records were removed. Invalid values, physiologically unrealistic values, and contextual outliers were removed and treated as gaps, except the latter two for R-to-R intervals, where the removed values were replaced using linear interpolation as routinely done to improve HRV analysis. Invalid and unrealistic values were also removed from posture and the number of steps, which were additionally acquired from VitalPatch to select data for heart rate recovery analysis.

Study II used a subset of 136 patients, 101 of which had reported PROs concurrently while using VitalPatch. Each participant used VitalPatch for 1–21 days, resulting in data of 1,297 days in total. A subset of 91 participants (30 healthy, and 9 PD, 6 HD, 12 IBD, 13 PSS, 10 RA and 11 SLE patients) had at least two 2 h

windows of sensor data (with >70% coverage of inspected signals, as compared to the number of expected samples) ending at a PRO response. This subset was employed to analyze the association between features derived from the sensor data and the PROs. Another subset of 73 participants (21 healthy, and 10 PD, 9 HD, 11 IBD, 8 PSS, 8 RA, and 6 SLE patients) had at least one uninstructed 6 min walking period (an average of at least 60 steps/min) followed by a one minute rest with full coverage heart rate. This subset was used for heart rate recovery analysis.

Study III used a smaller subset of the same data, including 82 individuals who had responded to a KSS questionnaire at least six times over 3–12 days while wearing the VitalPatch biosensor. The target for study III was to predict daytime sleepiness (sleepy or non-sleepy) from two hours of continuous respiratory rate signal. The subset included 8 PD, 6 HD, 10 IBD, 13 PSS, 7 RA, and 12 SLE patients and 26 healthy volunteers. The KSS comprised 10 options ranging from *extremely alert* (index 0) to *extremely sleepy* (index 9). The KSS responses were used as the prediction target so that *extremely alert* to *rather alert* (indices 0–3) were labelled non-sleepy and *neither alert nor sleepy* to *extremely sleepy* (indices 4–9) were labelled sleepy.

The respiratory rate data covering at least 90 % of the 2 hours preceding a KSS response was taken as the input data. Given the sampling rate of 0.25 Hz, the full 2 hour window would have consisted of 1800 samples. Therefore, the signal coverage was evaluated as the observed number of samples divided by 1800. Reduced coverage may have resulted from non-wear time or data cleaning.

4.3 Mortality risk prediction data

Study IV exploited a large database collected from three sources: the Pirkanmaa Hospital District EHR, the Tays Heart Hospital KARDIO registry, and the Finnish mortality registry collected by Statistics Finland. The first two extend back to the 1990's and early 2000's, respectively, and the date of death from the thirdly mentioned was included for the matching period. The database was collected automatically as a part of a retrospective registry study, MADDEC, and continued until January 2020 [78]. For Study IV, the data were anonymized in accordance with the Finnish legislation, leading to partial data loss, e.g., exact timestamps were replaced with days since birth. As a retrospective study, informed consent was not required.

The database consisted of CVD patients (various cardiac conditions) treated at the

Tays Heart Hospital. The anonymous database consisted of patient events, which were linked to the patient and their background information, such as birth year and residence, as well as event details according to the event type. The events were categorized as labs, diagnoses, medications, operations, procedures, measurements, hospital visits and wards, angiography, imaging, percutaneous coronary intervention (PCI), coronary care unit (CCU), transcatheter aortic valve implantation (TAVI), or resuscitation. Each event type was recorded with attributes specific to that event type. The anonymous database comprised a total of 72,680 patients, 9172 of which had diseased within six months of their last visit.

The event-oriented data were pre-processed into patient specific time series of events. Individual events were excluded from the time series if the event time was missing or overlapped with the date of death, or occurred before the patient turned eighteen. For each event type, the size of the event representation was limited to control time series length: attributes were completely disregarded if they were completely missing or were not within the nine best available attributes. Otherwise all event type specific attributes were included in all occurrences of such events, with any missing values set to none. The patient time series were tokenized with pre-trained tokenizers and special tokens were inserted once to each time series to the expected position (according to the selected transformer model). Numerical attributes were transformed to string type integers prior to tokenization, including the age of the patient at each event, which was additionally transformed to years. Finally, the sequences were truncated to 512 most recent tokens. Under-length sequences were padded with a specialized padding token determined by the tokenizer.

Study IV was formulated as a classification task between those who survived over six months after an event (negative cases) and those to those who died within 182 days of the last event (positive cases). Importantly, data were retrospective (including full history of those diseased) whereas the clinical use case would entail predictions at random patient encounters. To avoid bias due to the retrospective nature of the data, a random number of events were removed from the end of each patient time series, as long as (a) at least five events remained and (b) death was still within six months for the positive cases.

5 RESULTS

5.1 Accuracy of contact-free detection of simulated hypopnoea events

In Study I, the FMCW radar detected respiratory rate with an overall MAE of 1.4 rpm as compared to a certified medical device in laboratory settings. The radar derived respiratory rate demonstrated a 91 % Pearson correlation (p<0.01) with the reference, while the total RMSE was 3.1 rpm. The Bland-Altman plot is shown in Figure 5.1. The user specific MAE varied from 0.7 rpm to 5.1 rpm. As seen in Table 5.1, the highest errors clearly occurred for shallow respiratory motions, during the simulated hypopnoea. Moreover, the errors for this activity were notably higher in the lateral positions than in the supine or prone positions. It should, however, be

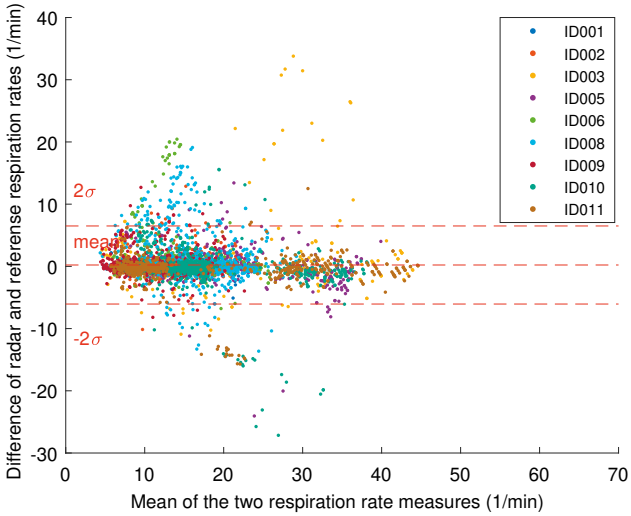

Figure 5.1 Bland-Altman plot of respiratory rate between the FMCW radar and the reference device.

Table 5.1 Mean absolute error in respiratory rate (rpm) from FMCW radar with respect to activity and lying position.

Position	Relaxed	Hypopnoea, Shallow	Hypopnoea, Normal	Post-exercise	Position MAE
Supine	1.002	1.270	1.149	1.149	1.086
Right lateral	1.096	3.588	0.682	-	1.656
Prone	1.256	1.458	0.707	-	1.225
Left lateral	1.732	3.454	1.001	-	2.064
Activity MAE	1.222	2.408	0.887	1.149	1.414*

noted that the reference RIP belt may have slightly dislocated and loosened when a participant rotated their body on the bed to change positions.

Table 5.2 Mean absolute error in interbeat interval (s) from FMCW radar with respect to activity and lying position.

Position	Relaxed	Hypopnoea, Shallow	Hypopnoea, Normal	Post-exercise	Position MAE
Supine	0.034	0.014	0.042	0.052	0.038
Right lateral	0.040	0.036	0.041	-	0.040
Prone	0.039	0.028	0.054	-	0.040
Left lateral	0.038	0.028	0.032	-	0.035
Activity MAE	0.037	0.026	0.042	0.052	0.038*

Monitoring the heart can reveal indications of developing CVDs due to sleep apnoea, thus offering valuable information when coupled with respiratory monitoring. The IBI measurements with the FMCW radar were typically overestimated with an overall MAE of 38 ms and a total RMSE of 84 s from the medical level reference, with an 81 % Pearson correlation to the reference values. Detailed inspection of the results revealed highest IBI measurement errors for arrhythmic heartbeats. The results are summarized with respect to the varying respiratory activities and positions in Table 5.2. The best accuracy was during shallow respiratory motions. The results demonstrate that the smaller the interference from the respiratory motions, the more accurate the IBI measurement can be. In terms of heart rate, the overall MAE was 1.1 bpm when ectopic beats were discarded.

Finally, the error in HRV analysis was evaluated for a selected group of time and frequency domain HRV features. The radar derived time domain features demon-

*Total MAE over all activites and positions

strated 79–98 % correlations with the reference, and frequency domain features showed 72–95 % correlations. For instance, MAE was 10 ms for both RMSSD and NN SD, and 20 ms for normal-to-normal intervals.

5.2 Continuous monitoring of measures describing the quality of life

In Study II, two-hour aggregates of physiological signals (HR, HRV, respiratory rate, temperature) were L5 normalized for each individual. The L5 windows were located at 1-min resolution. The L5 normalized data exhibited statistically significant correlations with immediately following patient reported outcomes describing fatigue and sleep. Several significant correlations were identified for all three cohorts (healthy, NDD, IMID) whether the measurements were normalized by the mean and SD averaged over all participant-specific L5 windows (see Figure 5.2), or just the nearest preceding L5 window (see Figure 5.3). Interestingly, only the correlation between minimum HR normalized by overall L5 metrics and sleepiness was significant in all three cohorts. In both disease cohorts, the respiratory rate minimum normalized by full L5 data correlated negatively with the reported sleep quality, while the healthy cohort did not show a significant correlation.

With normalization using all L5 data, the healthy and IMID patients showed several similar correlations. For both groups, mental activity level correlated with skin temperature maximum, and mental fatigue correlated with HR SD and R-to-R maximum, all negative correlations. Sleepiness showed negative correlations with HR minimum, maximum, mean, and SD (also with HRV derived HR maximum, mean, and SD instead of the directly available HR), respiratory rate maximum and SD, and NN CV. Additionally, sleepiness correlated positively with R-to-R and NN mean, and NN median.

Importantly, the healthy and NDD both showed significant correlations between sleep quality and NN range and HRV derived HR minimum, but of opposite directions. Wider NN range correlated with poor sleep quality in the healthy and good quality in the NDD patients. Higher HRV derived HR minimum correlated with better sleep quality in the healthy, and poor quality in the NDD patients.

In the NDD cohort, the significant correlations were mainly related to sleep quality, with only two significant correlations with both mental fatigue and mental activity level, and three significant correlations with sleepiness. The IMID cohort,

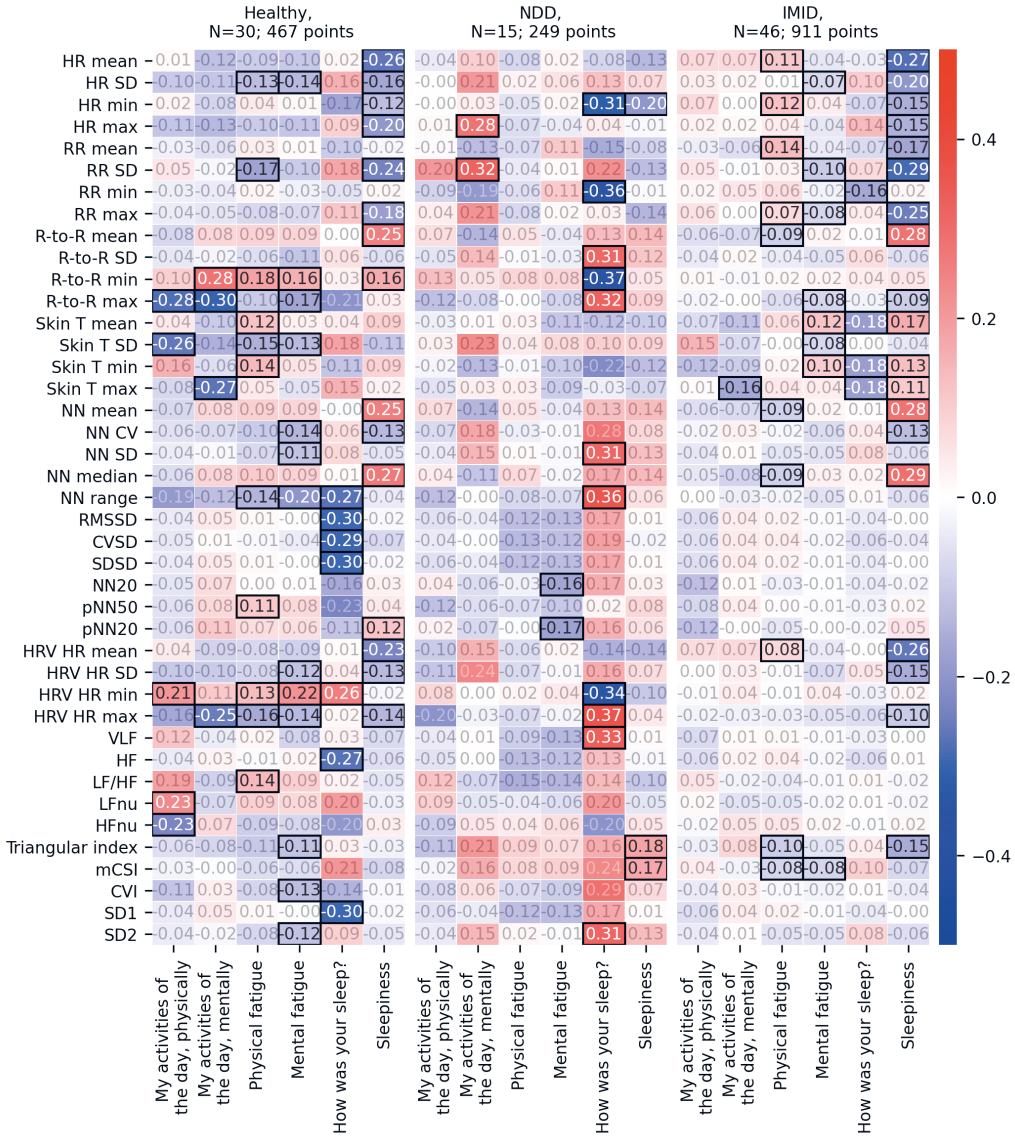

Figure 5.2 Repeated measures correlation r values between physiological quantities and PROs, for the healthy, NDD patients, and IMID patients. The statistically significant (p<0.05) are depicted on opaque colours. Here, the participant-specific normalization method used all available L5 data.

on the other hand, had only a few significant correlations with sleep quality, while most significant correlations were for sleepiness. The IMID group also demonstrated several significant correlations with both mental and physical fatigue. The healthy

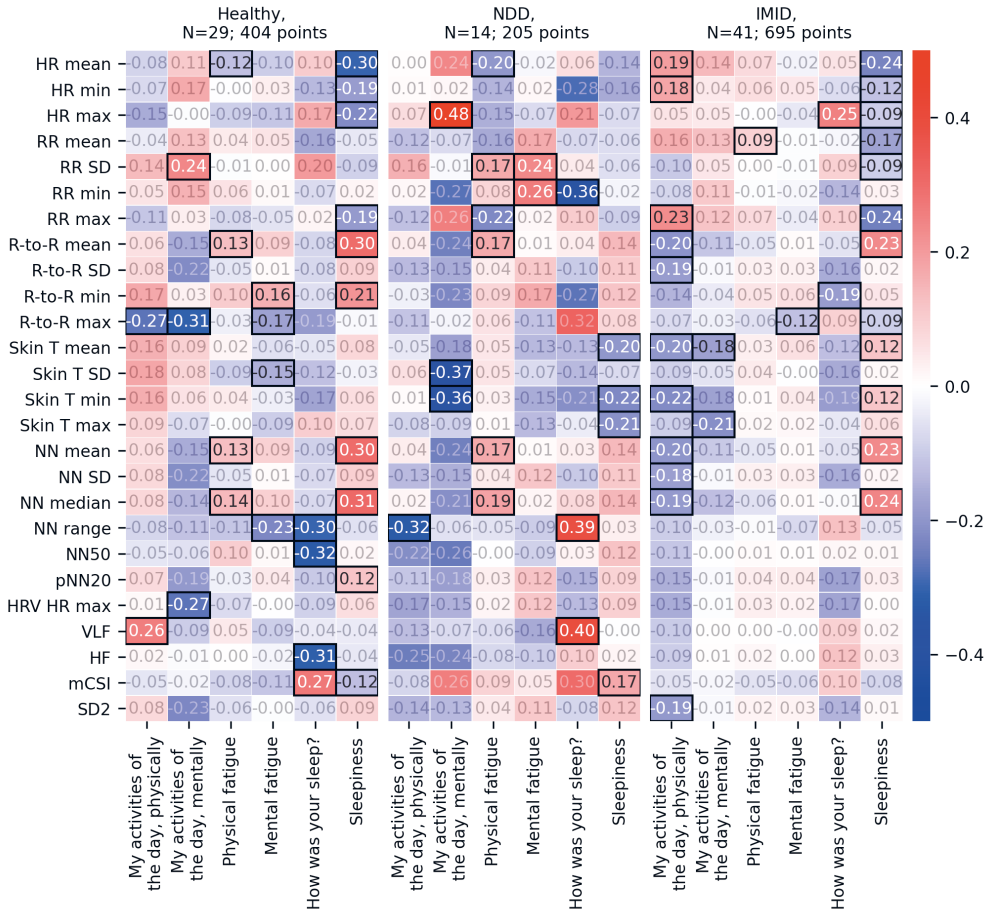

Figure 5.3 Repeated measures correlation r values between physiological quantities and PROs, for the healthy, NDD patients, and IMID patients. The statistically significant (p<0.05) are depicted on opaque colours. Here, the normalization method used the only the most recent available L5 data.

participants had several significant correlations with all PROs.

With normalization using only the latest L5 data, the healthy group exhibited mainly the same correlations as with the features normalized by all L5 data. Nevertheless, six new correlations emerged: (1) mCSI correlated with sleep quality and (2) sleepiness; (3) NN50 correlated with sleep quality; (4) NN median correlated with physical fatigue; (5) respiratory rate SD correlated with mental activity level; and (6) VLF correlated with physical activity level.

For NDD patients, only sleep quality correlations and the correlations between

HR max and mental activity level, and mCSI and sleepiness were similar to features normalized by all L5 data. Physical fatigue and physical activity level of the day showed significant correlations only with the latest L5 based normalization.

With latest L5 based normalization for IMID patients, the significant correlations found for mental activity level, mental and physical fatigue, and sleepiness also appeared with the other normalization approach, whereas two dissimilar correlations were identified for sleep quality. Interestingly, eleven significant correlations for physical activity level emerged, whereas no such correlations were observed with the full L5 based normalization.

5.2.1 Contextual features

Heart rate recovery over a restful period of 1 minute, following at least 6 minutes of walking, showed a statistically significant difference between cohort groups (F 5.68, p<0.006), as depicted in Figure 5.4 together with the walk durations. Partial η^2 indicated that cohort groups accounted for 14 % of HRR variance. Additionally, HRR in the healthy participants differed significantly from both the NDD (T 3.95, p=0.001) and IMID (T 2.51, p<0.038) patient groups.

Furthermore, HRR showed significant differences between low (scores 0–2) and high (scores 3–6) fatigue groups for both mental and physical fatigue, but only within the healthy participants. The two-sided KS test scored 0.77 (p 0.01) for physical

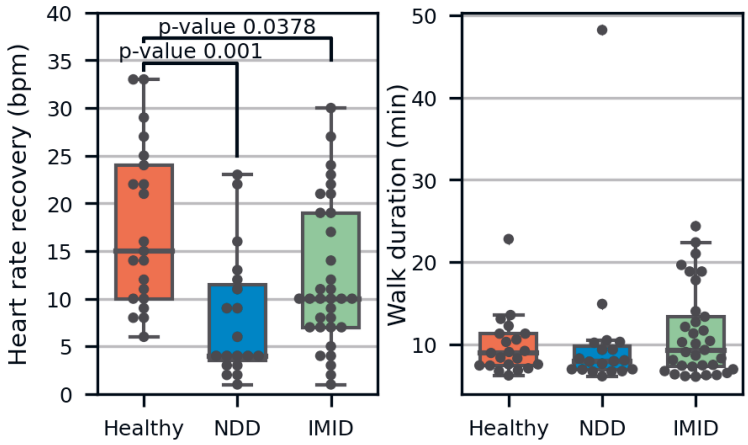

Figure 5.4 Heart rate recovery after at least one minute of walking (left) and the corresponding walk durations (right) in the healthy participants NDD patients, and IMID patients.

fatigue and 0.68 (p 0.02) for mental fatigue.

5.2.2 Statistical and deep learning based digital biomarkers

In Study III, the 1D CNN could classify 2-hour respiratory rate data between sleepy and non-sleepy with 62.6 % accuracy, 57.2 % sensitivity, and 69.2 % specificity in a test set comprising data from previously unseen participants (10 IMID and 2 NDD patients, and 5 healthy with roughly 52 % of samples representing the sleepy class). The model comprised 10,426 trainable parameters and was trained over 25 epochs. The input was standardized by the training set's average respiratory rate and standard deviation. Early stopping was applied with the condition of observing over 0.05 difference in training and validation set loss for three consecutive epochs.

5.3 Mortality risk prediction from electronic health records

The optimized BERT comprised 108,312,578 trainable parameters with 12 hidden layers, each with 12 attention heads. The optimized XLNet only used 6 hidden layers with 6 attention heads, totalling 5,482,130 trainable parameters. The final training was repeated five times to account for the effect of random initialization, and early stopping was applied with the condition that training loss would fail to improve by more than 0.0045 over 5 epochs. On each repetition, the training was stopped before reaching 50 epochs.

The final BERT model achieved the classification AUC of 75.5 %, precision of 19.2 %, and sensitivity of 73.3 %. The corresponding results for XLNet were 76.0 % (AUC), 15.9 % (precision), and 83.1 % (sensitivity). Overall, XLNet achieved a slightly better AUC and clearly better sensitivity, thus capturing more positive cases than BERT. Yet, the low precision scores reveal that both models produced mostly false positive predictions. The test results were well aligned with the cross-validation results during development apart from sensitivity. Sensitivity dropped in the test set due to the smaller relative amount of positive cases. The portion of positive cases in the test set matched the portion in the full data, whereas the training and validation data contained a larger portion due to down-sampling, which was applied to handle class imbalance during training.

The attention weights were inspected layer-by-layer for individual examples by using BertViz [95]. In an example case representing the positive class, correctly clas-

Figure 5.5 The attention weights in (a) XLNet and (b) BERT in selected layers (5th and 12th respectively) near the CLS token, for a given example of a positive case, correctly classified by XLNet and misclassified by BERT. The example is shown in text form in (c) with the first information available for XLNet and BERT highlighted in green and purple, respectively.

sified by XLNet and misclassified by BERT, XLNet attended to the patient's age and operation code towards the end of the timeline and BERT attended to lab results, as visualized in Figure 5.5. Another positive example case, depicted in Figure 5.6, was misclassified by XLNet and correctly classified by BERT. In this case, XLNet is mostly seen to attend to measurement context (e.g. suspected diagnosis) and measurement name, and BERT to the numerical values of lab tests in the vicinity of the CLS token. In both cases, BERT is able to include just slightly more information for the classification than XLNet, due to differences in tokenization.

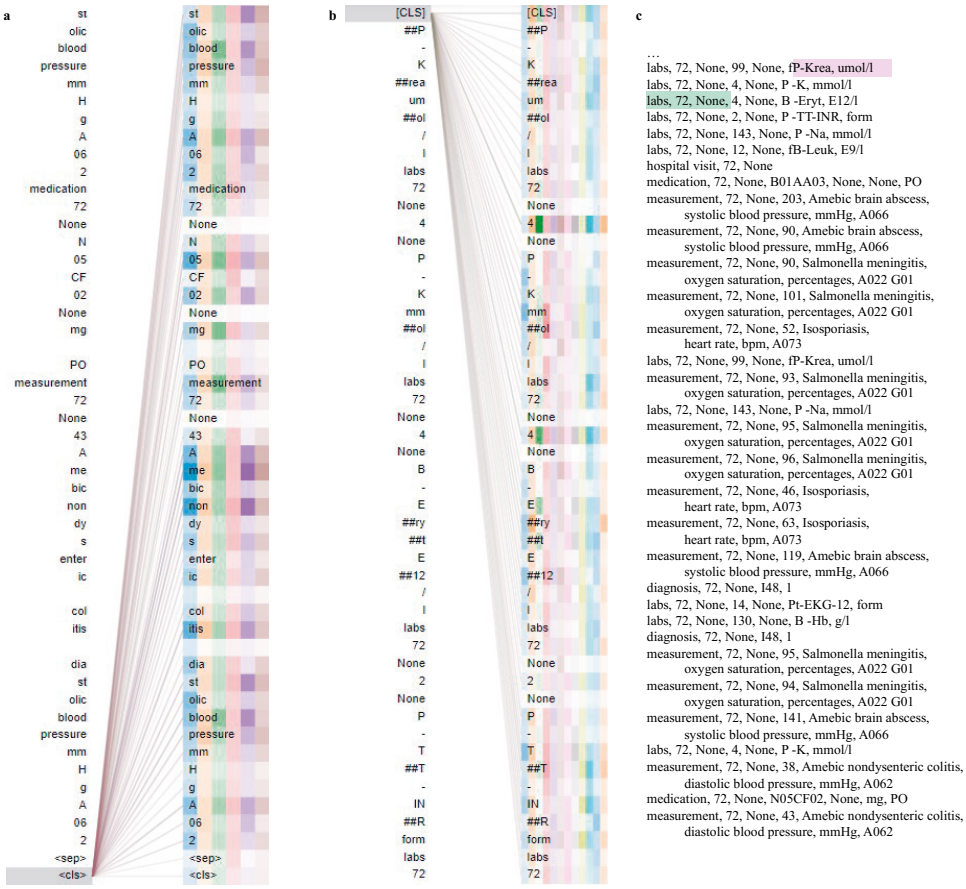

Figure 5.6 The attention weights in (a) XLNet and (b) BERT in selected layers (5th and 12th respectively) near the CLS token, for a given example of a positive case, misclassified by XLNet and correctly classified by BERT. The example is shown in text form in (c) with the first information available for XLNet and BERT highlighted in green and purple, respectively.

5.4 Data quality and availability across time spans

Study I was conducted in laboratory settings and the high coverage for the FMCW radar data may not be representative of a real-life setting. It is noted that the radar was sensitive to motion artefacts.

In Study II, the data were of good quality with less than 0.5 % outliers for HR, respiratory rate, and R-to-R intervals each, while 2.3 % of skin temperature were outliers. The median daily data coverage from the patch sensor was 78 % for HR, respiratory rate, and R-to-R intervals, and 77 % for skin temperature, after cleaning

the data. The median coverage was comparable across cohorts. Eleven participants stood out with missing skin temperature measurements and were revealed to have been collected at the same study site. Therefore, the missing measurements could be related to practicalities at one site.

The EHR data in Study IV was fundamentally different as compared to the sensor data in Studies I–III and hence the data quality cannot be evaluated similarly. The EHR data were only recorded when a patient used healthcare services, and they were partially incomplete and sometimes erroneous. The notation was not uniform across the database for several variables, body-mass index (BMI) included unit errors, and timestamps could be missing or recorded in reverse order. The data gaps were partly due to database anonymization. Sex was only available for roughly 36 % of patients. In the patient event time series, some event attributes were completely missing and excluded from the event attributes. In the 14 event types, residence was completely missing for TAVI events and in 59–94 % of the other events excluding CCU, resuscitation and hospital ward events where it was available in 67–79 % of cases. Diagnosis code and priority were only available in 36 % and 41 % of diagnosis events, respectively. While event type, start time, all operation attributes, times repeated, ward, sex, stenosis, imaging type, dialysis, temporary pacemaker, primary vasoactive medication, fluoroscopy time, and glomerular filtration rate attributes were all fully available for the related event types, the availability of other attributes varied starting from 56 % (with the exception of 1 % availability for textual lab attribute values).

6 DISCUSSION

The presented studies proposed decision support tools for chronic disease management. Temporal patterns and changes in health data may indicate patient deterioration and disease progression. Monitoring the evolution of health data may be used to benefit disease management and patient outcomes. The selected application time spans reflected the time needed to capture meaningful health events or physiological changes and, on the other hand, time required to make the output actionable.

Study I aimed for robust vital sign extraction across individuals. The Bland-Altman plot was selected to identify systematic and random error, while MAE, RMSE, and the Pearson correlation coefficient were used to quantify the agreement with the reference device. Studies II and III assumed chronic diseases with similar origins would exhibit reasonably similar physiological phenomena across patients whereas different disease groups might not. It was also assumed that the chronic diseases may affect the patients' daily lifestyle similarly. In Study II, normalization using participant specific L5 parameters was selected to promote inter-individual comparability and repeated measures correlation was selected to assess common within-individual association in a group of individuals while accounting for the repeated nature of the measurements. CNN was selected for Study III due to its high merits in spatial pattern recognition. Study IV aimed for temporal pattern recognition in episodic patient event data, among a heterogeneous group of CVD patients. Transformer models were chosen based on their abilities to learn bi-directional patterns in sequential data and use attention mechanisms to potentially resolve issues from the episodic and sparse nature of the data.

This chapter discusses the results while considering the research questions presented in Section 1.1. Section 6.1 focuses on research question 1 and reviews the algorithm performance results and technical challenges from each study. Section 6.2 inspects method applicability across time spans while assessing the presented solutions with respect to the requirements for ideal clinical algorithm determined, thus

addressing research questions 1 and 2. Section 6.3 delves into research questions 3 and 4 by discussing the feasibility of data collection methods with chronic disease patients and data collection requirements across time spans. Finally, Section 6.4 details the limitations of each presented study, touching their limitations related to clinical use and uptake to further address research question 2.

6.1 Algorithm performance

The presented work applied model based and data driven time series analytics to applications relevant to chronic diseases, and evaluated algorithm performance on real human subjects while reporting promising results. This subsection sums up the algorithm performance and scientific contributions of each substudy.

6.1.1 Contact-free monitoring applicable for monitoring abnormal respirations

Accurate contactless monitoring can offer a user-friendly and unobtrusive means of patient monitoring during sleep, at hospitals and in home care. Study I demonstrated high accuracy in a wide range of both respiratory and heart rates. FMCW vital sign monitoring obtained small MAE for both respiratory and IBI monitoring, and demonstrated potential for contactless HRV analysis.

The correlation between radar and reference respiratory rate was high at 91%. Across different scenarios, the respiratory rate error was highest for simulated hypopnoea where the reference respiratory rate was at highest. As seen in Figure 5.1, two participants stood out with largely underestimated (by over 20 rpm) respiratory rates from the radar for individual samples, when the reference value was high. Similarly, one participant stood out with some overestimated respiratory rates from the radar when the reference values exceeded 15 rpm. A larger study in real-life settings is needed to determine whether hypopnoea events of clinically significant duration could be missed due to the high error cases.

Using the autocorrelation function for respiratory rate extraction causes a delay equal to the maximum peak extraction buffer (15 seconds by default). In real-life applications, dynamic optimal range bin selection and DC removal would be needed, creating further latency. Improvements to decrease latency would be needed if this technology were used for any real-time interventional systems, with the intention to restart respiration during cessations.

Heart rate and HRV during hypopnoeas or apnoeas could act as early indicators of apnoea-induced pathological changes towards CVDs. Contact-free HR and HRV monitoring accuracy was highest at low respiratory motions, which insinuates technological applicability for such a use case. HR and HRV could also be used for sleep analysis throughout the time spent in bed. It should be noted that for monitoring IBI derived measures, the monitoring latency depends on the longest FFT window and the window used for iterative smoothing to remove uncertain observations. Here, the windows were 20 seconds and up to 4 minutes long, respectively, which may not be sufficient for interventional systems.

As compared to preceding studies, Study I achieved a high respiratory monitoring accuracy over a wide range of respiratory rates over several study participants while using a certified medical device as a reference. Prior studies were typically restricted to normal respiratory rates, seated positions, fewer participants, and/or commercial reference devices [99], [101], [103], [104]. Only Arsalan et al. included a sequence of holding one's breath in their measurements but only measured for 20 seconds and only reported errors for heart rate [102]. They also reported higher heart rate error during normal respiration as compared to holding one's breath.

More recent studies have investigated the feasibility of sleep apnoea detection with commercial FMCW radars and reported promising results obtained via ML and DL approaches, utilizing PSG reference [149], [150]. The studies included up to 44 participants and performed respiratory signal extraction but did not quantify the accuracy of the extracted signals. Choi et al. used 1 min signal segments and considered the apnoea-hypopnoea event severity in their analysis [149]. They noted that detection sensitivity was higher for severe apnoeas (87 %) but remained limited (54 %) for hypopnoeas.

6.1.2 Continuous monitoring offers objective measures to assess fatigue and sleep

Studies II and III addressed the lack of objective measurements to describe HRQoL in chronic diseases. The studies evaluated the agreement between a plethora of physiological measurements and fatigue and sleep PROs. Importantly, continuous measurements were conducted using multi-modal wearable sensor technology in free-living settings, and included both healthy and chronically ill patients. The data covers daily living and hence offers realistic insights of the feasibility of the proposed methods. The objective measurements could potentially augment the questionnaire based

HRQoL assessment that is currently used in healthcare and provide a consistent means to evaluate therapeutic outcomes.

Study II identified several promising physiological measures that correlated with PROs. All statistically significant correlations were modest, with absolute r values in 0.07 – 0.48. It should be noted that a subjectively reported reference may be influenced by many things, like recall bias and inter-individual differences in, e.g., tolerance to changes. Hence, high correlations with PROs are not typically reported. Additionally, the subjective measures may be confused with one another, whereas an objective measure typically describes a more well-defined (here, physiological) phenomenon. Additionally, here, two hours of physiological signals preceding a PRO were inspected, whereas some other windows relative to the PRO timing may be equally reasonable.

Three different groups (healthy, NDD and IMID patients) were studied, but only one correlation, that between minimum HR (full L5 based normalization) and sleepiness, was significant in all groups. This implies physiological differences between the cohorts. Some correlations appear significant in both the healthy and the IMID patients but not in NDD patients, and some correlations were common for healthy and NDD patients but the opposite direction. This could be indicative of the degeneration of central autonomic nuclei and/or pathways in NDDs.

Study II inspected two alternative approaches to normalize the physiological measures. While both relied on resting data for constructing the user-specific baseline, one used all available L5 windows and the other only the most recent L5 window. The firstly mentioned baseline may be more stable while the other shifts with daily changes. Indeed, the latter baseline revealed significant relationships to PROs the other normalization did not, especially in the disease groups; for physical activity and physical fatigue in NDD patients and for physical activity in IMID patients. Interestingly, some physiological measures were significant despite the normalization method.

Many of the observed correlations comply with normal physiological functions. For example, lower HR measures correlated with higher sleepiness in especially healthy participants and IMID patients, and normally heart rate slows down as the body prepares for sleep [8]. Also in the healthy participants, worse sleep quality was associated with lower HRV measures (NN range, RMSSD, CVSD, SDSD) [61]. Since the sleep quality question was prompted only in the morning, the analyzed

physiological measures were more likely to contain sleep.

Heart rate recovery analysis in Study II demonstrated that physiological measurements in a specific context in free-living settings may be useful in assessing patient status. As a measure of physiological fitness, HRR might be affected by fatigue. HRR showed differences between healthy and chronic disease patients, and may be indicative of mental and physical fatigue in the healthy. Prior work has suggested that laboratory HRR correlates with physical fatigue [151], [152]. A longer study period is needed to confirm the presented findings on a participant level, to enable repeated measures correlation analysis.

In addition to statistical aggregations, temporal patterns may capture more intricate indicators of decreased HRQoL. Study III was motivated by the natural association between yawning and sleepiness. It showed that deep learning methods hold promise in daytime sleepiness classification, although PROs were found a poor ground truth for a data-based model. The daytime sleepiness classification accuracy from respiratory rates was modest at roughly 63 %. Although respiratory rate was studied because it may be easily available from wearable devices, applying CNNs to the underlying respiratory signal may reach more informative patterns.

6.1.3 Bi-directional patterns in EHR data indicate increased risk of death

Predicting 6-month mortality can provide a significant advance for chronic disease management and treatment by flagging high risk patients in good time. Decision support tools that can be easily used during a doctor's appointment to detect increased risk of death may indicate changes in patient status, allowing preventive actions and interventions. Study IV demonstrated the promise of transformer models for mortality prediction with CVD patients, although many technical challenges remain before this could become a reality. Attention based algorithms like transformers are specifically interesting for healthcare applications because they have increased hopes of achieving explainable deep learning. The results from Study IV advocate the use of XLNet over BERT, underlining the need to expand future EHR based research beyond BERT and the need to pre-train XLNet resources for clinical applications.

Study IV showed moderate AUC results (75–76 %) for 6-month mortality prediction among CVD patients by applying transformer neural networks on multimodal patient time-series. Both BERT and XLNet produced mainly false positive predictions. XLNet achieved a reasonable sensitivity of 83 %, exceeding that of

BERT by nearly 10 %. Previous work has suggested the autoregressive XLNet may be better able to learn long-term dependencies as compared to BERT, which may explain why XLNet outperformed BERT [94]. XLNet additionally surpassed BERT in terms of size. The optimized XLNet network comprised 5.5 million trainable parameters; only 5 % of the amount required in the optimal BERT model. Generally, a smaller model is computationally less expensive and provides an easier starting point for productization and real-life use at hospitals.

The examined models used standard English tokenizers and were trained from scratch. Even though the results were promising, the models could benefit from tokenizers specialized in EHR vocabulary. As seen with Chat-GPT by OpenAI, pre-training transformers on extensive data can yield impressive results. Furthermore, the event-oriented data representation with a fixed number of attributes for each event type faced challenges due to computational restrictions. The most recent 512 tokens representing the patient history were used, which captured varying time periods from different patients. The proposed event-oriented representation could help discover novel dependencies in CVD disease progression but more concise representations are needed to capture more complete patient histories.

The patient population in Study IV was heterogeneous among CVD patients and highly imbalanced; only 6.57 % of the patients had died. In the future, harmonized EHR databases across hospitals and nations could enable more homogeneous patient data sets for more specialized and possibly more accurate models by providing a high number of positive examples despite the natural low prevalence of positive cases.

Other EHR based mortality studies building on attention typically predicted 24h mortality at admission and often focused on ICU patients, achieving varying results [119], [120], [125]. Study IV, on the other hand, examined chronically ill patients, who may benefit from early detection of increased mortality risk. Interestingly, non-deep learning models by Hernesniemi et al. achieved AUCs beyond 0.87 but only included a highly homogeneous subgroup of patients in their model [124]. Hence, more specialized models may perform better. Furthermore, more encouraging results have been reported on transformer models with more simple EHR input data representations [114].

Finally, Study IV experimented with model explainability by attempting to interpret the attention weights of individual model outputs. The deepness of the network and the number of attention heads were found to complicate model output inter-

pretability, and existing attention visualization tools like BertViz provide fairly complex output in contrast to quickly and easily comprehendable information needed for clinicians. More intuitive and user-friendly interfaces are needed for clinical use. When such tools are established, integrating them in to the EHR could make them a valuable part of the clinical workflow.

6.2 Analytical methods across time spans

The presented work employed model driven and data driven methods for applications with the time span of interest ranging from overnight to weekly and month-to-month. All solutions demonstrated dynamicity to capture temporal changes relevant to the application time span, although other requirements for the ideal clinical algorithm (explainability, preciseness, autonomy, fairness, and reproducibility), as defined by Loftus et al., were not always fully met [134].

For near-real time, overnight monitoring, the primary goal was high measurement accuracy and thus the ability to detect abnormalities over night. Despite their small latency, the selected model driven methods reached high accuracies for monitoring individual physiological events (respiration and heart beat), demonstrating dynamicity and preciseness. Deep learning methods could potentially match the reported accuracies and be equally autonomic but the presented solution benefits from explainability. Some participant-to-participant variation in the accuracy was observed. Hence, fairness could be further affirmed in future work. The result reproducibility could be estimated through the overall MAE in future studies.

For week-by-week monitoring, the aim was to find objective measures that reflect the health-related quality of life and the related changes. Both model driven and data driven methods showed modest correlation with patient reported outcomes, demonstrating ability to dynamically detect patterns in physiological events. In free-living settings, measuring in a specific context may provide further insights into patient status, as demonstrated in Study II.

The preciseness was challenging to evaluate in both studies II and III due to the limited-quality reference. PROs do not provide the high quality ground truth needed especially for data driven methods. Bearing that in mind, the results in Study III were quite promising and the method may prove more successful as a personalized model. This, however, requires far longer data collection and the implementation would

similarly suffer from a long training time to obtain the personalized model. The statistically based objective measures in Study II, on the other hand, attract with their explainability, although they too require some period of wear time to first collect a sufficient baseline (depending on the normalization method). Both solutions could operate autonomously.

It is noted that the proposed objective measures in Study II work differently between the participant cohorts. The fairness of the algorithm across other factors may depend on the wearable sensor technology; e.g. optical sensors accuracy may be affected by skin tone. Here, the sensor was ECG based and thus expected to be fair across patients.

In month-to-month monitoring, the objective was to predict increased risk of death. Data driven methods demonstrated reasonable ability to detect patterns between different EHR events and could provide a highly autonomous solution if integrated in the EHR system. The presented solution was able to exploit high volumes of heterogeneous data without feature engineering. The explainability of the model was insufficient with too complex visualization tools to explain the attention weights behind a single prediction. A previous study used traditional machine learning, including explainable models, to predict six month mortality in a more limited subset of CVD patients from the same database [124]. Patients with acute coronary syndrome comprised a considerably smaller (N=9066) and more homogeneous set of patients, and the study achieved AUCs above 80 %, notably higher than in Study IV. This observation emphasizes the challenge of learning meaningful dependencies among a wide spectrum of conditions with different symptoms and pathologies in a single model. On the other hand, collecting sufficiently large data sets of homogeneous patients can be very difficult, especially when the classification setting is highly imbalanced.

Additionally, the model in Study IV was trained on an ethnically homogeneous population, and its fairness and reproducibility at other locations with more diverse populations and possibly different EHR systems are not guaranteed.

Overall, model driven approaches (autocorrelation, cepstral analysis, statistical and HRV analysis) were generally more explainable as compared to data driven deep learning methods. Because of their explainability, model driven methods may be more easily accepted, especially in cases where data driven methods cannot be provided with a high quality reference to obtain high accuracies. Model driven ap-

proaches offer dynamic and precise solutions in applications for short-term time spans. On the other hand, data driven methods were able to capture more abstract patterns, and were more easily applied to sparse heterogeneous data. They may therefore be the more advantageous choice for applications related to longer time spans, which may rely on more complex dependencies across time and even across data sources.

6.3 Data collection for different time spans

The detection or prediction target sets the time span of interest and defines the minimum data collection period for algorithm development. The shorter the time span, the higher the required data frequency typically is. For example, in Study I, to detect abnormal physiological events the data collection frequency needs to be sufficient for accurate monitoring, as defined by the Nyquist theorem. Correspondingly, for long time spans, high level data may be sufficient. For instance, in Study IV the time span was six months, the data was episodic and each event represented by a rather small set of attributes. Like in Studies II and III, high frequency sensor data may be processed into less frequent derived metrics. For example, ECG signal may be processed to extract heart rate and further the resting heart rate, which can be a useful metric to monitor over longer periods.

Some of the collected data could not be employed in algorithm development. In studies I-III, some participants were excluded from the analysis due to problems with sensor data collection or reference data collection. In study II, the PRO coverage varied significantly between PROs. Moreover, out of 72.7k patients in Study IV, 57.4k were eligible for the study but only 23.5k patients could be included after downsampling. The actual amount of applicable data may be estimated before hand by considering the expected number of positive cases (if can be estimated) and the requirements of the analytical method together with the availability of computing resources. This can be especially important when the algorithm development depends on pre-existing data records.

The method of data collection plays a key role in fulfilling the ideal clinical algorithm requirements. The data collection technique (sampling frequency, accuracy, and overall quality including gaps) may directly contribute to the dynamicity, preciseness, fairness, and reproducibility of the algorithm. In Study I, two sampling

frequencies (110 Hz and 154 Hz) were tested and found to produce equally accurate results. In free-living settings, sensor data may include extra noise and gaps, as reported in Study II. In Study I, conducted in laboratory settings, coverage could not be realistically evaluated. However, another study from the same database as studies II and III reported a median coverage rate of 69 % over 76 participants for an ultra-wideband (UWB) radar, which is used similarly to the FMCW radar. They used WiFi for data transfer. In studies II-III, the wearable sensor data median daily coverage was at 78 %, indicating that some non-wear time was typical when using the selected wearable sensor. User experience may affect data coverage, especially in the case of wearables. With chronic disease patients, the data collection should not add burden or hinder the patients' daily living.

The EHR data in Study IV suffered from missing data, too, but also from human errors and inconsistent entries. Methods more robust towards missing data or noise and errors could possibly overcome some of the data accuracy or quality requirements but may create new requirements, e.g., bigger masses of training data.

6.4 Limitations

The substudies were based on limited data collections and the data collection methods imposed their own limitations. The proposed algorithms additionally have their own limitations.

Study I inspected healthy participants in controlled laboratory settings with a single person on a bed and is not directly applicable to distinguishing people from one another (such as two people on a double bed or a nurse in a hospital or nursing home). The number of participants was low and included but two females. Some of the medical device reference for respiration had to be replaced by non-medical reference due to poor signal quality. There were minimal extra motions, which will likely happen in free-living monitoring conditions. The study only included simulated hypopnoea periods, where the participants were instructed to perform shallow respiration. Moreover, future work should ensure that full cessations of respiration can be reliably distinguished from leaving the radar's field of view. Finally, the proposed algorithms introduce some delay and thus the approach may not be applicable for truly real-time monitoring.

Studies II and III inspected mostly participants of Caucasian ethnicity with only a

few non-Caucasian participants. The effect of the severity of the conditions was not investigated. The studies used a single wearable device and the HR, IBI, respiratory rate, and skin temperature readily provided by the device, processed from the relevant sensors using the manufacturer's algorithms. The HRV analysis was conducted over each full window inspected, instead of averaging short-term measures over the inspected periods, and can therefore describe different physiological phenomena as compared to typical commercial device HRV measures. The study analyzed two hour windows of physiological data leading up to each PRO instance, whereas other periods with respect to the PRO may be equally valid. The presented studies only combined different modalities for HRR analysis but did not assess combinations of signals/modalities with respect to PROs. Although, the presented results may help form more informed signal combinations in future work. Furthermore, an extended study period would allow for more variations in the PROs, hence enabling analysis of changes in biometrics with respect to changes in PROs.

The study data in Study IV consisted of an anonymous database, adding noise and gaps as compared to the original EHR. Sex was missing for the majority of patients. The examined models were trained from scratch using standard English tokenizers in order to perform a fair comparison between the models, whereas the strength of transformer models is typically in massive pre-trained models. For interpretability experiments, only individual examples were analyzed. Finally, in a real-life use case, mortality risk prediction may be an ethically questionable prediction target; it may not help direct more resources to high risk patients but instead might lead to pulling resources from high risk patients, even if an effective intervention or treatment existed and could change the outcome. The model users (clinicians) should be educated on the model principles and proper regulation would be needed to make sure the final decision makers are accountable for using the predicted information to the patient's benefit.

7 CONCLUSION

This thesis studied time series based algorithms for decision support in chronic diseases. Four studies were presented, covering time series analytics in three case studies where the application scope varied from overnight to monthly monitoring. The study data were collected from chronic disease patients and healthy individuals, focusing on cardiorespiratory data. The participants included patients visiting a hospital, outpatients in free-living settings, and healthy subjects in laboratory settings. The research questions urged the search for pragmatic and purposeful solutions that would be consequential for clinical algorithm development across application time spans, beyond any specific chronic disease. While the intentionally wide questions could not be fully answered within the scope of this thesis, the thesis presented relevant scientific contributions to address them.

The studies presented contact-free radar based algorithms for overnight respiratory and heart rate monitoring, wearable sensor based algorithms for fatigue, sleep, and daytime sleepiness assessment over weeks, and EHR based algorithms for 6-month mortality risk assessment. Study I presented algorithms for contact-free vital sign monitoring and tested them on a broad range of both respiratory and heart rates, providing insights on their performance in clinically meaningful conditions. Moreover, the work was among the first to demonstrate HRV analysis using an FMCW radar, valuable for further sleep analysis. Studies II and III explored the objective assessment of fatigue, sleep, and daytime sleepiness using a multi-modal wearable sensor. Quantifying the changes in HRQoL would be highly important for example for drug development. Studies II and III included chronic disease patients in a feasibility study conducted in free-living settings, addressing the lack of similar study data from chronic disease patients. Furthermore, Study II illustrated the promise that context-bound features from continuous monitoring can have in health state assessment. Study IV was one of the earliest works to apply transformer neural networks on multi-modal EHR data, furthermore focusing on chronic disease patients. The

study argued that transformer models beyond BERT can achieve improved results with a fraction (here, only 5 %) of the model size, and exemplified gaps in model interpretability that should be addressed before clinical use.

For algorithms closer to real-time applications, the requirements for accurate high frequency data collection and algorithm latency became more crucial, whereas longer application time spans could perform on episodic clinical data. In free-living settings, controlling for the measurement context provided additional value to the time series analytics, even though the context (low intensity activity) was achieved without any guidance to the participants. The quality of the reference data was a major limitation for both model and data driven algorithms. Accounting for the repeated measures study design in algorithm evaluation was found especially important when the reference was based on subjective evaluation.

Both model and data driven approaches provided reasonable performance and demonstrated the ability to capture time dependent variations in health data. Model driven algorithms were found more suitable to high frequency sensor signal processing. Deep learning methods enable the detection of complex time-dependent patterns but currently lack explainability, which typically hinders their clinical uptake. Deep learning methods benefit from good quality training reference, and from specialized training on similar inter-related physiological patterns rather than a wide spectrum of physiological phenomena. They may, nevertheless, be efficient in pattern recognition even from episodic multi-modal data.

The data collection method plays a major role in algorithm development and in real-life performance. Data quality and coverage may set limitations to the algorithm; how the input sequence can be selected and processed, whether model or data driven methods should be used, and how accurate the algorithm can be. Maximizing data coverage enables more flexibility in input sequence selection and more accurate analyses. Additionally, albeit not specifically inspected in this thesis, noise or gaps in high coverage data may convey actionable information as well. For example, noise in radar data may indicate motions providing information about sleep quality, or the lack of hospital visits may indicate a stable patient status.

Contact-free monitoring showed encouraging results with 91 % and 81 % correlations to reference in still laying positions for a range of respiratory and heart rates, respectively, whereas motion artefacts posed the largest limitation. Continuous monitoring with a wearable ECG based device achieved high wear-times with

78 % coverage and little noise (less than 0.5 %) for cardiorespiratory signals. User-friendly (easy and comfortable to use) sensors can avoid creating extra burden from the measurement and avoid excess non-wear time. Electronic health records can provide a pervasive view to patient history but systematic errors or gaps, due to privacy regulations or other reasons, may challenge algorithm development.

Time series analytics offer a variety of powerful methods to be harnessed for chronic disease management. The conclusions of this thesis are based on limited data and many questions remain around the practical clinical deployment of any of the presented algorithms. All the presented studies comprised mostly Caucasian ethnicities, and Study I included a small set of only healthy participants and should be further validated on sleep apnoea patients. Study I was conducted in controlled laboratory settings with low motion and for one participant at a time, Studies II and III exploited readily available derived measurements instead of raw signal data, and Study IV trained transformers from scratch with standard English tokenizers on highly gapped data.

In future work, improvements in algorithm delay and robustness to motion artefacts may be needed for real-time applications, along with improved means to exploit the information in data gaps. For assessing patient reported outcomes, other signal time windows with respect to the PRO timing should be evaluated. Future research may additionally consider semi-supervised data driven methods to overcome challenges with sub-optimal reference data, and longer study periods to explore personalized solutions. For long-term non-frequently sampled data, the explainability of attention based models need further exploration, and may prove especially useful in the case of multi-modal input data. Moreover, pre-trained models and tokenizers specialized in electronic health records are needed to fully exploit the potential of transformer models. Finally, major focus should be directed at the ethical principles of clinical time series analytics applications.

Future patient monitoring for chronic diseases can be expected to move away from the clinical environment and closer to the everyday life, collecting data where it matters the most. The large volumes of data from continuous monitoring could fill in gaps in the EHR data, linking multi-modal data such as lab values and imaging results to the free-living measurements. The combined depiction of the patient's life with the chronic disease may promote early detection of clinically meaningful changes in patient state and create more predictive and personalized health solutions.

REFERENCES

[1] S. James, *Wearable technology market to hit $186.14 billion by 2030: Grand view research, inc.* Accessed 14 July 2023, Nov. 2022. [Online]. Available: https://www.prnewswire.co.uk/news-releases/wearable-technology-market-to-hit-186-14-billion-by-2030-grand-view-research-inc-301687632.html.

[2] GBD 2017 Causes of Death Collaborators, "Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the global burden of disease study 2017," *The Lancet*, vol. 392, pp. 1736–1788, 10159 2018. DOI: 10.1016/S0140-6736(18)32203-7.

[3] GBD 2019 Diseases and Injuries Collaborators, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019," *The Lancet*, vol. 396, pp. 1204–1222, 10258 2020. DOI: 10.1016/S0140-6736(20)30925-9.

[4] I. Jaussent, J. Bouyer, M.-L. Ancelin, C. Berr, A. Foubert-Samier, K. Ritchie, M. M. Ohayon, A. Besset, and Y. Dauvilliers, "Excessive sleepiness is predictive of cognitive decline in the elderly," *Sleep*, vol. 35, no. 9, pp. 1201–1207, 2012. DOI: 10.5665/sleep.2070.

[5] M. R. Fu, "Real-time detection and management of chronic illnesses," *mHealth*, vol. 7, no. 1, Jan. 2021. DOI: 10.21037/mhealth-2020-2.

[6] J. G. Betts, P. Desaix, E. Johnson, J. E. Johnson, O. Korol, D. Kruse, B. Poe, J. A. Wise, M. Womble, and K. A. Young, *Anatomy and Physiology 2e*. OpenStax, 2022. [Online]. Available: https://openstax.org/details/books/anatomy-and-physiology-2e.

[7] M. M. Churpek, R. Adhikari, and D. P. Edelson, "The value of vital sign trends for detecting clinical deterioration on the wards," *Resuscitation*, vol. 102, May 2016. DOI: 10.1016/j.resuscitation.2016.02.005.

[8] F. Baschieri and P. Cortelli, "Circadian rhythms of cardiovascular autonomic function: Physiology and clinical implications in neurodegenerative diseases," *Autonomic Neuroscience*, vol. 217, pp. 91–101, 2019. DOI: 10.1016/j.autneu.2019.01.009.

[9] R. M. Ahmed, Y. D. Ke, S. Vucic, L. M. Ittner, W. Seeley, J. R. Hodges, O. Piguet, G. Halliday, and M. C. Kiernan, "Physiological changes in neurodegeneration — mechanistic insights and clinical utility," *Nature Reviews Neurology*, vol. 14, pp. 259–271, 2018. DOI: 10.1038/nrneurol.2018.23.

[10] J. Steventon, J. Collett, H. Furby, K. Hamana, C. Foster, P. O'Callaghan, A. Dennis, R. Armstrong, A. Németh, A. Rosser, K. Murphy, L. Quinn, M. Busse, and H. Dawes, "Alterations in the metabolic and cardiorespiratory response to exercise in Huntington's disease," *Parkinsonism & Related Disorders*, vol. 54, pp. 56–61, 2018. DOI: 10.1016/j.parkreldis.2018.04.014.

[11] K. B. Roberson, J. F. Signorile, C. Singer, K. A. Jacobs, M. Eltoukhy, N. Ruta, N. Mazzei, and A. N. Buskard, "Hemodynamic responses to an exercise stress test in Parkinson's disease patients without orthostatic hypotension," *Applied Physiology, Nutrition, and Metabolism*, vol. 44, no. 7, pp. 751–758, 2018. DOI: 10.1139/apnm-2018-0638.

[12] O. Dogdu, M. Yarlioglues, M. G. Kaya, I. Ardic, N. Oguzhan, M. Akpek, O. Sahin, L. Akyol, S. Kelesoglu, F. Koc, I. Ozdogru, and A. Oguzhan, "Deterioration of heart rate recovery index in patients with systemic lupus erythematosus," *The Journal of Rheumatology*, vol. 37, no. 12, pp. 2511–2515, 2010. DOI: 10.3899/jrheum.100163.

[13] B. Sarli, Y. Dogan, O. Poyrazoglu, A. O. Baktir, A. Eyvaz, E. Altinkaya, A. Tok, E. Donudurmaci, M. Ugurlu, A. Ortakoyluoglu, H. Saglam, and H. Arinc, "Heart rate recovery is impaired in patients with inflammatory bowel diseases," *Medical Principles and Practice*, vol. 25, pp. 363–367, 2016. DOI: 10.1159/000446318.

[14] P. Bienias, M. Ciurzyński, A. Chrzanowska, I. Dudzik-Niewiadomska, K. Irzyk, K. Oleszek, A. Kalińska-Bienias, B. Kisiel, W. Tłustochowicz, and P. Pruszczyk, "Attenuated post-exercise heart rate recovery in patients with systemic lupus erythematosus: The role of disease severity and beta-blocker treatment," *Lupus*, vol. 27, no. 2, pp. 217–224, 2018. DOI: 10.1177/0961203317716318.

[15] T. Peçanha, R. Rodrigues, A. J. Pinto, A. L. Sá-Pinto, L. Guedes, K. Bonfiglioli, B. Gualano, and H. Roschel, "Chronotropic incompetence and reduced heart rate recovery in rheumatoid arthritis," *JCR: Journal of Clinical Rheumatology*, vol. 24, no. 7, pp. 375–380, 2018. DOI: 10.1097/RHU.0000000000000745.

[16] A. C. Guyton and J. E. Hall, *Textbook of Medical Physiology*, 12th ed. Saunders Elsevier, 2011, ISBN: 978-1-4160-4574-8.

[17] E. Haug, O. Sand, Ø. V. Sjaastad, and K. C. Toverud, *Ihmisen fysiologia*, 5th ed. Sanoma Pro Oy, 2012, ISBN: 978-952-63-1129-6.

[18] R. Gordan, J. K. Gwathmey, and L.-H. Xie, "Autonomic and endocrine control of cardiovascular function," *World Journal of Cardiology*, vol. 7, no. 4, pp. 204–214, 2015. DOI: 10.4330/wjc.v7.i4.204.

[19] I. Nederend, M. Jongbloed, E. De Geus, N. Blom, and A. Ten Harkel, "Postnatal cardiac autonomic nervous control in pediatric congenital heart disease," *Journal of Cardiovascular Development and Disease*, vol. 3, no. 2, Apr. 2016. DOI: 10.3390/jcdd3020016.

[20] E. A. Wehrwein, H. S. Orer, and S. M. Barman, "Overview of the anatomy, physiology, and pharmacology of the autonomic nervous system," *Comprehensive Physiology*, vol. 6, pp. 1239–1278, 2016. DOI: 10.1002/cphy.c150037.

[21] K. Yamakawa, P. S. Rajendran, T. Takamiya, D. Yagishita, E. L. So, A. Mahajan, K. Shivkumar, and M. Vaseghi, "Vagal nerve stimulation activates vagal afferent fibers that reduce cardiac efferent parasympathetic effects," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 309, no. 9, H1579–H1590, 2015. DOI: 10.1152/ajpheart.00558.2015.

[22] S. Bibevski and M. E. Dunlap, "Evidence for impaired vagus nerve activity in heart failure," *Heart Failure Reviews*, vol. 16, pp. 129–135, 2011. DOI: 10.1007/s10741-010-9190-6.

[23] M. Vaseghi, S. Salavatian, P. S. Rajendran, D. Yagishita, W. R. Woodward, D. Hamon, K. Yamakawa, T. Irie, B. A. Habecker, and K. Shivkumar, "Parasympathetic dysfunction and antiarrhythmic effect of vagal nerve stimulation following myocardial infarction," *JCI Insight*, vol. 2, no. 16, 2017. DOI: 10.1172/jci.insight.86715.

[24] P. Zalewski, J. Słomko, and M. Zawadka-Kunikowska, "Autonomic dysfunction and chronic disease," *British Medical Bulletin*, vol. 128, no. 1, pp. 61–74, Dec. 2018. DOI: 10.1093/bmb/ldy036.

[25] I. G. Boelhouwer, W. Vermeer, and T. van Vuuren, "Work ability, burnout complaints, and work engagement among employees with chronic diseases: Job resources as targets for intervention?" *Frontiers in Psychology*, vol. 11, 1805 2020. DOI: 10.3389/fpsyg.2020.01805.

[26] J. Schram, M. Schuring, K. Oude Hengel, A. Burdorf, and S. Robroek, "The influence of chronic diseases and poor working conditions in working life expectancy across educational levels among older employees in the Netherlands," *Scandinavian journal of work, environment & health*, vol. 48, pp. 391–398, 5 2022. DOI: 10.5271/sjweh.4028.

[27] J. H. Fong, "Disability incidence and functional decline among older adults with major chronic diseases," *BMC Geriatrics*, vol. 19, no. 323, 2019. DOI: 10.1186/s12877-019-1348-z.

[28] Y. Goërtz, A. Braamse, M. Spruit, and et al., "Fatigue in patients with chronic disease: Results from the population-based lifelines cohort study," *Scientific Reports*, vol. 11, no. 20977, 2021. DOI: 10.1038/s41598-021-00337-z.

[29] A. W. Vaes, Y. M. J. Goërtz, M. van Herck, R. J. H. C. G. Beijers, M. van Beers, C. Burtin, D. J. A. Janssen, A. M. W. J. Schols, and M. A. Spruit, "Physical and mental fatigue in people with non-communicable chronic diseases," *Annals of Medicine*, vol. 54, no. 1, pp. 2522–2534, 2022. DOI: 10.1080/07853890.2022.2122553.

[30] W. Lee, S. Lee, H. Ryu, Y. Chung, and W. Kim, "Quality of life in patients with obstructive sleep apnea: Relationship with daytime sleepiness, sleep quality, depression, and apnea severity," *Chronic Respiratory Disease*, vol. 13, no. 1, pp. 33–39, 2016. DOI: 10.1177/1479972315606312.

[31] G. Deuschl, E. Beghi, F. Fazekas, T. Varga, K. A. Christoforidi, E. Sipido, C. L. Bassetti, T. Vos, and V. L. Feigin, "The burden of neurological diseases in europe: An analysis for the global burden of disease study 2017," *The Lancet Public Health*, vol. 5, no. 10, e551–e567, 2020. DOI: 10.1016/S2468-2667(20)30190-0.

[32] S. Javaheri, F. Barbe, F. Campos-Rodriguez, J. A. Dempsey, R. Khayat, S. Javaheri, A. Malhotra, M. A. Martinez-Garcia, R. Mehra, A. I. Pack, V. Y. Polotsky, S. Redline, and V. K. Somers, "Sleep apnea: Types, mechanisms, and clinical cardiovascular consequences," *Journal of the American College of Cardiology*, vol. 69, no. 7, pp. 841–858, 2017. DOI: 10.1016/j.jacc.2016.11.069.

[33] A. S. Jordan, D. G. McSharry, and A. Malhotra, "Adult obstructive sleep apnoea," *Lancet*, vol. 383, no. 9918, pp. 736–747, Feb. 2014. DOI: 10.1016/S0140-6736(13)60734-5.

[34] C. Arnaud, T. Bochaton, J.-L. Pépin, and E. Belaidi, "Obstructive sleep apnoea and cardiovascular consequences: Pathophysiological mechanisms," *Archives of Cardiovascular Diseases*, vol. 113, no. 5, pp. 350–358, May 2020. DOI: 10.1016/j.acvd.2020.01.003.

[35] P. Maresova, J. Hruska, B. Klimova, S. Barakovic, and O. Krejcar, "Activities of daily living and associated costs in the most widespread neurodegenerative diseases: A systematic review," *Clinical Interventions in Aging*, vol. 15, pp. 1841–1862, Apr. 2020. DOI: 10.2147/CIA.S264688.

[36] A. H. Hristova and W. C. Koller, "Early Parkinson's disease: What is the best approach to treatment," *Drugs & Aging*, vol. 17, no. 3, pp. 165–181, Sep. 2000. DOI: 10.2165/00002512-200017030-00002.

[37] M. Politis, K. Wu, S. Molloy, P. G. Bain, K. R. Chaudhuri, and P. Piccini, "Parkinson's disease symptoms: The patient's perspective," *Movement Disorders*, vol. 25, no. 11, pp. 1646–1651, Aug. 2010. DOI: 10.1002/mds.23135.

[38] W. Yang, J. L. Hamilton, C. Kopil, J. C. Beck, C. M. Tanner, R. L. Albin, E. Ray Dorsey, N. Dahodwala, I. Cintina, P. Hogan, and T. Thompson, "Current and projected future economic burden of Parkinson's disease in the U.S.," *npj Parkinson's Disease*, vol. 6, no. 1, pp. 1–9, Jul. 2020. DOI: 10.1038/s41531-020-0117-1.

[39] I. Rodríguez-Santana, T. Mestre, F. Squitieri, R. Willock, A. Arnesen, A. Clarke, B. D'Alessio, A. Fisher, R. Fuller, J. L. Hamilton, H. Hubberstey, C. Stanley, L. Vetter, M. Winkelmann, M. Doherty, Y. Wu, A. Finnegan, and S. Frank, "Economic burden of Huntington disease in Europe and the USA: Results from the Huntington's disease burden of illness study," *European Journal of Neurology*, vol. 30, no. 4, pp. 1109–1117, Apr. 2023. DOI: 10.1111/ene.15645.

[40] R. Matos, L. Lencastre, V. Rocha, S. Torres, F. Vieira, M. R. Barbosa, J. Ascenção, and M. P. Guerra, "Quality of life in patients with inflammatory bowel disease: The role of positive psychological factors," *Health Psychology and Behavioral Medicine*, vol. 9, no. 1, pp. 989–1005, Jan. 2021. DOI: 10.1080/21642850.2021.2007098.

[41] D. R. Van Langenberg and P. R. Gibson, "Systematic review: Fatigue in inflammatory bowel disease," *Alimentary Pharmacology & Therapeutics*, vol. 32, no. 2, pp. 131–143, Apr. 2010. DOI: 10.1111/j.1365-2036.2010.04347.x.

[42] T. Both, V. A. Dalm, P. M. Van Hagen, and P. L. Van Daele, "Reviewing primary Sjögren's syndrome: Beyond the dryness - from pathophysiology to diagnosis and treatment," *International Journal of Medical Sciences*, vol. 14, no. 3, pp. 191–200, 2017. DOI: 10.7150/ijms.17718.

[43] A.-L. Stefanski, C. Tomiak, U. Pleyer, T. Dietrich, G. R. Burmester, and T. Dörner, "The diagnosis and treatment of Sjögren's syndrome," *Deutsches Ärzteblatt international*, May 2017. DOI: 10.3238/arztebl.2017.0354.

[44] W. Grassi, R. De Angelis, G. Lamanna, and C. Cervini, "The clinical features of rheumatoid arthritis," *European Journal of Radiology*, vol. 27, S18–S24, May 1998. DOI: 10.1016/S0720-048X(98)00038-2.

[45] A.-F. Radu and S. G. Bungau, "Management of rheumatoid arthritis: An overview," *Cells*, vol. 10, no. 11, Oct. 2021. DOI: 10.3390/cells10112857.

[46] C.-F. Kuo, I. J. Chou, F. Rees, M. J. Grainge, P. Lanyon, G. Davenport, C. D. Mallen, T.-T. Chung, J.-S. Chen, W. Zhang, and M. Doherty, "Temporal relationships between systemic lupus erythematosus and comorbidities," *Rheumatology*, vol. 58, no. 5, pp. 840–848, May 2019. DOI: 10.1093/rheumatology/key335.

[47] P. P. Smith and C. Gordon, "Systemic lupus erythematosus: Clinical presentations," *Autoimmunity Reviews*, vol. 10, no. 1, pp. 43–45, Nov. 2010. DOI: 10.1016/j.autrev.2010.08.016.

[48] C. Yildirim-Toruner and B. Diamond, "Current and novel therapeutics in the treatment of systemic lupus erythematosus," *Journal of Allergy and Clinical Immunology*, vol. 127, no. 2, pp. 303–312, Feb. 2011. DOI: 10.1016/j.jaci.2010.12.1087.

[49] H. Ritchie, F. Spooner, and M. Roser, "Causes of death," *Our World in Data*, 2023, Accessed 14 July 2023. [Online]. Available: https://ourworldindata.org/causes-of-death.

[50] A. Nanjo, H. Evans, K. Direk, A. C. Hayward, A. Story, and A. Banerjee, "Prevalence, incidence, and outcomes across cardiovascular diseases in homeless individuals using national linked electronic health records," *European Heart Journal*, vol. 41, no. 41, pp. 4011–4020, Nov. 2020. DOI: 10.1093/eurheartj/ehaa795.

[51] H. A. DeVon, K. Vuckovic, C. J. Ryan, S. Barnason, J. J. Zerwic, B. Pozehl, P. Schulz, Y. Seo, and L. Zimmerman, "Systematic review of symptom clusters in cardiovascular disease," *European Journal of Cardiovascular Nursing*, vol. 16, no. 1, pp. 6–17, Jan. 2017. DOI: 10.1177/1474515116642594.

[52] M. Altini, P. Casale, J. Penders, and O. Amft, "Cardiorespiratory fitness estimation in free-living using wearable sensors," *Artificial Intelligence in Medicine*, vol. 68, pp. 37–46, 2016. DOI: 10.1016/j.artmed.2016.02.002.

[53] D. Spathis, I. Perez-Pozuelo, T. I. Gonzales, Y. Wu, S. Brage, N. Wareham, and C. Mascolo, "Longitudinal cardio-respiratory fitness prediction through wearables in free-living environments," *npj Digital Medicine*, vol. 5, no. 1, 2022. DOI: 10.1038/s41746-022-00719-1.

[54] D. J. Miller, C. Sargent, and G. D. Roach, "A validation of six wearable devices for estimating sleep, heart rate and heart rate variability in healthy adults," *Sensors*, vol. 22, no. 16, 2022. DOI: 10.3390/s22166317.

[55] H. Kinnunen, A. Rantanen, T. Kenttä, and H. Koskimäki, "Feasible assessment of recovery and cardiovascular health: accuracy of nocturnal HR and HRV assessed via ring PPG in comparison to medical grade ECG," *Physiological Measurement*, vol. 41, no. 4, 2020. DOI: 10.1088/1361-6579/ab840a.

[56] A. J. Boe, L. L. McGee Koch, M. K. O'Brien, N. Shawen, J. A. Rogers, R. L. Lieber, K. J. Reid, P. C. Zee, and A. Jayaraman, "Automating sleep stage classification using wireless, wearable sensors," *npj Digital Medicine*, vol. 2, no. 1, 2019. DOI: 10.1038/s41746-019-0210-1.

[57] J. M. Kortelainen, M. van Gils, and J. Pärkkä, "Multichannel bed pressure sensor for sleep monitoring," in *2012 Computing in Cardiology*, Sep. 2012, pp. 313–316.

[58] F. Shaffer, Z. M. Meehan, and C. L. Zerr, "A critical review of ultra-short-term heart rate variability norms research," *Frintiers in Neuroscience*, vol. 14, no. 594880, 2020. DOI: 10.3389/fnins.2020.594880.

[59] S.-C. Fang, Y.-L. Wu, and P.-S. Tsai, "Heart rate variability and risk of all-cause death and cardiovascular events in patients with cardiovascular disease: A meta-analysis of cohort studies," *Biological Research For Nursing*, vol. 22, no. 1, pp. 45–56, Jan. 2020. DOI: 10.1177/1099800419877442.

[60] P. Hämmerle, C. Eick, S. Blum, V. Schlageter, A. Bauer, K. D. Rizas, C. Eken, M. Coslovsky, S. Aeschbacher, P. Krisai, P. Meyre, J.-M. Vesin, N. Rodondi, E. Moutzouri, J. Beer, G. Moschovitis, R. Kobza, M. Di Valentino, V. D. A. Corino, R. Laureanti, L. Mainardi, L. H. Bonati, C. Sticherling, D. Conen, S. Osswald, M. Kühne, C. S. Zuern, and S.-A. S. Investigators, "Heart rate variability triangular index as a predictor of cardiovascular mortality in patients with atrial fibrillation," *Journal of the American Heart Association*, vol. 9, no. 15, Aug. 2020. DOI: 10.1161/JAHA.120.016075.

[61] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in Public Health*, vol. 5, 2017. DOI: 10.3389/fpubh.2017.00258.

[62] Task Force Of The European Society Of Cardiology The North American Society of Pacing Electrophysiology, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, Mar. 1996. DOI: 10.1161/01.CIR.93.5.1043.

[63] S. Mahdiani, V. Jeyhani, M. Peltokangas, and A. Vehkaoja, "Is 50 Hz high enough ECG sampling frequency for accurate HRV analysis?" In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biol-*

ogy Society (EMBC), Milan: IEEE, Aug. 2015, pp. 5948–5951. DOI: 10.1109/EMBC.2015.7319746.

[64] A. Nicolò, C. Massaroni, E. Schena, and M. Sacchetti, "The importance of respiratory rate monitoring: From healthcare to sport and exercise," *Sensors*, vol. 20, no. 21, 2020. DOI: 10.3390/s20216396.

[65] X. Xu, J. Zhu, C. Chen, X. Zhang, Z. Lian, and Z. Hou, "Application potential of skin temperature for sleep-wake classification," *Energy and Buildings*, vol. 266, 2022. DOI: 10.1016/j.enbuild.2022.112137.

[66] K. A. Herborn, J. L. Graves, P. Jerem, N. P. Evans, R. Nager, D. J. McCafferty, and D. E. McKeegan, "Skin temperature reveals the intensity of acute stress," *Physiology & Behavior*, vol. 152, pp. 225–230, Dec. 2015. DOI: 10.1016/j.physbeh.2015.09.032.

[67] S. Jardak, T. Kiuru, M. Metso, P. Pursula, J. Häkli, M. Hirvonen, S. Ahmed, and M.-S. Alouini, "Detection and localization of multiple short range targets using FMCW radar signal," in *2016 Global Symposium on Millimeter Waves (GSMM) & ESA Workshop on Millimetre-Wave Technology and Applications*, Jun. 2016. DOI: 10.1109/GSMM.2016.7500332.

[68] T. Kiuru, M. Metso, S. Jardak, P. Pursula, J. Häkli, M. Hirvonen, and R. Sepponen, "Movement and respiration detection using statistical properties of the FMCW radar signal," in *2016 Global Symposium on Millimeter Waves (GSMM) & ESA Workshop on Millimetre-Wave Technology and Applications*, 2016, pp. 1–4. DOI: 10.1109/GSMM.2016.7500331.

[69] A. Ometov, V. Shubina, L. Klus, J. Skibińska, S. Saafi, P. Pascacio, L. Flueratoru, D. Q. Gaibor, N. Chukhno, O. Chukhno, A. Ali, A. Channa, E. Svertoka, W. B. Qaim, R. Casanova-Marqués, S. Holcer, J. Torres-Sospedra, S. Casteleyn, G. Ruggeri, G. Araniti, R. Burget, J. Hosek, and E. S. Lohan, "A survey on wearable technology: History, state-of-the-art and current challenges," *Computer Networks*, vol. 193, 2021. DOI: 10.1016/j.comnet.2021.108074.

[70] D. Fuller, E. Colwell, J. Low, K. Orychock, M. A. Tobin, B. Simango, R. Buote, D. Van Heerden, H. Luan, K. Cullen, L. Slade, and N. G. A. Taylor, "Reliability and validity of commercially available wearable devices for mea-

suring steps, energy expenditure, and heart rate: Systematic review," *JMIR mHealth and uHealth*, vol. 8, no. 9, 2020. DOI: 10.2196/18694.

[71] E. E. Dooley, N. M. Golaszewski, and J. B. Bartholomew, "Estimating accuracy at exercise intensities: A comparative study of self-monitoring heart rate and physical activity wearable devices," *JMIR mHealth and uHealth*, vol. 5, no. 3, 2017. DOI: 10.2196/mhealth.7043.

[72] C. M. Areia, M. Santos, S. Vollam, M. Pimentel, L. Young, C. Roman, J. Ede, P. Piper, E. King, O. Gustafson, M. Harford, A. Shah, L. Tarassenko, and P. Watkinson, "A chest patch for continuous vital sign monitoring: Clinical validation study during movement and controlled hypoxia," *Journal of Medical Internet Research*, vol. 23, no. 9, Sep. 2021. DOI: 10.2196/27547.

[73] P. L. Rajbhandary and G. Nallathambi, "Feasibility of continuous monitoring of core body temperature using chest-worn patch sensor," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Montreal, QC, Canada: IEEE, Jul. 2020, pp. 4652–4655. DOI: 10.1109/EMBC44109.2020.9175579.

[74] M. Lehne, J. Sass, A. Essenwanger, J. Schepers, and S. Thun, "Why digital medicine depends on interoperability," *npj Digital Medicine*, vol. 2, no. 1, Aug. 2019. DOI: 10.1038/s41746-019-0158-1.

[75] H. Tolonen, V. Salomaa, J. Torppa, J. Sivenius, P. Immonen-Räihä, A. Lehtonen, and R. Finstroke, "The validation of the Finnish hospital discharge register and causes of death register data on stroke diagnoses," *European journal of cardiovascular prevention and rehabilitation*, vol. 14, no. 3, pp. 380–385, 2007. DOI: 10.1097/01.hjr.0000239466.26132.f2.

[76] P. Pajunen, H. Koukkunen, M. Ketonen, T. Jerkkola, P. Immonen-Räihä, P. Kärjä-Koskenkari, M. Mähönen, M. Niemelä, K. Kuulasmaa, P. Palomäki, J. Mustonen, A. Lehtonen, M. Arstila, T. Vuorenmaa, S. Lehto, H. Miettinen, J. Torppa, J. Tuomilehto, Y. A. Kesäniemi, K. Pyörälä, and V. Salomaa, "The validity of the Finnish hospital discharge register and causes of death register data on coronary heart disease," *European Journal of Cardiovascular Prevention & Rehabilitation*, vol. 12, no. 2, pp. 132–137, 2005. DOI: 10.1097/00149831-200504000-00007.

[77] M. A. Vuori, J. A. Laukkanen, A. Pietilä, A. S. Havulinna, M. Kähönen, V. Salomaa, T. J. Niiranen, and the FinnGen investigators, "The validity of heart failure diagnoses in the Finnish hospital discharge register," *Scandinavian Journal of Public Health*, vol. 48, no. 1, pp. 20–28, 2020. DOI: 10.1177/1403494819847051.

[78] J. A. Hernesniemi, S. Mahdiani, L.-P. Lyytikäinen, T. Lehtimäki, M. Eskola, K. Nikus, K. Antila, and N. Oksala, "Cohort description for maddec – mass data in detection and prevention of serious adverse events in cardiovascular disease," in *EMBEC & NBC 2017*, H. Eskola, O. Väisänen, J. Viik, and J. Hyttinen, Eds., vol. 65, Singapore: Springer Singapore, 2018, pp. 1113–1116. DOI: 10.1007/978-981-10-5122-7_278.

[79] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967. DOI: 10.1109/TAU.1967.1161901.

[80] M. Peltola, "Role of editing of R-R intervals in the analysis of heart rate variability," *Frontiers in Physiology*, vol. 3, 2012. DOI: 10.3389/fphys.2012.00148.

[81] A. Voss, R. Schroeder, A. Heitmann, A. Peters, and S. Perz, "Short-term heart rate variability—influence of gender and age in healthy subjects," *PLOS ONE*, vol. 10, no. 3, A. V. Hernandez, Ed., Mar. 2015. DOI: 10.1371/journal.pone.0118308.

[82] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer, 2006.

[83] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data mining, inference, and prediction*, 2nd ed. Springer, 2008.

[84] G. Hinton, *A practical guide to training restricted Boltzmann machines*, Aug. 2012. [Online]. Available: https://www.csrc.ac.cn/upload/file/20170703/1499052743888438.pdf.

[85] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

[86] C. Huyen, *Designing machine learning systems: An iterative process for production-ready applications*, 1st ed. O'Reilly Media, Inc., 2022.

[87] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[88] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015. [Online]. Available: https://www.robots.ox.ac.uk/~vgg/publications/2015/Simonyan15/simonyan15.pdf.

[89] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241. DOI: 10.48550/arXiv.1505.04597.

[90] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. DOI: 10.48550/ARXIV.1706.03762.

[91] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations*, 2016. DOI: 10.48550/arXiv.1409.0473.

[92] L. Weng, *Attention? Attention!* Accessed 14 July 2023, Jun. 2018. [Online]. Available: https://lilianweng.github.io/posts/2018-06-24-attention/.

[93] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186. DOI: 10.48550/ARXIV.1810.04805.

[94] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems*, vol. 32, 2019. DOI: 10.48550/arXiv.1906.08237.

[95] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 37–42. DOI: 10.18653/v1/P19-3007.

[96] J. Z. Bakdash and L. R. Marusich, "Repeated measures correlation," *Frontiers in Psychology*, vol. 8, 2017. DOI: 10.3389/fpsyg.2017.00456.

[97] J. Bland and D. Altman, "Comparing methods of measurement: Why plotting difference against standard method is misleading," *The Lancet*, vol. 346, no. 8982, pp. 1085–1087, Oct. 1995. DOI: 10.1016/S0140-6736(95)91748-9.

[98] L. Chen, X. Ma, M. Chatterjee, J. M. Kortelainen, T. Ahmaniemi, W. Maetzler, P. Wang, and D. Zhang, "Fatigue and sleep assessment using digital sleep trackers: Insights from a multi-device pilot study," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Jul. 2022, pp. 1133–1136. DOI: 10.1109/EMBC48229.2022.9870923.

[99] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, "Smart homes that monitor breathing and heart rate," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15, New York, NY, USA: Association for Computing Machinery, 2015, pp. 837–846. DOI: 10.1145/2702123.2702200.

[100] A. A. Pramudita, D.-B. Lin, S.-N. Hsieh, E. Ali, H. H. Ryanu, T. Adiprabowo, and A. T. Purnomo, "Radar system for detecting respiration vital sign of live victim behind the wall," *IEEE Sensors Journal*, vol. 22, no. 15, pp. 14 670–14 685, Aug. 2022. DOI: 10.1109/JSEN.2022.3188165.

[101] S. Wang, A. Pohl, T. Jaeschke, M. Czaplik, M. Köny, S. Leonhardt, and N. Pohl, "A novel ultra-wideband 80 GHz FMCW radar system for contactless monitoring of vital signs," 2015, pp. 4978–4981. DOI: 10.1109/EMBC.2015.7319509.

[102] M. Arsalan, A. Santra, and C. Will, "Improved contactless heartbeat estimation in FMCW radar via Kalman filter tracking," *IEEE Sensors Letters*, vol. 4, no. 5, pp. 1–4, 7001304 May 2020. DOI: 10.1109/LSENS.2020.2983706.

[103] L. Anitori, A. de Jong, and F. Nennie, "FMCW radar for life-sign detection," in *2009 IEEE Radar Conference*, 2009, pp. 1–6. DOI: 10.1109/RADAR.2009.4976934.

[104] S. Lim, G. S. Jang, W. Song, B.-h. Kim, and D. H. Kim, "Non-contact vital signs monitoring of a patient lying on surgical bed using beamforming FMCW radar," *Sensors*, vol. 22, no. 21, 8167 Nov. 2022. DOI: 10.3390/s22218167.

[105] M. Weenk, H. v. Goor, B. Frietman, L. J. Engelen, C. J. v. Laarhoven, J. Smit, S. J. Bredie, and T. H. v. d. Belt, "Continuous monitoring of vital signs using wearable devices on the general ward: Pilot study," *JMIR mHealth and uHealth*, vol. 5, no. 7, e7208, Jul. 2017. DOI: 10.2196/mhealth.7208.

[106] S. Soon, H. Svavarsdottir, C. Downey, and D. G. Jayne, "Wearable devices for remote vital signs monitoring in the outpatient setting: An overview of the field," *BMJ Innovations*, vol. 6, no. 2, 2020. DOI: 10.1136/bmjinnov-2019-000354.

[107] E.-S. Väliaho, J. A. Lipponen, P. Kuoppa, T. J. Martikainen, H. Jäntti, T. T. Rissanen, M. Castrén, J. Halonen, M. P. Tarvainen, T. M. Laitinen, T. P. Laitinen, O. E. Santala, O. Rantula, N. S. Naukkarinen, and J. E. K. Hartikainen, "Continuous 24-h photoplethysmogram monitoring enables detection of atrial fibrillation," *Frontiers in Physiology*, vol. 12, 2022. DOI: 10.3389/fphys.2021.778775.

[108] R. Abdolkhani, K. Gray, A. Borda, and R. DeSouza, "Patient-generated health data management and quality challenges in remote patient monitoring," *JAMIA Open*, vol. 2, no. 4, pp. 471–478, Dec. 2019. DOI: 10.1093/jamiaopen/ooz036.

[109] D. Sokas, B. Paliakaitė, A. Rapalis, V. Marozas, R. Bailón, and A. Petrėnas, "Detection of walk tests in free-living activities using a wrist-worn device," *Frontiers in Physiology*, vol. 12, 2021. DOI: 10.3389/fphys.2021.706545.

[110] F. M. Iqbal, K. Lam, M. Joshi, S. Khan, H. Ashrafian, and A. Darzi, "Clinical outcomes of digital sensor alerting systems in remote monitoring: A systematic review and meta-analysis," *npj Digital Medicine*, vol. 4, no. 7, 2021. DOI: 10.1038/s41746-020-00378-0.

[111] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "RETAIN: interpretable predictive model in healthcare using reverse time attention mechanism," *CoRR*, vol. abs/1608.05745, 2016. DOI: 10.48550/arXiv.1608.05745.

[112] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proceedings of the 1st Machine Learning for Healthcare Conference*, Northeastern University, Boston, MA, USA: PMLR, Aug. 2016, pp. 301–318. DOI: 10.48550/arXiv.1511.05942.

[113] J. Shang, T. Ma, C. Xiao, and J. Sun, "Pre-training of graph augmented transformers for medication recommendation," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019*, ser. IJCAI International Joint Conference on Artificial Intelligence, 2019, pp. 5953–5959. DOI: 10.24963/ijcai.2019/825.

[114] Y. Li, S. Rao, J. R. A. Solares, *et al.*, "BEHRT: Transformer for electronic health records," *Scientific Reports*, vol. 10, no. 7155, 2020.

[115] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *npj Digital Medicine*, vol. 4, no. 86, 2021. DOI: 10.1038/s41746-021-00455-y.

[116] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," 2019. DOI: 10.48550/ARXIV.1904.05342.

[117] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific Reports*, vol. 6, no. 26094, May 2016. DOI: 10.1038/srep26094.

[118] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," *Journal of Biomedical Informatics*, vol. 69, pp. 218–229, 2017. DOI: 10.1016/j.jbi.2017.04.001.

[119] A. Rajkomar *et al.*, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 18, 2018. DOI: 10.1038/s41746-018-0029-1.

[120] E. Choi, Z. Xu, Y. Li, M. W. Dusenberry, G. Flores, Y. Xue, and A. M. Dai, "Learning the graphical structure of electronic health records with graph convolutional transformer," 2019. DOI: 10.48550/ARXIV.1906.04716.

[121] C. B. Hilton, A. Milinovich, C. Felix, N. Vakharia, T. Crone, C. Donovan, A. Proctor, and A. Nazha, "Personalized predictions of patient outcomes during and after hospitalization using artificial intelligence," *npj Digital Medicine*, vol. 3, no. 51, pp. 1–8, Apr. 2020. DOI: 10.1038/s41746-020-0249-z.

[122] H. Hemingway, F. W. Asselbergs, J. Danesh, R. Dobson, N. Maniadakis, A. Maggioni, G. J. M. van Thiel, M. Cronin, G. Brobert, P. Vardas, S. D. Anker, D. E. Grobbee, S. Denaxas, B. C. o. 2. a. Innovative Medicines Initiative 2nd programme Big Data for Better Outcomes, and industry partners including ESC, "Big data from electronic health records for early and late translational cardiovascular research: challenges and potential," *European Heart Journal*, vol. 39, no. 16, pp. 1481–1495, Aug. 2017. DOI: 10.1093/eurheartj/ehx487.

[123] M. Motwani, D. Dey, D. S. Berman, G. Germano, S. Achenbach, M. H. Al-Mallah, D. Andreini, M. J. Budoff, F. Cademartiri, T. Q. Callister, H.-J. Chang, K. Chinnaiyan, B. J. Chow, R. C. Cury, A. Delago, M. Gomez, H. Gransar, M. Hadamitzky, J. Hausleiter, N. Hindoyan, G. Feuchtner, P. A. Kaufmann, Y.-J. Kim, J. Leipsic, F. Y. Lin, E. Maffei, H. Marques, G. Pontone, G. Raff, R. Rubinshtein, L. J. Shaw, J. Stehli, T. C. Villines, A. Dunning, J. K. Min, and P. J. Slomka, "Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: A 5-year multicentre prospective registry analysis," *European Heart Journal*, vol. 38, no. 7, pp. 500–507, Jun. 2016. DOI: 10.1093/eurheartj/ehw188.

[124] J. A. Hernesniemi, S. Mahdiani, J. A. Tynkkynen, L.-P. Lyytikäinen, P. P. Mishra, T. Lehtimäki, M. Eskola, K. Nikus, K. Antila, and N. Oksala, "Exten-

sive phenotype data and machine learning in prediction of mortality in acute coronary syndrome – the MADDEC study," *Annals of Medicine*, vol. 51, no. 2, pp. 156–163, 2019. DOI: 10.1080/07853890.2019.1596302.

[125] F. Yang, J. Zhang, W. Chen, Y. Lai, Y. Wang, and Q. Zou, "DeepMPM: a mortality risk prediction model using longitudinal EHR data," *BMC Bioinformatics*, vol. 23, no. 423, Oct. 2022. DOI: 10.1186/s12859-022-04975-6.

[126] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Medical concept representation learning from electronic health records and its application on heart failure prediction," 2016. DOI: 10.48550/ARXIV.1602.03686.

[127] J. Chu, W. Dong, and Z. Huang, "Endpoint prediction of heart failure using electronic health records," *Journal of Biomedical Informatics*, vol. 109, no. 103518, Sep. 2020. DOI: 10.1016/j.jbi.2020.103518.

[128] M. J. Kolek, A. J. Graves, M. Xu, A. Bian, P. L. Teixeira, M. B. Shoemaker, B. Parvez, H. Xu, S. R. Heckbert, P. T. Ellinor, E. J. Benjamin, A. Alonso, J. C. Denny, K. G. M. Moons, A. K. Shintani, F. E. Harrell, D. M. Roden, and D. Darbar, "Evaluation of a prediction model for the development of atrial fibrillation in a repository of electronic medical records," *JAMA Cardiology*, vol. 1, no. 9, pp. 1007–1013, Dec. 2016. DOI: 10.1001/jamacardio.2016.3366.

[129] F. Rodriguez, S. Chung, M. R. Blum, A. Coulet, S. Basu, and L. P. Palaniappan, "Atherosclerotic cardiovascular disease risk prediction in disaggregated Asian and Hispanic subgroups using electronic health records," *JAHA*, vol. 8, no. e011874, Jul. 2019. DOI: 10.1161/JAHA.118.011874.

[130] "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension," vol. 26, pp. 1364–1374, Sep. 2020. DOI: 10.1038/s41591-020-1034-x.

[131] S. C. Shelmerdine, O. J. Arthurs, A. Denniston, and N. J. Sebire, "Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare," *BMJ Health & Care Informatics*, vol. 28, no. e100385, Aug. 2021. DOI: 10.1136/bmjhci-2021-100385.

[132] I. Scott, S. Carter, and E. Coiera, "Clinician checklist for assessing suitability of machine learning applications in healthcare," *BMJ Health & Care Informatics*, vol. 28, no. e100251, Feb. 2021. DOI: 10.1136/bmjhci-2020-100251.

[133] M. Van Smeden, G. Heinze, B. Van Calster, F. W. Asselbergs, P. E. Vardas, N. Bruining, P. De Jaegere, J. H. Moore, S. Denaxas, A. L. Boulesteix, and K. G. M. Moons, "Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease," *European Heart Journal*, vol. 43, no. 31, pp. 2921–2930, Aug. 2022. DOI: 10.1093/eurheartj/ehac238.

[134] T. J. Loftus, P. J. Tighe, T. Ozrazgat-Baslanti, J. P. Davis, M. M. Ruppert, Y. Ren, B. Shickel, R. Kamaleswaran, W. R. Hogan, J. R. Moorman, G. R. Upchurch, P. Rashidi, and A. Bihorac, "Ideal algorithms in healthcare: Explainable, dynamic, precise, autonomous, fair, and reproducible," *PLOS Digital Health*, vol. 1, no. 1, e0000006, Jan. 2022. DOI: 10.1371/journal.pdig.0000006.

[135] D. Fotopoulos, D. Filos, and I. Chouvarda, "Towards explainable and trustworthy AI for decision support in medicine: An overview of methods and good practices," *Aristotle Biomedical Journal*, vol. 3, no. 2, Feb. 2021. DOI: 10.26262/ABJ.V3I2.8105.

[136] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017. DOI: 10.48550/ARXIV.1705.07874.

[137] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[138] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20, New York, NY, USA: Association for Computing Machinery, Feb. 2020, pp. 180–186. DOI: 10.1145/3375627.3375830.

[139] B. Bai, J. Liang, G. Zhang, H. Li, K. Bai, and F. Wang, "Why attentions may not be interpretable?" In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Virtual Event Singapore: ACM, Aug. 2021, pp. 25–34. DOI: 10.1145/3447548.3467307.

[140] C. Gilon, J.-M. Grégoire, J. Hellinckx, S. Carlier, and H. Bersini, "Reproducibility of machine learning models for paroxysmal atrial fibrillation onset prediction," in *Computing in Cardiology 2022*, vol. 49, 2022. DOI: 10.22489/CinC.2022.171.

[141] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in ML-based science," 2022. DOI: 10.48550/ARXIV.2207.07048.

[142] *Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.)* May 2017. [Online]. Available: http://data.europa.eu/eli/reg/2017/745/oj.

[143] *EU AI Act: First regulation on artificial intelligence*, Last updated 14 June 2023. Accessed 5 Dec 2023, Jun. 2023. [Online]. Available: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[144] European Commission and Directorate-General for Communications Networks, Content and Technology, *Ethics guidelines for trustworthy AI*. Publications Office, 2019. DOI: 10.2759/346720.

[145] *Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment*, Last updated 8 March 2021. Accessed 5 Dec 2023, Jul. 2020. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[146] *Embla Titanium Clinical Manual*. 2007, Revision 3.0.

[147] The IDEA-FAST consortium, *D2.1: First study subject approvals package of the feasibility study (FS)*, Jul. 2020. [Online]. Available: https://idea-fast.eu/wp-content/uploads/2021/01/IDEA-FAST_D2.1_FS-Approvals_v1.0_IMI.pdf.

[148] E. Vildjiounaite, J. Kallio, V. Kyllönen, M. Nieminen, I. Määttänen, M. Lindholm, J. Mäntyjärvi, and G. Gimel'farb, "Unobtrusive stress detection on the basis of smartphone usage data," *Personal and Ubiquitous Computing*, vol. 22, no. 4, pp. 671–688, Aug. 2018. DOI: 10.1007/s00779-017-1108-z.

[149] J. W. Choi, D. H. Kim, D. L. Koo, Y. Park, H. Nam, J. H. Lee, H. J. Kim, S.-N. Hong, G. Jang, S. Lim, and B. Kim, "Automated detection of sleep apnea-hypopnea events based on 60 GHz frequency-modulated continuous-wave radar using convolutional recurrent neural networks: A preliminary report of a prospective cohort study," *Sensors*, vol. 22, no. 19, Sep. 2022. DOI: 10.3390/s22197177.

[150] Z. Zhuang, F. Wang, X. Yang, L. Zhang, C.-H. Fu, J. Xu, C. Li, and H. Hong, "Accurate contactless sleep apnea detection framework with signal processing and machine learning methods," *Methods*, vol. 205, pp. 167–178, Sep. 2022. DOI: 10.1016/j.ymeth.2022.06.013.

[151] R. P. Lamberts, J. Swart, B. Capostagno, T. D. Noakes, and M. I. Lambert, "Heart rate recovery as a guide to monitor fatigue and predict changes in performance parameters: Heart rate recovery to monitor of performance," *Scandinavian Journal of Medicine & Science in Sports*, vol. 20, no. 3, pp. 449–457, Jun. 2009. DOI: 10.1111/j.1600-0838.2009.00977.x.

[152] L. Djaoui, M. Haddad, K. Chamari, and A. Dellal, "Monitoring training load and fatigue in soccer players with physiological markers," *Physiology & Behavior*, vol. 181, pp. 86–94, Nov. 2017. DOI: 10.1016/j.physbeh.2017.09.004.

PUBLICATIONS

PUBLICATION
I

Vital sign monitoring using FMCW radar in various sleeping scenarios

E. Turppa*, J. M. Kortelainen, O. Antropov, and T. Kiuru

Sensors, vol. 20, no. 22:6505
DOI: 10.3390/s20226505

sensors

MDPI

Article

Vital Sign Monitoring Using FMCW Radar in Various Sleeping Scenarios

Emmi Turppa *, Juha M. Kortelainen, Oleg Antropov and Tero Kiuru

VTT Technical Research Centre of Finland Ltd., P.O. Box 1300, 33101 Tampere, Finland; juha.m.kortelainen@vtt.fi (J.M.K.); oleg.antropov@vtt.fi (O.A.); tero.kiuru@vtt.fi (T.K.)
* Correspondence: emmi.turppa@vtt.fi

Received: 2 October 2020; Accepted: 12 November 2020; Published: 14 November 2020

Abstract: Remote monitoring of vital signs for studying sleep is a user-friendly alternative to monitoring with sensors attached to the skin. For instance, remote monitoring can allow unconstrained movement during sleep, whereas detectors requiring a physical contact may detach and interrupt the measurement and affect sleep itself. This study evaluates the performance of a cost-effective frequency modulated continuous wave (FMCW) radar in remote monitoring of heart rate and respiration in scenarios resembling a set of normal and abnormal physiological conditions during sleep. We evaluate the vital signs of ten subjects in different lying positions during various tasks. Specifically, we aim for a broad range of both heart and respiration rates to replicate various real-life scenarios and to test the robustness of the selected vital sign extraction methods consisting of fast Fourier transform based cepstral and autocorrelation analyses. As compared to the reference signals obtained using Embla titanium, a certified medical device, we achieved an overall relative mean absolute error of 3.6% (86% correlation) and 9.1% (91% correlation) for the heart rate and respiration rate, respectively. Our results promote radar-based clinical monitoring by showing that the proposed radar technology and signal processing methods accurately capture even such alarming vital signs as minimal respiration. Furthermore, we show that common parameters for heart rate variability can also be accurately extracted from the radar signal, enabling further sleep analyses.

Keywords: biomedical monitoring; biomedical signal processing; contactless; health monitoring; heart rate; heart rate variability; millimeter wave radar; respiratory rate

1. Introduction

Monitoring vital signs is routine practice to detect patient deterioration at healthcare facilities. Changes in vital signs can indicate serious medical problems, and catching the early signs may improve survival rates for the relevant conditions [1]. Lately, the general population has become more interested in self-monitoring, which has provoked the emergence of numerous commercial wearable devices, particularly ones specialized in heart rate monitoring. Such wearable devices have also been examined in the context of monitoring healthcare patients [2,3]. Yet, wearable and other attachable devices can cause eczema and they depend on a sufficient contact to operate. In contrast, remote monitoring is contactless, unobtrusive, and could monitor several vital signs simultaneously while providing more user-friendly monitoring in various environments [4–7]. Remote monitoring with radar technology could reform sleep monitoring at home and nursing homes by removing the often disturbing tactile sensation of a wearable device and the wired sensors that tend to detach. It can also be a cost-effective solution as it does not require disposable elements such as electrodes. Ultimately, remote measurements could ease monitoring in critical care taking place in hospitals [8,9].

Periodic variations in the measured radar signal, which are caused by micromotions on the body surface, can convey information regarding the two vital signs considered herein: heart rate and

respiration rate. In this paper, we study a frequency modulated continuous wave (FMCW) radar developed at VTT Technical Research Centre of Finland [6,10]. Similar millimetre wave chipsets and development boards capable of time domain multiplexing are also available commercially [11,12]. The radar transmits frequency-modulated electromagnetic waves and can detect the phase of the received signal with about one degree accuracy. Its high resolution enables the detection of microscopic vascular pulsations on the skin.

Prior studies on FMCW radars have already established the potential for vital sign monitoring applications [4,13–15]. Among the heart rates extracted by Anitori et al. 60% were within 10% of their reference values [13]. Alizadeh et al. later achieved 94% and 80% accuracies for respiration rate and heart rate, respectively [15]. However, while both studies examined monitoring in lying positions, Alizadeh et al. included only one subject as opposed to the six participants by Anitori et al. Additionally, both studies used a commercial, non-medical device for reference signals. In contrast, Wang et al. and Adib et al. studied ten or more subjects in seated positions, using medical devices for reference [4,14]. Wang et al. reported approximately 5–31% and 11–20% relative errors for respiration and heart rate, respectively, depending on the exact position [14]. Adib et al. demonstrated median accuracies of 99.4% and 99% for respiration and heart rate, respectively, and were also able to measure multiple targets simultaneously [4].

In this study, we explore the potential of FMCW radar technology for the special application of nocturnal vital sign monitoring by emulating diverse real-life scenarios. Unlike previous studies, we pursue a wide range of both heart and respiration rates to discuss the applicability of FMCW radars to monitor people with different conditions. Capturing a wide range of vital signs is essential for sleep analysis and for monitoring sleep disorders, such as hypopnoea, and for following the effects of possible interventions [16–18]. Whereas previous works have established suitable accuracies for commercial use at home and office environments, we demonstrate the applicability of an FMCW radar in the aforementioned clinical applications by showing that it can accurately capture even alarmingly anomalous vital signs, such as shallow respiration.

We include ten volunteers in our study in order to account for the natural differences between individuals and to ensure a level of robustness in our vital sign extraction methods [19]. The participants are monitored in varying lying positions while performing simple activities, emulating vital sign variations of real-life sleeping scenarios. We extract their interbeat intervals (IBI) and respiratory rates, and compare to reference data acquired using Embla Titanium, a certified medical device.

Despite the deliberately challenging study setting, we are able to surpass the results of previous studies in heart rate monitoring accuracy. Moreover, we establish high accuracy in respiration monitoring even with minimal respiratory motion, which we expect to promote radar-based monitoring in clinical settings. We provide further grounds for such clinical applications by demonstrating, for the first time to our knowledge, accurate radar-based extraction of features commonly used in heart rate variability (HRV) analysis, an essential tool in modern stress monitoring applications [20–23].

2. Materials and Methods

The workflow of our study is schematically illustrated in Figure 1. In this section, we elaborate on each item step-by-step. We used an in-house developed FMCW radar (Section 2.1) and reference devices (Section 2.2) to measure each participant in the study group (Section 2.3) during a set of activities resembling real-life sleeping scenarios (Section 2.4). The measured data were analysed using robust state-of-the-art approaches to extract interbeat interval, heart rate, and heart rate variability parameters, as well as respiration rate (Sections 2.5–2.7). Finally, the accuracy assessment of the retrieved estimates was performed using the methods described in Section 2.8.

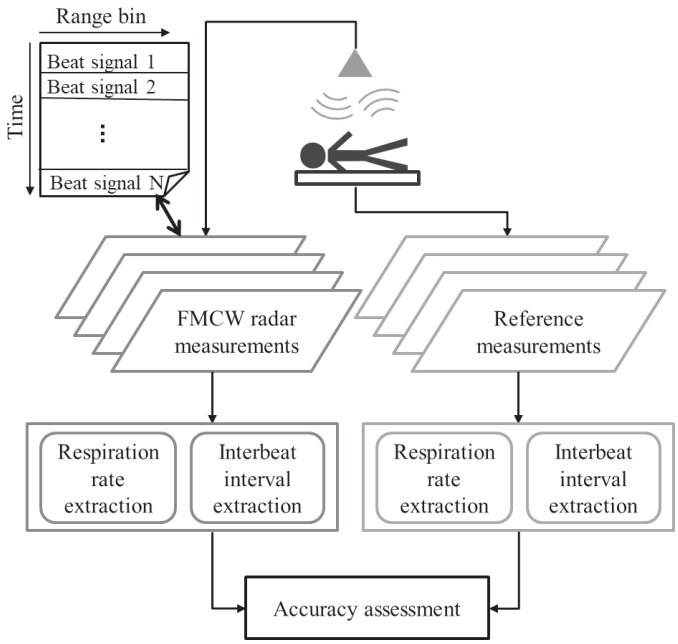

Figure 1. The overall measurement and evaluation workflow. The radar and reference data are processed separately. A beat signal describes the difference between the transmitted and received signals.

2.1. Frequency Modulated Continuous Wave Radar

An FMCW radar transmits a frequency-modulated continuous signal and detects its reflection. As visualized in Figure 2, the distance to an object can be computed based on the beat frequency, i.e., the frequency difference between the transmitted and received signals [24].

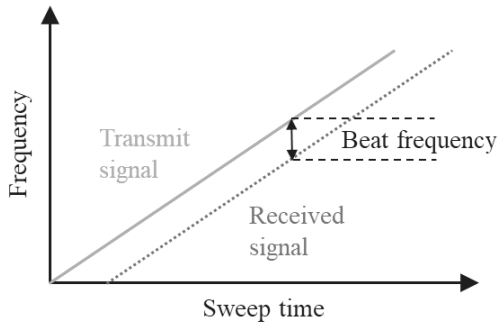

Figure 2. Illustration of the frequency modulated continuous wave (FMCW) radar principle [6]. Using a frequency sweep allows to compute the distance from the radar to the target.

An FMCW radar measurement can be divided into chirps, or frequency sweeps, where the transmit signal frequency modulates (or sweeps) through the specified frequency band [10]. The instantaneous profile of the observed distances, also known as the complex range profile, can be extracted from the collected samples of beat signals by applying a Fast Fourier Transform (FFT) to each

set of samples from the same chirp. Using the resulting set of complex FFTs, the beat signal amplitude and phase can be extracted for the desired range bin. Furthermore, the complex range profiles from consecutive chirps can be stacked into the range slow-time matrix, which contains the phase information of the beat signal as a function of time, the main signal needed for vital sign extraction.

The in-house FMCW radar used in this study operated at the carrier frequency of 24 GHz with a 250 MHz bandwidth. The radar has a range resolution of 60 cm, micromotion detection accuracy below 1 µm, and receiver noise figure of 12 dB. In this study, we explore two chirp repetition frequencies, i.e., sampling frequencies: 110 Hz and 154 Hz. While the maximum operable frequency of 154 Hz can capture more detailed information, the lower one is more stable to operate with the existing software. Thus, we take the opportunity to examine whether lowering the sampling frequency deteriorates performance in vital sign monitoring.

The radar was mounted on the ceiling above a bed, facing downwards towards the subject above the torso, at a fixed distance of about 2 m, as portrayed in Figure 3. The radar antenna 3 dB beam width was 65° along the length of the bed and 26° along the perpendicular direction. The field of coverage was configured to 3 m to reduce noise.

Figure 3. The measurement setting. The FMCW radar (highlighted in brighter tones and pointed out with a red arrow) is mounted on the ceiling above the bed. The treadmill beside the bed was used for exercising during the measurement session.

2.2. Reference Devices

Reference signals were collected simultaneously with the radar data, using the Embla titanium portable polysomnography (PSG) system, a CE certified class II device in use in many physiological studies worldwide [25]. Two electrocardiographic (ECG) electrodes were attached to the subject to collect the reference ECG signal at 256 Hz sampling frequency. One electrode was attached under the

right-side collarbone and the other on the lower left part of the thoracic cage. One respiratory inductive plethysmography (RIP) belt on the thorax was used to collect the reference respiration signals at 32 Hz sampling frequency.

Because of technical issues that sometimes occur in the measurements and downgrade the signal-to-noise ratio, an additional reference was collected at 110 Hz using VTT's ballistocardiography (BCG) based sensor sheet installed beneath the mattress topper. The sensor sheet can detect respiration rate with 1.5% error relative to RIP belts [26].

2.3. Study Group

We measured eleven participants from age 25 to 55 (37 on average, 2 female), who signed an informed consent form prior to the measurement, after receiving information about the measurement protocol and the study objectives. The study did not intervene with the physical integrity of the volunteers and the study setting was not harmful or otherwise disturbing.

2.4. Measurement Protocol

As presented in Table 1, the measurement protocol was a combination of three distinct activities, each measured for two minutes at a time: relaxed respiration, hypopnoea simulation, and recovering after physical exercise. The hypopnoea simulation comprised one minute of shallow respiration and another minute of normal respiration. These sub-activities are presented separately in Table 1 for clarity. Before the final activity of recovering after exercise, the participants walked on a treadmill with roughly 15% inclination at 2 km/h, for two minutes. The participants were not measured during the exercise, and they were allowed to interrupt at any time. Nevertheless, all participants exercised the full two minutes.

Table 1. Vital sign measurement protocol.

Activity	Position	Duration (min)	
		110 Hz	154 Hz
Relaxed respiration	Supine	2	2
Relaxed respiration	Right lateral	2	2
Relaxed respiration	Prone	2	2
Relaxed respiration	Left lateral	2	2
Relaxed respiration	Supine	2	2
Hypopnoea simulation, shallow respiration	Supine	1	-
Hypopnoea simulation, normal respiration	Supine	1	-
Hypopnoea simulation, shallow respiration	Right lateral	1	-
Hypopnoea simulation, normal respiration	Right lateral	1	-
Hypopnoea simulation, shallow respiration	Prone	1	-
Hypopnoea simulation, normal respiration	Prone	1	-
Hypopnoea simulation, shallow respiration	Left lateral	1	-
Hypopnoea simulation, normal respiration	Left lateral	1	-
Recovering after exercise [a]	Supine	2	2
Total measurement time (min)		20	12

[a] Preceded by a two-minute exercise (not measured).

The relaxed respiration and hypopnoea simulation activities were measured once in all four different positions: supine, left lateral recumbent, right lateral recumbent, and prone. This was true with one exception: the relaxed respiration activity was repeated in the supine position to ease the participant's transition to the next activity. The final activity (after the exercise) was only measured in the supine position. The participants were given sufficient transition time for each change of position.

The protocol was repeated with two sampling frequencies using a reduced protocol with the more unstable 154 Hz sampling frequency (see Table 1). Measurement segments using the different sampling frequencies were performed sequentially but in alternating order between participants.

In total, 32 min of activity data were collected per participant (20 min with 110 Hz and 12 min with 154 Hz). This comprises 14 min in the supine position and 6 min in each of the three other positions.

2.5. Heart Rate Extraction

The interbeat interval was extracted from the reference and radar devices using different methods (see Figure 1). The R-to-R interval, used as reference IBI, was extracted from the ECG signal using the `findpeaks` function by MATLAB® (minimum peak distance 0.3 s) after trend removal. To extract IBI from the radar signal, cepstral analysis, a variant of spectral analysis, was applied [27]. It is able to emphasize the significantly small heartbeat-induced motions on the body surface by using a logarithmic transformation. However, the performance of the FFT-based method is deteriorated by both spectral variance and the natural variations in the pulse shape and IBI. To minimize spectral variance, we average over multiple contemporaneous range signals. To overcome the non-stationary nature of the heartbeat, we use a set of different FFT window lengths in parallel to compose a summary cepstrum. The proposed method was first developed for IBI extraction from multichannel BCG (covered by US patent 2010/0249628) [28]. In the current study, we adapt the method for the radar application.

Figure 4 illustrates the radar IBI extraction process. We selected $N = 24$ radar range bins to provide the input signals. Six different length Hamming windows W_i ($i = 1, \ldots, K$, $K = 6$) were used in parallel, each on every one of the N signals, to apply FFT and obtain the signal spectra. Each window was applied as a sliding window with strong overlap (0.1 s interval). The K window lengths ranged from 3.5 s to 20 s. The shortest windows were used to capture IBI approximately equal to half the window length, whereas the longest windows containing more than the optimal two heartbeats were used to detect IBI of rather constant pulse shape and interval. The FFT length was set to 40 s of samples for all windows and zero padding was used to improve resolution for the upcoming peak selection. To boost computational speed, the FFT length was rounded up to be divisible by sixteen.

Subsequently, the inverse FFT (IFFT) over the natural logarithm of the averaged spectra were computed to obtain the K cepstra. The cepstrum C_i is defined as the real part of the inverse Fourier transform $\mathcal{F}-1\{\cdot\}$ taken over the natural logarithm of an amplitude spectrum $|S_i|$

$$C_x = real(\mathcal{F}^{-1}\{log(|S_x|)\}) \tag{1}$$

as described in [27]. Whereas the spectrum S_i contains peaks at the harmonic frequencies of the fundamental heartbeat frequency, in the cepstrum the harmonic spectral peaks appear as a single peak at the corresponding lag time, or quefrency [27].

The K cepstra are averaged to form the summary cepstrum. The overlaps between neighbouring cepstra were taken into consideration by applying a weighting window designed to produce equal sensitivity on each quefrency upon averaging.

Finally, the peaks in the summary cepstrum were taken as the IBI estimates. The quefrency resolution of the summary cepstrum and thus the IBI is directly the inverse of sampling frequency, and the slow-time resolution equals the sliding window interval. The peak selection was performed over the quefrency range from 0.5 s to 1.5 s in the cepstogram. It was initialized by taking the peaks that were strong with respect to both quefrency and time. Next, while weighting each initial IBI estimate by the corresponding peak height, a time averaged IBI S_{IBI} was computed using a 60 s sliding Hamming window. Lastly, uncertain IBI were removed based on the cepstral peak height and distance to S_{IBI}; only the most prominent peaks were selected. The entire peak selection routine was repeated iteratively up to four times to allow S_{IBI} to stabilize into the most prominent signal shape.

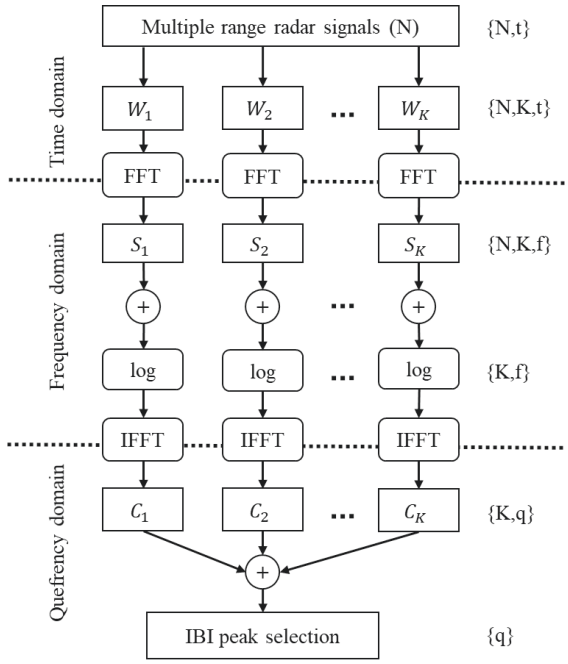

Figure 4. Interbeat interval extraction from a set of N range signals using K Fast Fourier Transform (FFT) windows. The process is repeated using overlapping FFT windows. The notations on the right represent the data dimensions at each phase; t denotes time, f frequency, and q the cepstrum lag time, or quefrency.

We note that our method can yield more than one estimates per actual interbeat interval. Thus, the average heart rate in the unit of beats per minute (bpm) was calculated by dividing 60 s by the average of the corresponding IBI estimates.

2.6. Heart Rate Variablity Analysis

HRV analysis employs a collection of features describing the beat-to-beat signal. The 13 time-domain features and 7 frequency domain features selected for this study are described in Table 2. In the context of frequency domain features, we chose to use the square root of power to rather present information scaled by the IBI signal amplitude than power itself. Welch's method (30 s windows with trend removal, 75% window overlap) was used to estimate the power spectral density after resampling IBI to a constant 10 Hz sample frequency using cubic interpolation.

HRV features are commonly extracted from normal-to-normal peak intervals (NNI). Therefore, abnormal IBI were removed to obtain NNI estimates and replaced by linearly interpolated values [29]. A reference interval was considered abnormal if it changed more than 15% with respect to the previous one, whereas IBI from the radar were allowed a 20% change to account for the irregularity of the extracted IBI values.

Table 2. Heart rate variability features explained.

Domain	Feature	Description
Time	Mean NNI	Average over all normal-to-normal peak intervals (NNI)
	Median NNI	Median over all NNI
	RMSSD	Root mean square of consecutive differences of adjacent NNI
	SDSD	Standard deviation of consecutive differences of adjacent NNI
	SDNN	Standard deviation of NNI
	CVNNI	Coefficient of variation
	CVSD	Coefficient of variation for successive differences
	pNN20	Percentage of interval differences exceeding 20 ms
	pNN50	Percentage of interval differences exceeding 50 ms
	HR	Heart rate
	STD_{HR}	Standard deviation of heart rate
	Min HR	Minimum heart rate
	Max HR	Maximum heart rate
Frequency	$\sqrt{}$Total power	Square root of total power
	$\sqrt{}$VLF	Square root of very low frequency power (0.0033–0.04 Hz)
	$\sqrt{}$LF	Square root of low frequency power (0.4–0.15 Hz)
	$\sqrt{}$HF	Square root of high frequency power (0.15–0.4 Hz)
	$\sqrt{}$LF/HF ratio	Square root of the ratio of low and high frequency power
	LFnu	Normalized low frequency power
	HFnu	Normalized high frequency power

2.7. Respiration Rate Extraction

As indicated in Figure 1, distinct methods were used to extract the respiratory vital sign from the reference devices and the radar. The reference respiration rate was derived from the reference signal through detrending and peak detection. The subtracted, smoothed trend was estimated using a 15 s Hann window, and the `findpeaks` function by MATLAB® was used for peak detection (peak distances ranging from 1.4 s to 20 s were allowed). Local respiratory cycles were extracted from subsequent maxima and minima separately. Artefacts were identified based on the length and amplitude of the respiration cycle. Consecutive distorted cycles were combined when possible to better match the preceding and following five respiration cycles.

For the radar data, the respiratory motion was captured from the change of phase between consecutive chirps in the complex range profile, hereafter referred to as the phase signal. Specifically, the autocorrelation function (ACF) is applied on the phase signal of a selected range bin. The ACF has been previously proved to work in respiration monitoring on a single subject [15]. In this study, we aim for a robust implementation accurate for several subjects.

The respiration rate extraction process is presented schematically in Figure 5. The optimal range bin corresponds to the distance where the primary target is located. The participants in this study were mostly stationary, which made it possible to select an optimal range-bin for each sub-measurement (row in Table 1). The range bin with the global maximum over the profiles in the range slow-time matrix was selected as the optimal range bin. Subsequently, the DC component was estimated globally from the full slow-time signal and removed. The slow-time profile in the optimal range bin was further used to extract an instantaneous phase signal, which was then unwrapped to remove $\pm 2\pi$ phase jumps.

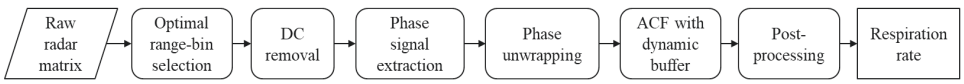

Figure 5. Respiration signal extraction.

Because the phase signal closely follows periodic variation of the respiratory motion, we used the autocorrelation function to extract and quantify it. Unlike a periodogram, it can work with both long and short signals. The ACF at lag $k < n$ can be written as

$$R_k = \left[\sum_{i=1}^{n-k} (s_i - \mu)(s_{i+k} - \mu)\right] / \left[\sum_{i=1}^{n} (s_i - \mu)^2\right], \quad (2)$$

where the input sequences $[s_1, s_2, ..., s_n]$ are generated by a sliding window function, and μ denotes their mean [30]. Given a suitable input sequence size, the lag of the maximum peak directly provides an estimate of the breathing interval. Thus, approximating one respiratory cycle as the average of 100 consecutive estimates, the window size was dynamically adjusted to contain 2.2 respiratory cycles.

Finally, post-processing focused on the removal of non-reliable estimates, such as outliers (over three standard deviations apart from the mean) and estimates with unstable phase due to other movements.

2.8. Performance Evaluation

Vital sign values (IBI or respiration interval) extracted from the radar data were each compared against the reference value closest in time. In all cases, a reference value resided within a maximum of 1.5 s temporal distance from the value.

Mean absolute error (MAE), root mean square error (RMSE), and Pearson correlation coefficient were selected for performance evaluation. While MAE is easy to interpret, RMSE emphasizes large errors, conveying information on where the most blatant errors occur. MAE and RMSE were computed individually for each participant and sub-measurement (row in Table 1). The resulting errors were weighted by the sub-measurement duration to aggregate representative error metrics for participants, activities, and lying positions. The aggregation methods are further described in Supplementary Material. For visual analysis, Bland-Altman plots were chosen to depict the agreement between the suggested methods and the reference. In contrast to correlation, the Bland-Altman plot describes both random and systematic error [31,32].

3. Results

The two sampling frequencies produced equally accurate results: the difference in RMSE was 0.001 s when measuring IBI and 0.169 1/min when measuring respiration rate. Thus, the measurements of either frequency are included in the remaining analysis.

The vital sign extraction results encompass ten participants. One of the 11 subjects (ID004) was excluded from the analysis due to unsuccessful data collection. Also, for three other participants (ID002, ID005, and ID010), the PSG respiration reference showed poor signal quality and was replaced with the secondary BCG-based reference. One participant (ID007) was excluded from the respiratory rate analysis due to poor quality reference in several sub-measurements.

3.1. Heart Rate

The average measured interbeat interval over all measurements was 1.041 s (standard deviation SD 0.160 s, 5th percentile 0.820 s, 95th percentile 1.313 s), corresponding roughly 57.7 bpm, using the reference device. Respectively, the radar measured average was 1.053 s (SD 0.152 s, 5th percentile 0.832 s, 95th percentile 1.310 s), or roughly 57.0 bpm. Largest variations in the IBI were recorded when recovering from physical exercise; SD of 0.178 s was observed (0.164 s using the radar). Figure 6 illustrates samples of the extracted IBI signals with respect to the reference IBI for two participants. More samples are provided in the Supplementary Figure S1.

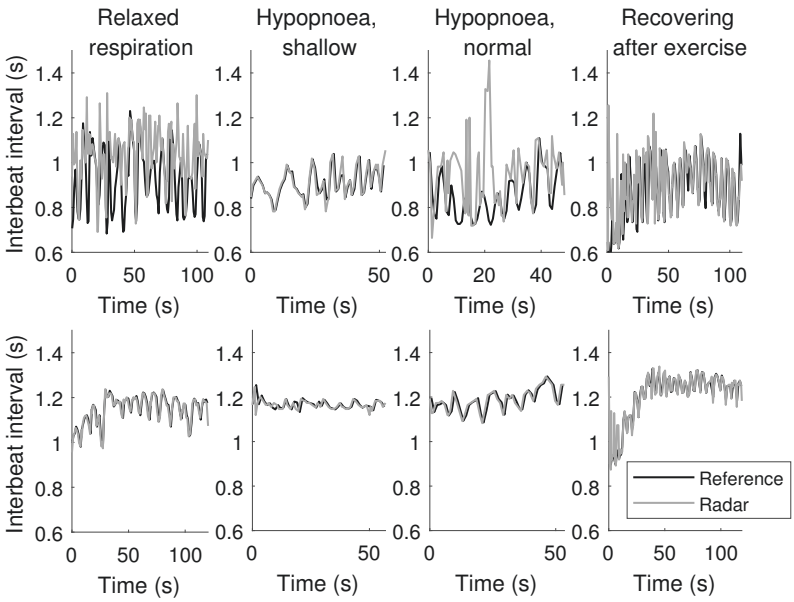

Figure 6. Interbeat interval samples of two participants (ID003 on the top, ID011 on the bottom) for each activity in the supine lying position.

Figure 7 illustrates the resulting differences between the radar and reference IBI in a Bland-Altman plot. The mean difference between the radar-derived and reference IBI is 0.013 s (SD 0.083 s), which roughly corresponds to a 0.71 bpm difference in the instantaneous heart rate.

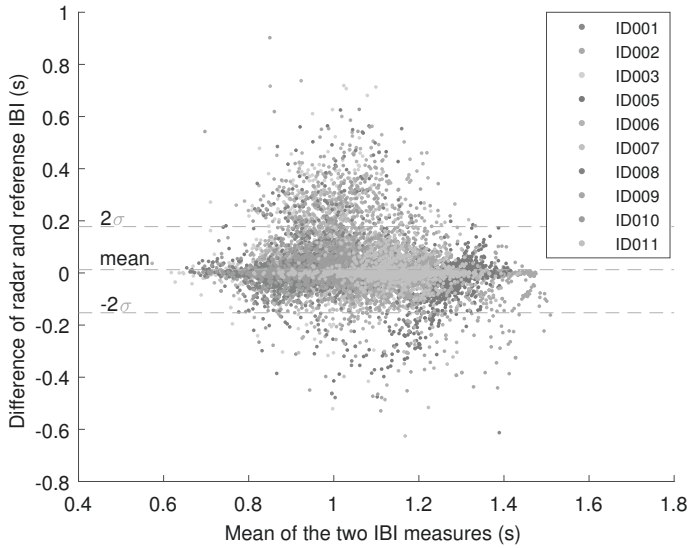

Figure 7. The Bland-Altman plot for interbeat interval. The dashed lines indicate the mean and the interval containing 95% of the samples.

Table 3 presents mean absolute error for each participant and activity. Differences between participants were below 0.07 s. To complement these results, we compensated for the varying number of heartbeat events per participant by taking an average not weighted by measurement duration. Also in this case, the participant MAE demonstrate differences under 0.07 s (and an overall average MAE of 0.038 s).

Considering MAE for each activity in Table 3, the error was at its largest when the participant was recovering after a short exercise session. Table 4 further describes the increase in MAE during the recovering activity as compared to the other activities measured in the same position. Additionally, Table 4 shows that differences between positions are in the order of milliseconds.

Table 3. Mean absolute error (MAE, s) for interbeat intervals with respect to activity and participant.

Participant ID	Relaxed	Hypopnoea, Shallow	Hypopnoea, Normal	Recovering	Participant MAE
ID001	0.016	0.013	0.011	0.018	0.015
ID002	0.040	0.015	0.028	0.050	0.034
ID003	0.068	0.032	0.067	0.058	0.061
ID005	0.029	0.032	0.023	0.095	0.038
ID006	0.073	0.074	0.061	0.116	**0.077**
ID007	0.023	0.010	0.057	0.027	0.027
ID008	0.041	0.019	0.029	0.032	0.036
ID009	0.046	0.009	0.090	0.061	0.051
ID010	0.023	0.032	0.020	0.019	0.023
ID011	0.018	0.014	0.017	0.020	0.018
Activity MAE	0.037	0.026	0.042	**0.052**	0.038 [a]

The largest activity and participant MAEs are bolded. [a] The total MAE over all activities and participants

Table 4. Mean absolute error (s) for interbeat intervals with respect to activity and lying position.

Position	Relaxed	Hypopnoea, Shallow	Hypopnoea, Normal	Recovering [a]	Position MAE
Supine	0.034	0.014	0.042	0.052	0.038 [b]
Right lateral	0.040	0.036	0.041	-	0.040
Prone	0.039	0.028	0.054	-	**0.040**
Left lateral	0.038	0.028	0.032	-	0.035
Activity MAE	0.037	0.026	0.042	**0.052**	0.038 [c]

The largest mean MAEs are bolded. [a] Recovering was only measured in the supine position. [b] 0.033 s if the recovering activity is not considered. [c] The total MAE over all activities and positions.

Considering all participants, activities, and positions, the IBI extracted from the radar exhibited an overall MAE of 0.038 s (SD 0.074 s, median absolute error 0.008 s) and RMSE of 0.084 s. The RMSE results presented in the Supplementary Tables S1 and S2 display similar trends as MAE. The Supplementary Figure S2 exemplifies the difference of the two metrics. Furthermore, the extracted IBI demonstrated a statistically significant Pearson correlation of 0.862 (p-value less than 0.01) as compared to the reference IBI. The IBI extracted from the radar are illustrated with respect to the reference values in the Supplementary Figure S4.

As for averaged heart rate in the units of beats per minute, MAE varied from 0.816 to 1.384 bpm for the different activities. These results are in line with the IBI results. Ranking the participant-wise errors, the order of some participants were reversed (e.g., ID003 and ID006), showing a small effect from the timestamp misalignment between the radar-derived and reference IBI. The mean absolute temporal distance between the two estimates was 0.24 s (SD 0.13 s), while maximum temporal misalignment was 0.74 s.

Overall, the heart rate analysis gave a MAE of 1.031 bpm, which corresponds to an average MAE of 0.016 s for the IBI. After removing ectopic beats, the results remained similar with an overall MAE of 1.079 bpm.

3.2. Heart Rate Variability

The HRV features were computed and evaluated for each sub-measurement individually. The results are summarized over all measurements in Table 5.

Table 5. Comparison of the heart rate variability features.

Feature	Mean ± Standard Deviation		MAE	Correlation
	Radar	Reference		
Time-domain features				
Mean NNI	1.06 ± 0.13	1.05 ± 0.14	0.02	0.98
Median NNI	1.06 ± 0.14	1.05 ± 0.15	0.02	0.98
RMSSD	0.05 ± 0.02	0.04 ± 0.02	0.01	0.81
SDNN	0.07 ± 0.03	0.07 ± 0.04	0.01	0.88
SDSD	0.05 ± 0.02	0.05 ± 0.02	0.01	0.81
CVNNI	0.07 ± 0.03	0.07 ± 0.04	0.01	0.89
CVSD	0.05 ± 0.02	0.04 ± 0.02	0.01	0.84
pNNI20	50.59 ± 17.75	50.92 ± 25.88	11.12	0.81
pNNI50	28.10 ± 15.03	25.31 ± 21.08	9.19	0.84
Mean HR	57.68 ± 7.06	58.52 ± 7.66	1.15	0.97
STD_{HR}	4.03 ± 1.93	4.09 ± 2.44	0.77	0.86
Min HR	48.87 ± 5.95	51.03 ± 6.67	2.61	0.87
Max HR	70.24 ± 9.79	70.87 ± 12.02	3.93	0.79
Frequency-domain features				
$\sqrt{}$Total power	0.09 ± 0.04	0.09 ± 0.04	0.01	0.95
$\sqrt{}$VLF	0.04 ± 0.02	0.05 ± 0.02	4.9×10^{-3}	0.94
$\sqrt{}$LF	0.07 ± 0.03	0.07 ± 0.03	0.01	0.93
$\sqrt{}$HF	0.04 ± 0.01	0.04 ± 0.02	0.01	0.86
$\sqrt{}$LF/HF ratio	1.58 ± 0.37	1.66 ± 0.41	0.20	0.72
LFnu	68.94 ± 13.27	70.70 ± 13.28	5.44	0.78
HFnu	31.06 ± 13.27	29.30 ± 13.28	5.44	0.78

Most of the time-domain features exhibited notable correlation between the radar-derived and reference features. However, the MAE indicated notable 9–11% mean absolute errors in the pNNI20 and pNNI50. The remaining features agreed well with the reference features, exhibiting high correlation and small errors. For deviation-based features, MAE were lower than the standard deviation of the mean reference and for other features MAE was at most 5.5% (max HR) of the mean reference value.

Mean NNI, median NNI, and mean heart rate exhibited similar trends in terms of MAE between different participants as already observed in Table 3. Additionally, no large differences between neither the lying positions nor the activities were observed, although the recovering activity showed the largest error consistently.

The frequency-domain features showed mostly high correlations and low errors as well. The $\sqrt{}$LF/HF ratio, LFnu, and HFnu features exhibited the most moderate correlations and the largest errors among the selected features. Yet, a 5% error in the normalized low or high frequency power may be considered acceptable. The errors for each activity, participant, or position did not seem to differ much for any of the frequency-domain features.

3.3. Respiration Rate

The measured respiration rates were on average 15.634 1/min (SD 7.492 1/min, 5th percentile 6.548 1/min, 95th percentile 32.535 1/min) for the reference devices. Correspondingly, the radar measured average was 15.855 1/min (SD 7.223 1/min, 5th percentile 7.138 1/min, 95th percentile 32.000 1/min). Most variation was recorded during the shallow respiration part of the hypopnoea simulation, showing SD of 10.010 1/min (9.097 1/min for the radar measured values). Samples of the

extracted respiration rates are illustrated in Figure 8 with respect to the reference signals. More samples are depicted in the Supplementary Figure S3.

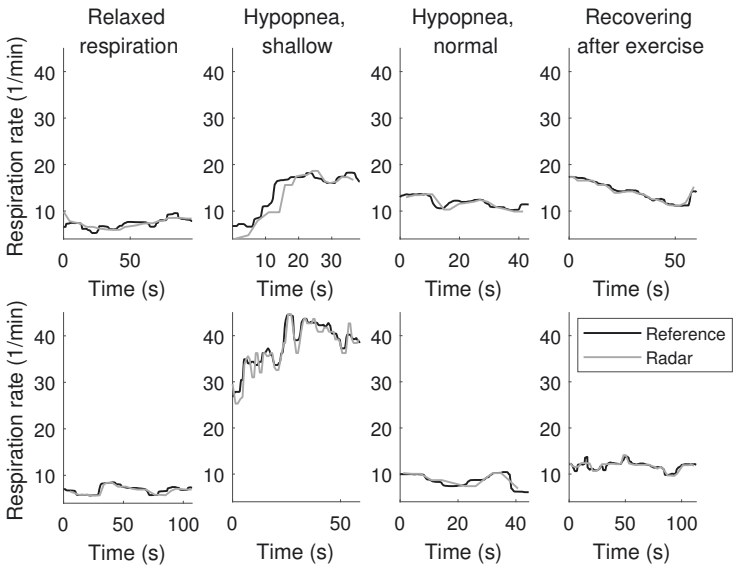

Figure 8. Respiration signal samples of two participants (ID003 on the top, ID011 on the bottom) for each activity in the supine lying position.

Figure 9 presents the Bland-Altman plot comparing respiration rates from the radar to those given by the reference devices. It exhibits a mean error of 0.221 breaths per minute (SD 3.137 1/min).

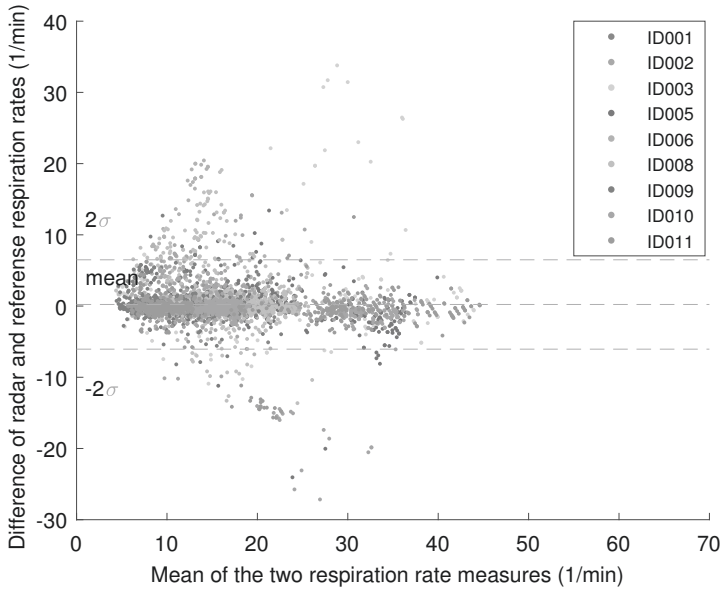

Figure 9. The Bland-Altman plot for respiration rate. The dashed lines indicate the mean and the interval containing 95% of the samples.

As presented in Table 6, the largest observed difference between participants in terms of MAE was 1.820 1/min. Additionally, when compensating for the different number of respiratory events per participant by averaging without weighting by measurement duration, the largest difference in participant MAE reduced to 1.487 1/min (with an overall average MAE of 1.354 1/min). For different activities, the smallest respiratory motions (during hypopnoea, shallow) exhibited the largest errors (especially ID003 and ID006).

Table 6. Mean absolute error for respiration rates with respect to activity and participant.

Participant ID	Relaxed	Hypopnoea, Shallow	Hypopnoea, Normal	Recovering	Participant MAE
ID001	0.367	0.725	0.418	0.251	0.487
ID002	0.940	1.694	1.124	0.611	1.088
ID003	1.305	5.118	0.505	0.339	2.272
ID005	0.751	2.419	0.934	0.617	1.075
ID006	0.856	3.852	0.369	0.536	1.331
ID008	2.626	2.310	1.601	1.170	**2.308**
ID009	1.049	2.610	0.869	1.833	1.336
ID010	0.737	2.516	0.752	3.094	1.392
ID011	1.311	1.155	0.981	0.388	1.114
Activity MAE	1.222	**2.408**	0.887	1.149	1.414 [a]

The largest activity and participant MAEs are bolded. [a] The total MAE over all activities and participants.

In different positions, the differences in MAE are below 1.000 1/min, as presented in Table 7. However, the lateral measurement positions, especially the left lateral position, exhibited higher MAE as compared to the other two positions.

Table 7. Mean absolute error for respiration rates with respect to activity and lying position.

Position	Relaxed	Hypopnoea, Shallow	Hypopnoea, Normal	Recovering [a]	Position MAE
Supine	1.002	1.270	1.149	1.149	1.086 [b]
Right lateral	1.096	3.588	0.682	-	1.656
Prone	1.256	1.458	0.707	-	1.225
Left lateral	1.732	3.454	1.001	-	**2.064**
Activity MAE	1.222	**2.408**	0.887	1.149	1.414 [c]

The largest mean MAEs are bolded. [a] Recovering was only measured in the supine position. [b] 1.061 1/min if the recovering activity is not considered. [c] The total MAE over all activities and positions.

We obtained an overall MAE of 1.414 1/min (SD 2.810 1/min, median absolute error 0.515 1/min) and RMSE of 3.145 1/min for respiration rate. Detailed RMSE results are presented in the Supplementary Tables S3 and S4. Furthermore, the measurements exhibited a significantly high Pearson correlation of 0.910 (p-value less than 0.01), as visualized in the Supplementary Figure S5.

4. Discussion

The radar derived IBI tend to be slightly larger than the reference, although this systematic error varies considerably between participants (see Figure 7). Differences between participants in the normal ECG waveform, heart rate, and heart rate variability are all expected due to varying physiological factors. Thus, it comes as no surprise that the IBI of some participants may be more difficult to extract than that of others. The error in IBI for each participant is reasonable, in the order of tens of milliseconds.

As for different activities, the error in the extracted IBI was at its largest when a participant was recovering after an exercise session. This is when the largest respiratory motions are expected. Consistently, during the hypopnoea-mimicking shallow respiration, with close to no respiratory movement, the error was at its smallest. After the physical exercise, the participants were somewhat out of breath and it came naturally to take quick, deep breaths. Thus, the respiratory rate came closer

to the expected range of the heart rate. As both motions are periodic, it became more difficult to distinguish the two. However, MAE for the recovering activity is only 0.052 s, thus showing good performance despite the challenging circumstances.

The overall error of 0.038 s indicates strong performance for the presented IBI extraction method in various scenarios. The different lying position did not affect the accuracy of the extracted IBI (see Table 4).

The presented results exceed previous achievements in heart rate monitoring obtained for lying positions. Anitori et al. presented an FFT method achieving a 10% error for heart rate, whereas we obtained 3.6% error for instantaneous heart rate [13]. Alizadeh et al. obtained a correlation of 80% for a single person, whereas we demonstrate an 86% correlation for ten participants [15]. Our results also compete with the results by Adib et al. who achieved a 99% median accuracy over a variety of measurement distances for participants in sitting positions [4]. At a similar distance of 2 m, they obtained a median accuracy of 98.7%, whereas our overall median absolute error of 0.008 s corresponds to a 99.2% accuracy and the mean absolute error of 0.038 s corresponds to 96.3% accuracy.

In HRV analysis, most time and frequency-domain features obtained from the IBI estimates demonstrated high agreement with the reference values. The pNNI20 and pNNI50 time-domain features were the notable exceptions. The error for these parameters presumably followed from the numerous estimates for each actual heartbeat event given by our IBI extraction method, which shifts the number of intervals exceeding 20 or 50 ms as compared to the total number of intervals. The extra estimates might have also affected the error in minimum and maximum heart rates. Yet, many HRV parameters remain useful when extracted from the radar.

As for respiratory rate, smallest respiratory motions were the most difficult to detect using the radar (see Table 6). Yet, the average MAE remained comparable to that of normal respiration (1.222 1/min), supporting that the method is reliable in various real-life scenarios.

Respiration rate extraction was found to be more complicated in lateral positions; the highest error was measured in one of the two positions for all expect one participant. Notably, two participants (ID003 and ID006) exhibited exceptionally high error (especially RMSE) in one of the lateral positions during the shallow respiration period of the hypopnoea simulation, contributing notably to the overall error.

The observed differences in the error of the two lateral positions may be tracked back to the measurement setting. The RIP belt used to measure the primary respiration reference is an elastic band around the participants thorax; the change of posture could loosen the RIP belt by sliding it from its original location. As described in the protocol (Table 1), the participants always visited the right lateral position before the left, which may have resulted in larger error in the left lateral position. Additionally, the error may be higher in lateral positions as compared to the other two because of the smaller prevalence of the respiratory motion in the observed area.

Altogether, our respiration extraction method is comparable to the state-of-the-art methods. As compared to Alizadeh et al. who used a similar autocorrelation approach to extract respiration with 94% correlation with the reference, we expand the method for several participants and achieve a 91% correlation. Adib et al. on the other hand achieved a 99.4% median accuracy in seated positions, while we show an overall 96.5% median absolute accuracy and a 91.0% mean absolute accuracy in lying positions [4]. However, in contrast to previous studies, our study considered a wide range of respiration rates including breathing with minimal motion [4,13–15]. Despite the challenging setup, our method performed robustly in the different scenarios. The high correlation between the radar-extracted and reference estimates is a remarkable result given the wide range of recorded respiratory motions.

The chosen methods of cepstrum for IBI extraction and autocorrelation for respiration rate extraction are closely related, as both can be formulated as an inverse Fourier transform from the power spectra [33]. Cepstral analysis emphasizes the harmonic frequencies of a spectrum. The signal power of rapid bursts, such as heartbeats, is mainly carried by the harmonic spectral peaks, which are further emphasized in the logarithm of the spectrum. In contrast to heartbeats, the signal power of

respiratory motion is concentrated on the base frequency, making autocorrelation a suitable approach to extract respiration [33].

The presented vital sign extraction methods are limited with respect to real time applications. The IBI extraction requires a delay equal to the longest FFT window (20 s) in addition to the delay due to the iterative smoothing to remove uncertain estimates (upto 4 min). The IBI extraction performance could however be improved if it was implemented in parallel with another method, such as the data fusion method described in [34]. The respiration rate extraction is restricted by the maximum delay equal to the maximum ACF peak extraction buffer (default of 15 s). Furthermore, the optimal range bin selection and DC removal were computed globally for each sub-measurement, and would need to be performed adaptively to account for changes of position during sleep.

The presented results are limited by the restricted set of participants and thus the methods may not generalize as well for broader groups. Although data collection in a controlled environment allowed us to capture a wide range of vital signs, the natural next step would be to test the presented methods on a large study group in over-night measurements. The suggested techniques could also be optimized for personal vital sign patterns to improve performance for individuals. For future work, we note that while here the clean reference ECG enabled the use of the standard MATLAB® tool `findpeaks`, the Pan-Tompkins algorithm is suggested for reference R peak extraction. In addition, the results were obtained on subjects lying still and therefore do not directly apply to moving subjects. However, applying noise removal prior to the vital sign extraction methods could increase performance for moving subjects.

5. Conclusions

Our results suggest that the cost-effective 24 GHz FMCW radar together with the proposed vital sign extraction methods represent a solution that can deliver accurate results for nocturnal vital sign monitoring even during various conditions, such as sleep apnoea. We obtained state-of-the-art level accuracies for heart rate monitoring while, to the best of our knowledge, being the first to report as low errors in recording instantaneous interbeat intervals using a similar device [4]. Moreover, we demonstrated the radar's feasibility in heart rate variability analysis. Finally, we presented remarkably accurate results in respiration monitoring, maintaining a reasonable error level from abnormally shallow respiration to high-volume gasping. As far as we know, this is the first study to include uncommonly small respiratory motions in the studied respiratory range and evaluate the FMCW radar technology for apnoea indication.

While our study focused on nocturnal vital sign monitoring applications, the technology can be applicable to various other purposes where the subjects remain mainly still, such as monitoring bedridden patients or the elderly, or finding victims trapped under constructions at disaster scenes. In the future, the methods can be tested on authentic nocturnal measurements and adjusted for more advanced 60 GHz radars, enabling the measurement of multiple subjects simultaneously, despite close proximity [35]. Other remaining challenges include, e.g., decreasing the effect of motion artefacts and reducing the time delay during signal extraction to enable real-time applications.

Supplementary Materials: The following are available online at http://www.mdpi.com/1424-8220/20/22/6505/s1; Aggregated error metrics, Figure S1: Examples of interbeat intervals extracted for each subject during relaxed respiration in the supine lying position, Table S1: Root mean square error for interbeat intervals with respect to activity and participant, Table S2: Root mean square error for interbeat intervals with respect to activity and lying position, Figure S2: Example of interbeat intervals extracted from an arrhythmic sequence, Figure S3: Examples of respiration signals for each subject during relaxed respiration in the supine lying position, Table S3: Root mean square for respiration rates with respect to activity and participant, Table S4: Root mean square error for respiration rates with respect to activity and lying position, Figure S4: Correlation of the interbeat interval derived from the radar signal and the reference IBI, Figure S5: Correlation of the respiration rate derived from the radar signal and the reference respiration rates.

Author Contributions: Conceptualization, E.T., J.M.K. and T.K.; methodology, E.T., J.M.K. and O.A.; software, J.M.K and O.A.; validation, E.T., J.M.K. and O.A.; formal analysis, E.T.; investigation E.T.; resources, J.M.K., O.A. and T.K.; data curation, J.M.K.; writing–original draft preparation, E.T.; writing–review and editing, J.M.K., O.A. and T.K.; visualization, E.T.; supervision, T.K.; project administration, T.K.; funding acquisition, T.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Academy of Finland HISENS project under grant 310879 and Business Finland MAIWAY project under grant 2638/31/2018.

Acknowledgments: The authors thank Johanna Närväinen, Kati Pettersson, Lic. Sc. (Tech.) Johan Plomp, and Mark van Gils of VTT Technical Research Centre of Finland Ltd, and Moncef Gabbouj of Tampere University, Finland, for their valuable comments on the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Ludikhuize, J.; Smorenburg, S.M.; de Rooij, S.E.; de Jonge, E. Identification of deteriorating patients on general wards; measurement of vital parameters and potential effectiveness of the Modified Early Warning Score. *J. Crit. Care* **2012**, *27*, 424.e7–424.e13. [CrossRef] [PubMed]
2. Weenk, M.; van Goor, H.; Frietman, B.; Engelen, L.J.; van Laarhoven, C.J.; Smit, J.; Bredie, S.J.; van de Belt, T.H. Continuous Monitoring of Vital Signs Using Wearable Devices on the General Ward: Pilot Study. *JMIR Mhealth Uhealth* **2017**, *5*, e91. [CrossRef] [PubMed]
3. Chan, A.M.; Selvaraj, N.; Ferdosi, N.; Narasimhan, R. Wireless patch sensor for remote monitoring of heart rate, respiration, activity, and falls. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; pp. 6115–6118. [CrossRef]
4. Adib, F.; Mao, H.; Kabelac, Z.; Katabi, D.; Miller, R.C. Smart Homes That Monitor Breathing and Heart Rate. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Korea, 18–23 April 2015; ACM: New York, NY, USA, 2015; pp. 837–846. [CrossRef]
5. Lubecke, O.B.; Ong, P.; Lubecke, V.M. 10 GHz Doppler radar sensing of respiration and heart movement. In Proceedings of the IEEE 28th Annual Northeast Bioengineering Conference (IEEE Cat. No.02CH37342), Philadelphia, PA, USA, 21 April 2002; pp. 55–56. [CrossRef]
6. Kiuru, T.; Metso, M.; Jardak, S.; Pursula, P.; Häkli, J.; Hirvonen, M.; Sepponen, R. Movement and respiration detection using statistical properties of the FMCW radar signal. In Proceedings of the 2016 Global Symposium on Millimeter Waves (GSMM) ESA Workshop on Millimetre-Wave Technology and Applications, Espoo, Finland, 6–8 June 2016; pp. 1–4. [CrossRef]
7. Mercuri, M.; Lorato, I.R.; Liu, Y.H.; Wieringa, F.; Van Hoof, C.; Torfs, T. Vital-sign monitoring and spatial tracking of multiple people using a contactless radar-based sensor. *Nat. Electron.* **2019**, *2*, 252–262. [CrossRef]
8. Mok, W.Q.; Wang, W.; Liaw, S.Y. Vital signs monitoring to detect patient deterioration: An integrative literature review. *Int. J. Nurs. Pract.* **2015**, *21*, 91–98. [CrossRef] [PubMed]
9. Cretikos, M.A.; Bellomo, R.; Hillman, K.; Chen, J.; Finfer, S.; Flabouris, A. Respiratory rate: The neglected vital sign. *Med J. Aust.* **2008**, *188*, 657–659. [CrossRef] [PubMed]
10. Jardak, S.; Kiuru, T.; Metso, M.; Pursula, P.; Häkli, J.; Hirvonen, M.; Ahmed, S.; Alouini, M. Detection and localization of multiple short range targets using FMCW radar signal. In Proceedings of the 2016 Global Symposium on Millimeter Waves (GSMM) ESA Workshop on Millimetre-Wave Technology and Applications, Espoo, Finland, 6–8 June 2016; pp. 1–4. [CrossRef]
11. AWR1243 76-GHz to 81-GHz High-Performance Automotive MMIC TI.com. Available online: http://www.ti.com/product/AWR1243 (accessed on 11 September 2020).
12. 60 GHz 4TX4TR MIMO—Silicon Radar GmbH. Available online: https://siliconradar.com/products/single-product/60-ghz-4tx4tr-mimo/ (accessed on 11 September 2020).

13. Anitori, L.; de Jong, A.; Nennie, F. FMCW radar for life-sign detection. In Proceedings of the 2009 IEEE Radar Conference, Pasadena, CA, USA, 4–8 May 2009; pp. 1–6. [CrossRef]
14. Wang, S.; Pohl, A.; Jaeschke, T.; Czaplik, M.; Köny, M.; Leonhardt, S.; Pohl, N. A novel ultra-wideband 80 GHz FMCW radar system for contactless monitoring of vital signs. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015, pp. 4978–4981. [CrossRef]
15. Alizadeh, M.; Shaker, G.; Almeida, J.C.M.D.; Morita, P.P.; Safavi-Naeini, S. Remote Monitoring of Human Vital Signs Using mm-Wave FMCW Radar. *IEEE Access* **2019**, *7*, 54958–54968. [CrossRef]
16. Parati, G.; Lombardi, C.; Castagna, F.; Mattaliano, P.; Filardi, P.P.; Agostoni, P. Heart failure and sleep disorders. *Nat. Rev. Cardiol.* **2016**, *13*, 389–403. [CrossRef] [PubMed]
17. Gutierrez, G.; Williams, J.; Alrehaili, G.A.; McLean, A.; Pirouz, R.; Amdur, R.; Jain, V.; Ahari, J.; Bawa, A.; Kimbro, S. Respiratory rate variability in sleeping adults without obstructive sleep apnea. *Physiol. Rep.* **2016**, *4*, e12949. [CrossRef] [PubMed]
18. Stein, P.K.; Pu, Y. Heart rate variability, sleep and sleep disorders. *Sleep Med. Rev.* **2012**, *16*, 47–66. [CrossRef] [PubMed]
19. Quer, G.; Gouda, P.; Galarnyk, M.; Topol, E.J.; Steinhubl, S.R. Inter- and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: Retrospective, longitudinal cohort study of 92,457 adults. *PLoS ONE* **2020**, *15*, e0227709. [CrossRef] [PubMed]
20. Peake, J.M.; Kerr, G.; Sullivan, J.P. A Critical Review of Consumer Wearables, Mobile Applications, and Equipment for Providing Biofeedback, Monitoring Stress, and Sleep in Physically Active Populations. *Front. Physiol.* **2018**, *9*, 743. [CrossRef] [PubMed]
21. Parak, J.; Korhonen, I. Evaluation of wearable consumer heart rate monitors based on photopletysmography. In Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; pp. 3670–3673. [CrossRef]
22. Teisala, T.; Mutikainen, S.; Tolvanen, A.; Rottensteiner, M.; Leskinen, T.; Kaprio, J.; Kolehmainen, M.; Rusko, H.; Kujala, U.M. Associations of physical activity, fitness, and body composition with heart rate variability–based indicators of stress and recovery on workdays: A cross-sectional study. *J. Occup. Med. Toxicol.* **2014**, *9*. [CrossRef] [PubMed]
23. Kinnunen, H.O.; Koskimäki, H. 0312 The HRV Of The Ring—Comparison Of Nocturnal HR And HRV Between A Commercially Available Wearable Ring And ECG. *Sleep* **2018**, *41*, A120. [CrossRef]
24. Charvat, G.L. *Small and Short-Range Radar Systems*; CRC Press: Boca Raton, FL, USA, 2014.
25. *Embla Titanium Clinical Manual*; Revision 3.0; Micromed S.p.A.: Zona Industriale S.p.z., Italy, 2007.
26. Kortelainen, J.M.; van Gils, M.; Pärkkä, J. Multichannel bed pressure sensor for sleep monitoring. In Proceedings of the 2012 Computing in Cardiology, Krakow, Poland, 9–12 September 2012; pp. 313–316.
27. Bogert, B.P.; Healy, M.J.R.; Tukey, J.W. The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In *Time Series Analysis*; John Wiley & Sons, Inc.: New York, NY, USA, 1963; Chapter 15, pp. 209–243.
28. Kortelainen, J.M.; Virkkala, J. FFT averaging of multichannel BCG signals from bed mattress sensor to improve estimation of heart beat interval. In Proceedings of the 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, 22–26 August 2007; pp. 6685–6688. [CrossRef]
29. Peltola, M. Role of editing of R-R intervals in the analysis of heart rate variability. *Front. Physiol.* **2012**, *3*, 148. [CrossRef] [PubMed]
30. Box, G.E.P.; Jenkins, G.M. *Time Series Analysis: Forecasting and Control*, 2nd ed.; Holden-Day: San Francisco, CA, USA, 1976.
31. Ludbrook, J. Special article comparing methods of measurement. *Clin. Exp. Pharmacol. Physiol.* **1997**, *24*, 193–203. [CrossRef] [PubMed]
32. Bland, J.; Altman, D. Comparing methods of measurement: Why plotting difference against standard method is misleading. *Lancet* **1995**, *346*, 1085–1087. [CrossRef]
33. Oppenheim, A.V.; Schafer, R.W. *Discrete-Time Signal Processing*, 1st ed.; Prentice-Hall, Inc.: Upper sader River, NJ, USA, 1989.

34. Brüser, C.; Kortelainen, J.M.; Winter, S.; Tenhunen, M.; Pärkkä, J.; Leonhardt, S. Improvement of Force-Sensor-Based Heart Rate Estimation Using Multichannel Data Fusion. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 227–235. [CrossRef] [PubMed]
35. Forsten, H.; Kiuru, T.; Hirvonen, M.; Varonen, M.; Kaynak, M. Scalable 60 GHz FMCW frequency-division multiplexing MIMO radar. *IEEE Trans. Microw. Theory Tech.* **2020**, *68*, 2845–2855. [CrossRef]

PUBLICATION
II

Assessing fatigue and sleep in chronic diseases using physiological signals from wearables: A pilot study

E. Antikainen, H. Njoum, J. Kudelka, D. Branco, R. Z. U. Rehman, V. Macrae, K. Davies, H. Hildesheim, K. Emmert, R. Reilmann, C. J. van der Woude, W. Maetzler, W.-F. Ng, P. O'Donnell, G. Van Gassen, F. Baribaud, I. Pandis, N. V. Manyakov, M. van Gils, T. Ahmaniemi, and M. Chatterjee

Frontiers in Physiology, vol. 13, no. 968185
DOI: 10.3389/fphys.2022.968185

frontiers | Frontiers in Physiology

TYPE Original Research
PUBLISHED 14 November 2022
DOI 10.3389/fphys.2022.968185

OPEN ACCESS

EDITED BY
Jordi Aguilo,
Universitat Autònoma de Barcelona, Spain

REVIEWED BY
Youngsun Kong,
University of Connecticut, United States
Junichiro Hayano,
Heart Beat Science Lab, Co., Ltd., Japan

*CORRESPONDENCE
Emmi Antikainen,
emmi.antikainen@gmail.com
Meenakshi Chatterjee,
mchatte4@ITS.JNJ.com

SPECIALTY SECTION
This article was submitted to Physiological Logging, a section of the journal Frontiers in Physiology

RECEIVED 13 June 2022
ACCEPTED 31 October 2022
PUBLISHED 14 November 2022

CITATION
Antikainen E, Njoum H, Kudelka J, Branco D, Rehman RZU, Macrae V, Davies K, Hildesheim H, Emmert K, Reilmann R, Janneke van der Woude C, Maetzler W, Ng W-F, O'Donnell P, Van Gassen G, Baribaud F, Pandis I, Manyakov NV, van Gils M, Ahmaniemi T and Chatterjee M (2022), Assessing fatigue and sleep in chronic diseases using physiological signals from wearables: A pilot study. *Front. Physiol.* 13:968185. doi: 10.3389/fphys.2022.968185

Assessing fatigue and sleep in chronic diseases using physiological signals from wearables: A pilot study

Emmi Antikainen[1]*, Haneen Njoum[2], Jennifer Kudelka[3], Diogo Branco[4], Rana Zia Ur Rehman[5], Victoria Macrae[6], Kristen Davies[5], Hanna Hildesheim[3], Kirsten Emmert[3], Ralf Reilmann[7,8,9], C. Janneke van der Woude[10], Walter Maetzler[3], Wan-Fai Ng[5,6], Patricio O'Donnell[11], Geert Van Gassen[12], Frédéric Baribaud[13], Ioannis Pandis[14], Nikolay V. Manyakov[15], Mark van Gils[16], Teemu Ahmaniemi[1] and Meenakshi Chatterjee[17]* on behalf of the IDEA-FAST project consortium

[1]VTT Technical Research Centre of Finland Ltd., Tampere, Finland, [2]Sanofi R&D, Frankfurt, Germany, [3]Department of Neurology, University Hospital Schleswig-Holstein, Kiel University, Kiel, Germany, [4]LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal, [5]Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle Upon Tyne, United Kingdom, [6]NIHR Newcastle Biomedical Research Centre and NIHR Newcastle Clinical Research Facility, Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle Upon Tyne, United Kingdom, [7]George-Huntington-Institute, University of Münster, Münster, Germany, [8]Department of Clinical Radiology, University of Münster, Münster, Germany, [9]Department of Neurodegenerative Diseases and Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany, [10]Department of Gastroenterology and Hepatology, Erasmus MC, Rotterdam, Netherlands, [11]Department of Psychiatry, Harvard Medical School, McLean Hospital, Belmont, MA, United States, [12]Takeda Belgium, Zavente, Belgium, [13]Bristol Myers Squibb, New York, NY, United States, [14]Janssen Research & Development, London, United Kingdom, [15]Janssen Research & Development, Beerse, Belgium, [16]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland, [17]Janssen Research & Development, Cambridge, MA, United States

Problems with fatigue and sleep are highly prevalent in patients with chronic diseases and often rated among the most disabling symptoms, impairing their activities of daily living and the health-related quality of life (HRQoL). Currently, they are evaluated primarily *via* Patient Reported Outcomes (PROs), which can suffer from recall biases and have limited sensitivity to temporal variations. Objective measurements from wearable sensors allow to reliably quantify disease state, changes in the HRQoL, and evaluate therapeutic outcomes. This work investigates the feasibility of capturing continuous physiological signals from an electrocardiography-based wearable device for remote monitoring of fatigue and sleep and quantifies the relationship of objective digital measures to self-reported fatigue and sleep disturbances. 136 individuals were followed for a total of 1,297 recording days in a longitudinal multi-site study conducted in free-living settings and registered with the German Clinical Trial Registry (DRKS00021693). Participants comprised healthy individuals (N = 39) and patients with neurodegenerative disorders (NDD, N = 31) and immune mediated inflammatory diseases (IMID, N = 66). Objective physiological

measures correlated with fatigue and sleep PROs, while demonstrating reasonable signal quality. Furthermore, analysis of heart rate recovery estimated during activities of daily living showed significant differences between healthy and patient groups. This work underscores the promise and sensitivity of novel digital measures from multimodal sensor time-series to differentiate chronic patients from healthy individuals and monitor their HRQoL. The presented work provides clinicians with realistic insights of continuous at home patient monitoring and its practical value in quantitative assessment of fatigue and sleep, an area of unmet need.

KEYWORDS

wearabe sensors, chronic disease, biomedical signal analysis, fatigue, sleep disturbance, continuous monitoring, neurodegenerative diseases, immune-mediated inflammatory disease

1 Introduction

Health-related quality of life (HRQoL) and ability to conduct activities of daily living (ADL) are greatly impaired in patients with chronic diseases, such as neurodegenerative disorders (NDD) and immune mediated inflammatory diseases (IMID) (Kluger et al., 2013; Zielinski et al., 2019). Fatigue and sleep disturbances are known to be key factors predicting poor HRQoL or reduced ADLs, and as such alleviation of these symptoms may significantly improve patient's health and quality of life (Center for Disease Control and Prevention, 2000). Current evaluations rely primarily on patient reported outcomes (PROs) which are subjective and prone to recall biases and poorly capture variability over time (Stone et al., 2002). Sensors, such as wearable technology or standalone sensors using a wide range of technologies, can perform continuous real-world monitoring of patient health and thus offer the opportunity to provide digital measures that are objective, potentially reliable and more sensitive to change over time (Bangerter et al., 2020a; 2020b; Luo et al., 2020).

Fatigue is defined as a multi-dimensional phenomenon in which the biophysiological, cognitive, motivational and emotional state of the body is affected resulting in significant impairment of the individual's ability to function in their normal capacity (Davies et al., 2021). Specifically, in NDD and IMID patients, such as those with Huntington's Disease (HD), Parkinson's Disease (PD), Inflammatory Bowel Diseases (IBD), Primary Sjögren's Syndrome (PSS), Rheumatoid Arthritis (RA), and Systemic Lupus Erythematosus (SLE), fatigue and sleep disturbances are highly prevalent (Hewlett et al., 2011; Lendrem et al., 2014; Siciliano et al., 2018; Chavarría et al., 2019). Previous studies assessing fatigue through digital measurement technologies are relatively sparse, especially in chronic disease populations. Changes in physical activity levels such as daily and bouted moderate to vigorous physical activity (MVPA) minutes and no bouts of MVPA have been found to be associated with fatigue in RA, SLE and Crohn's disease (Legge et al., 2017). Fatigue has also been shown to be correlated with changes in the frequency spectrum of EEG signals (Zhang et al., 2020). Individuals with chronic fatigue syndrome were found to have lower heart rate variability (HRV) measures such as standard deviation of the interbeat intervals of normal sinus beats (SDNN), power spectrum densities of low frequency (LF) and high frequency (HF) compared to controls, while total HRV power within the frequency range of 0–0.4 Hz was shown to be negatively associated with fatigue (Boissoneault et al., 2019; Escorihuela et al., 2020).

Sleep disorders such as decreased sleep efficiency and increased fragmentation are the second most frequent complaint in PD (Stefani and Högl, 2019). In HD, sleep and circadian rhythm alterations have been reported to correlate with depression and cognitive impairment (Aziz et al., 2010). Sleep disturbances, also common in RA, SLE, IBD, and PSS, have been attributable to changes in circadian rhythms or disease symptoms such as pain, discomfort, respiratory and movement disorders sleep, with disruptions in sleep associated with further worsening of disease symptoms (Swanson and Burgess, 2017). Recording night ECG allows evaluation of the fluctuation of the sympathetic and parasympathetic nervous system functions, which physiologically happen during sleep. The LF (frequency range 0.04–0.15 Hz) reflects both sympathetic and vagal modulations, which decrease with the depth of sleep. The HF (frequency range 0.15–0.4 Hz) is associated with respiration and reflects the activity of the parasympathetic nervous system, which increases in deep sleep (Somers et al., 1993).

Digital measures that can objectively assess HRQoL-related factors, such as sleep and fatigue will be invaluable for drug development. Despite the advent of wearable sensors, there is limited understanding of fatigue and sleep assessment using objective measurements in these patient population, with existing work primarily focusing on the relationship between physical activity measured from accelerometers with fatigue PROs. Even among healthy cohorts, only few recent studies have utilized wearable sensors (Luo et al., 2020) such as inertial measurement units and heart rate monitors to assess

fatigue, majority with smaller sample size or under tightly controlled experimental settings. Building on these challenges, the IDEA-FAST project (https://idea-fast.eu/) aims to utilize multiple sensing modalities and technologies at home to identify digital endpoints of fatigue and sleep in the six NDD and IMID populations–HD, PD, IBD, PSS, RA, and SLE.

In this paper, we present insights from a feasibility study of IDEA-FAST (The IDEA-FAST project consortium, 2020) and focus specifically on evaluating the promise of capturing digital measures of fatigue and sleep from biophysiological signals collected in patients and healthy groups at home from a wearable ECG device. Specifically, signal quality and coverage of digital measures were assessed and their agreement with sleep and fatigue PROs were investigated. Furthermore, heart rate recovery (HRR) periods were estimated, among patients and healthy participants, as a metric to assess physiological fitness which could potentially be impacted by fatigue. Post-exercise heart rate recovery reflects the interplay between the sympathetic and parasympathetics parts of the autonomic nervous system (Qiu et al., 2017). It is an important predictor of all-cause mortality and related to fatal cardiovascular events (Qiu et al., 2017). Decrease in HRR is shown to be associated with physical fatigue (Lamberts et al., 2009; Djaoui et al., 2017) and has been typically measured in controlled laboratory settings. Here we explored if HRR quantified from free-living environments can distinguish between NDD, IMID and healthy groups and those with varying levels of fatigue.

2 Materials and methods

The presented data was obtained as a part of the IDEA-FAST project (The IDEA-FAST project consortium, 2020; Chen et al., 2022). Nine different candidate technologies measuring different modalities (activity trackers, ECG-sensors, sleep trackers) were explored in a feasibility study aiming to assess fatigue and sleep disorders. Additionally, the participants' social activity, cognitive skills, and PROs were captured with smartphone applications. This paper focuses on the continuously measured physiological signals collected from the ECG-based VitalPatch sensor and the concurrently collected PROs. The digital measures from VitalPatch included heart rate (HR), R-to-R interval, respiratory rate (RR), skin temperature (skin T), number of steps, and posture. The first three are mainly derived from the ECG measurement and are the main focus of this study.

2.1 Ethical approvals

Ethical approval was first granted by the Ethical Committee of the Medical Faculty of Kiel University (D491/20) in June 2020 and then by the Research Ethics Committees of all other study sites: Newcastle upon Tyne Hospitals National Health Service (NHS) Foundation Trust/Newcastle University in August 2020, Erasmus University Medical Centre in Rotterdam in November 2020, and George-Huntington-Institute in Münster in September 2020. The study was registered with the German Clinical Trial Registry (DRKS00021693) and was conducted according to the principles of the Declaration of Helsinki (version of 2013).

2.2 Study participants

Potential participants were identified during routine clinical visits at the hospitals and through public outreach at information events or support groups. After providing information about the study and obtaining informed consent, the participants were screened for eligibility. Inclusion criteria required age over 18 years, consent to participate in the study for up to 60 days and according to the study protocol, use of a smartphone in the past 3 months, and ability to follow written and oral instructions in the local language, to walk, sit, and stand independently and to socialize and communicate. Another inclusion criterion was a score of over 15 points in the Montreal Cognitive Assessment (MoCA), which was used to evaluate cognitive abilities (Nasreddine et al., 2005). Participants were excluded if they had certain comorbidities like major sleep disorders, chronic fatigue syndrome, respiratory, cardiovascular or metabolic disorders or physical traumas with hospitalization in the past 3 months, diagnosis of cancer in the past 3 years, major psychiatric disorders, suicidal attempt in the past 5 years or suicidal ideation in the past 6 months, substance or ethanol abuse or severe visual impairment.

The study was conducted at four different sites: Rotterdam (E), Kiel (K), Muenster (G), and Newcastle (N). The study start date was between July 2020 (Kiel) and November 2020 (Rotterdam), depending on the date of ethical approval of the study site. The last visit of the final participant took place in December 2021. The participants were either healthy or suffered from one of six diseases, which we have divided into two groups: NDDs including HD and PD, and IMIDs including IBD, PSS, RA, and SLE. Thus, the study inspects three participant categories: 1) the healthy participants, 2) the NDD patients, and 3) the IMID patients.

2.3 Study design

Participants were enrolled in the study for up to 60 days. Demographic information was collected during a baseline visit conducted at the study center or at the participant's home. Subsequently, the participants were provided with a detailed explanation of the devices and the applications. In addition, they received informational materials and telephone support by the

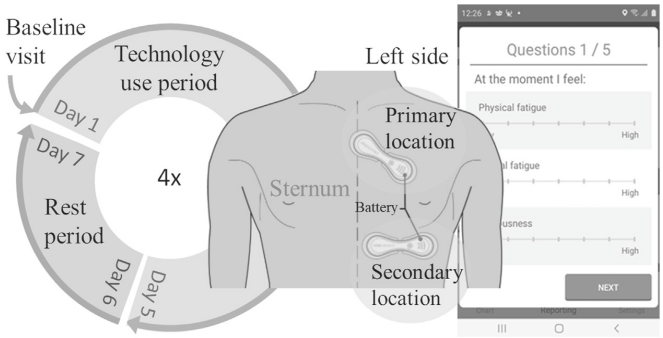

FIGURE 1
The study cycle. Participants went through a baseline visit at the beginning of the study, followed by four technology use periods (the circle) when they were instructed to use different sensors and report certain outcomes with a mobile phone application (example screenshot on the right). This study focuses on the participants using the depicted wearable device, VitalPatch (two alternative wear locations depicted).

study team. Optional home visits were conducted to further ensure accurate use of the devices.

Over a period of five consecutive days, participants wore the VitalPatch biosensor in their home environment and were instructed to carry out their usual daily activities. This constituted one technology use period that was followed by at least two rest days, after which a new technology use period could be started. Participants were able to opt for a prolonged resting period. The study cycle, illustrated in Figure 1, was repeated up to four times per participant. During the technology use period, participants were asked to report their perceptions of fatigue and sleep quality four times daily in an e-diary using the VTT Stress Monitor Application (SMA) (Vildjiounaite et al., 2018).

2.4 Measurement setup

VitalPatch is a wireless wearable patch sensor designed for remote patient monitoring (Areia et al., 2021). The fully disposable 12-cm patch adheres to the skin and is worn on the left chest. It contains a zinc-air battery that lasts up to 7 days. Once the measurement is started, it continues whenever the device is in skin contact, until the battery runs out. After one patch sensor is disposed, the measurement can be continued with a new patch. VitalPatch has CE certification as a Class IIa medical device and FDA clearance.

The VitalPatch biosensor incorporates a single-lead ECG, a tri-axial accelerometer, and a thermistor. It records ECG at 125 Hz sampling frequency, with derived heart rate, R-to-R interval, and respiratory rate (partly derived from the accelerometer) sampled at 0.25 Hz. The accelerometer is used for step counting and posture detection at 1 Hz. The thermistor collects skin temperature at 0.25 Hz. The recorded data is encrypted and transferred with a latency in the order of seconds *via* a wireless connection to a cloud-based patient monitoring platform. If the connection is interrupted, the device can store up to 10 h of data until the connection is re-established.

2.5 Patient reported outcomes

PROs were collected using the VTT Stress Monitor Application (SMA), an Android smartphone application that provides a user interface for questionnaires (Vildjiounaite et al., 2018). PROs were collected four times a day (at 9:00, 13:00, 17:00, and 21:00 local time). The response could be submitted within 3 hours of the prompted question, except in the evening as those responses were set due at 23:30. To promote compliance, the application prompted a new notification again every 15 min if the user had opened the application but did not submit the responses, and the application had gone out of active view. Throughout the day, the participants were requested to respond to a total of 14 different PROs, as detailed in Table 1. All Likert items had seven options from low (zero) to high (six). An example of the Likert item interface is presented in Figure 1.

2.6 Data pre-processing

The HR, R-to-R interval, RR, and skin T data were pre-processed in two steps. First, timestamps were sorted, and duplicates removed. Second, the data were cleaned from 1) invalid values (unsuccessfully measured) predefined by the manufacturer, 2) physiologically unrealistic values, and 3) contextual outliers. Such values were removed and considered

TABLE 1 Patient reported outcomes collected with the VTT Stress Monitor Application.

		Questionnaire time			
Patient reported outcome	**Type**	**Morning (9–12)**	**Early after-noon (13–16)**	**Late after-noon (17–20)**	**Evening (21–23:30)**
Physical fatigue	Likert item	X	X	X	X
Mental fatigue	Likert item	X	X	X	X
Anxiousness	Likert item	X	X	X	X
Depression	Likert item	X	X	X	X
Pain	Likert item	X	X	X	X
I went to bed at	Clock	X			
I woke up	Clock	X			
How was your sleep?	Likert item	X			
Time to fall asleep	Drop-down menu	X			
Time awake during night	Drop-down menu	X			
Sleepiness, current feeling	Drop-down menu		X	X	X
My activities of the day, physically	Likert item				X
My activities of the day, mentally	Likert item				X
Other comments	Free text				X

TABLE 2 Accepted range for each physiological feature. The selected ranges were validated visually and by comparing them against the 1st and 99th percentiles of the collected data.

	Heart rate (bpm)	**R-to-R interval (ms)**	**Respiratory rate (bpm)**	**Skin temperature (°C)**
Minimum	30	300	4	28
Maximum	200	2000	60	40

gaps in the data, except for cases 2–3 for R-to-R interval, which were replaced using linear interpolation to improve heart rate variability (HRV) analysis (detailed below). The first pre-processing step along with the removal of invalid values were also applied to the number of steps and posture.

To exclude any physiologically unrealistic values, a range of acceptable values was defined for each feature independently. The selected limits are presented in Table 2. The limits for HR and R-to-R interval are adopted from previous studies (Tanaka et al., 2001; Zhai et al., 2020). The range for respiratory rate, on the other hand, was defined broadly, including abnormal hypo- and hyperventilation scenarios, such as exercise (Cretikos et al., 2008; Gutierrez et al., 2016; Nicolò et al., 2017, 2020). Finally, skin temperature is presumed to obtain lower values as compared to core body temperature but is allowed a range that can capture abnormal physiological states (Martinez-Nicolas et al., 2015; Rajbhandary and Nallathambi, 2020). Restricting the range is expected to exclude notably exceptional measurement conditions, even though the thermal sensor itself has been reported to work accurately across a wider range (Selvaraj et al., 2018).

Contextual outliers were removed to reduce unlikely variation within short time periods. Using a sliding window, the value at the centre of the window was inspected: if it was not within a predefined range of the window mean, it was considered a contextual outlier (Karlsson et al., 2012). The ranges were 30% for all features except for RR; threshold for respiration rate was 50%. As respiration can be controlled at will, it is more prone to larger variations. The size of the sliding window was 1 min for HR and R-to-R interval, 3 min for RR, and 5 min for skin T.

Patient reported outcomes did not require pre-processing, apart from the sleep times: they were collected with a 24-h clock user interface, which was discovered prone to 12-h shifts in the user input, especially when reporting late hours (12–24). Bedtimes that exceeded the waking up or occurred considerably late with respect to the wake-up time, were considered as input errors and shifted by 12 h.

TABLE 3 Heart rate and heart rate variability features.

Abbreviation	Domain	Description
NN mean	Time	Mean of normal-to-normal peak intervals (NN)
NN CV	Time	Coefficient of variation of NN
NN SD	Time	Standard deviation of NN
NN median	Time	Median of NN
NN range	Time	Difference between maximum and minimum of NN
RMSSD	Time	Root mean square of consecutive differences in adjacent NN
CVSD	Time	Coefficient of variation of consecutive differences in adjacent NN
SDSD	Time	Standard deviation of consecutive differences in adjacent NN
NN50	Time	Number of interval differences greater than 50 ms
NN20	Time	Number of interval differences greater than 20 ms
pNN50	Time	Percentage of interval differences greater than 50 ms
pNN20	Time	Percentage of interval differences greater than 20 ms
HRV HR mean	Time	Heart rate mean
HRV HR SD	Time	Heart rate standard deviation
HRV HR min	Time	Heart rate minimum
HRV HR max	Time	Heart rate maximum
VLF	Frequency	Power spectral density in very low frequencies (0.003–0.04 Hz)
LF	Frequency	Power spectral density in low frequencies (0.04–0.15 Hz)
HF	Frequency	Power spectral density in high frequencies (0.15–0.40 Hz)
Total power	Frequency	Total power spectral density; sum of VLF, LF, and HF
LF/HF	Frequency	The ratio of LF and HF
LFnu	Frequency	LF normalized to the sum of LF and HF
HFnu	Frequency	HF normalized to the sum of LF and HF
Triangular index	Geometrical	Number of all NN divided by the maximum of the NN density distribution
CSI	Non-linear	Cardiac sympathetic index
mCSI	Non-linear	Modified cardiac sympathetic index
CVI	Non-linear	Cardiac vagal index
SD1	Non-linear	Poincaré plot, SD1
SD2	Non-linear	Poincaré plot, SD2
SD2/SD1	Non-linear	SD2 to SD1 ratio

2.7 Data quality assessment

The quality of the digital measures was assessed *via* the extent of pre-processing necessary [corresponding to items (1), (2), (3) as described in Section 2.6], and the data coverage after pre-processing. For HR, RR and skin T, coverage was calculated based on the expected number of samples. Coverage of R-to-R interval was estimated *via* the sum of recorded R-to-R interval values divided by the duration of the actual measurement period. Data quality was first evaluated on participant-level and then averaged over cohorts or participant groups. The participant-level coverage was computed as the mean of midnight-to-midnight coverage values.

Additionally, PRO coverage during VitalPatch wear periods was evaluated for each PRO as compared to the expected number of responses. PRO coverage was also evaluated midnight-to-midnight for each participant and then averaged over participant subgroups.

2.8 Feature aggregates

The features were segmented into time windows of interest (see Section 2.9) and aggregated into statistical descriptors, to summarize the physiological feature time series into single values, which could be compared to the corresponding PROs.

TABLE 4 Demographics of study participants using VitalPatch.

Cohort group	Cohort	Sites	N	Female	Male	Years since diagnosis, mean (SD)	Years since diagnosis	Age, mean (SD)	Age range	BMI, mean (SD)
Healthy	Healthy	All	39	20	19	—	—	47.3 (16.3)	21–77	26.3 (4.9)
NDD	HD	G, K	13	7	6	4.8 (2.7)[a]	0–8[a]	44.2 (9.6)	30–60	26.3 (7.1)[b]
	PD	K	18	7	11	7.8 (5.9)	1–18[a]	62.3 (11.0)	37–80	24.3 (2.4)
IMID	IBD	E	18	9	9	12.9 (10.8)	1–35	36.7 (11.3)	22–55	24.7 (3.4)
	PSS	N	18	16	2	11.6 (5.4)	4–27	62.6 (13.1)	37–82	21.9 (10.3)
	RA	K, N	14	11	3	14.1 (9.4)	3–35	64.6 (12.2)	39–79	29.5 (7.9)
	SLE	K, N	16	16	0	16.7 (9.9)[a]	4–34[a]	48.3 (13.1)	31–80	23.1 (10.5)
Total	*7 cohorts*	*4 sites*	*136*	*86 (63.2%)*	*50 (36.8%)*	*11.2 (8.4)*	*0–35*	*51.6 (16.1)*	*21–82*	*25.2 (7.0)*

[a]Four HD, patients, one PD, patient, and nine SLE, patients with unknown years since diagnosis.
[b]Four HD, patients with unknown BMI.

The selected statistical aggregations were the mean, standard deviation (SD), minimum, and maximum.

Additionally, HRV parameters were computed from the R-to-R interval data over each full window. Furthermore, the feature coverage within each window was computed for reliability evaluation. For HRV analysis, the R-to-R interval data was further cleaned to achieve normal-to-normal (NN) intervals by replacing ectopic peaks using linear interpolation (Peltola, 2012). This was performed *via* the Malik method: intervals deviating more than 20% from the previous interval were replaced (Malik et al., 1996). Both time and frequency domain HRV features were computed, as well as geometric and non-linear features (Malik et al., 1996; Champseix, 2021). The included HRV features are described in Table 3. Details of these widely-used HRV parameters and their implications in health and performance can be found in existing literature (Shaffer and Ginsberg, 2017).

2.9 Time windows of interest

The feature aggregates were computed over 2-h windows preceding the time at which a PRO response was obtained. Thus, the aggregation represents the participants' physiological features leading to the questionnaire response.

To estimate physiological measures during rest, major rest periods were identified for each participant, using step count and posture information available from VitalPatch. As a proxy for major rest periods, the L5 metric was calculated which corresponded to the least active 5-h (L5) periods in the day (Witting et al., 1990). A maximum of 100 steps was allowed and a minimum of 80% laying down was used as a threshold. The starting times of the 5 h long resting windows were located at 1 min resolution, and the best option among overlapping consecutive windows was selected by maximizing the laying down percentage.

2.10 Feature normalization

Physiological parameters are affected by the subject's age and sex (and physical fitness) and the inter-individual differences can be significant (Voss et al., 2015; Garavaglia et al., 2021). Therefore, the 2-h feature aggregates were normalized on a subject-by-subject basis to alleviate the differences. Previous studies have normalized HRV features by adjusting the feature according to feature baseline and range, adjusted with the 5th and 95th percentiles to account for outlier effects (Wijsman et al., 2011; Xiao et al., 2013; Altini et al., 2014; Altini and Kinnunen, 2021). In this study, the features are normalized relative to the L5 aggregates, according to

$$x_{norm} = \frac{x - \mu_{L5}}{\sigma_{L5}},$$

where x is a feature aggregate (over a 2h window of interest), x_{norm} is the normalized feature aggregate, μ_{L5} is the mean feature value and σ_{L5} its standard deviation obtained as the mean and SD (a) from the nearest previous L5 window, or (b) averaged over all subject specific L5 windows. In approach (a), the specific instance of x was excluded if no previous L5 window existed. Normalization was applied to all physiological feature aggregates (excluding the feature coverage).

2.11 Feature association with patient reported outcomes

The association between the above-described 2-h feature aggregates and the PROs were studied through repeated measures correlation, to account for intra-individual dependencies in the data (Bakdash and Marusich, 2017). Significance level α was set to 0.05. Feature aggregates

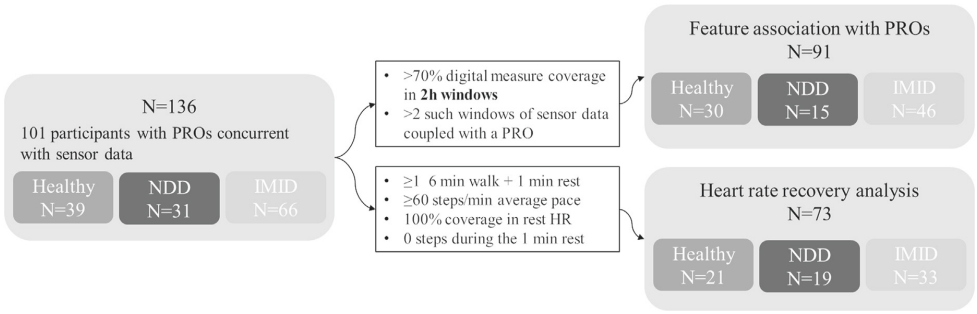

FIGURE 2
Participant flow diagram. The number of participants (N) in total and included in different analyses, presented with the inclusion criteria (black-bordered box).

demonstrating lower than 70% coverage over the window of interest were excluded from the association analysis. Moreover, only participants with at least three pairs of PROs and feature aggregates were included. Repeated measures correlation values close to one indicate linear correlation between the two compared measures.

2.12 Heart rate recovery

Heart rate recovery (HRR) was defined as the maximum difference in the HR signal provided by the VitalPatch sensor that was observed during a 1 min resting period after a 6-min walk, similarly to a six-minute walking test (Roberts et al., 2006; Bellet et al., 2012). Because the measurements were conducted in free-living settings, applicable sequences were retrospectively detected from the clean (non-aggregated) sensor data. The walking periods were identified *via* the "walking" posture, as classified by the wearable sensor. A walk was required to last at least 6 min, but no upper limit was applied. Small pauses in walking and changes of posture lasting up to 3 s were ignored (Del Din et al., 2016). However, a minimum average cadence of 60 steps/min was required (Sokas et al., 2021). For the 1-min resting periods, we required 100% heart rate coverage and zero taken steps. In case of multiple applicable sequences, the highest HRR for a participant was selected as the representative value.

2.13 Statistical analysis

One-way and two-sided Analysis of Covariance (ANCOVA) was used to assess whether the HRR differs significantly among the three participant groups (healthy, NDD, and IMID). The significance level α was again 0.05. Age and gender were taken as covariates and the effect size was evaluated using partial η^2 (eta squared). Pairwise differences between the groups were analysed in post hoc tests performed with Tukey's method, which adjusts the *p*-values for multiple comparisons.

All presented boxplots depict the median as the horizontal line within the box, the interquartile range (IQR) *via* the box limits, and 1.5 times the IQR through the whiskers (points falling outside this range are displayed individually as outliers).

3 Results

3.1 Participant number and demographics

Continuous physiological monitoring of VitalPatch was conducted on 136 participants, 101 of which responded to PROs collected during the patch measurement period. Participants recorded VitalPatch data on 1–21 days, summing up to a total of 1,297 days. Table 4 describes the demographics in each cohort. The patients were diagnosed on average 11.2 years before participation (SD 8.4, ranging from less than a year to 35 years, excluding 14 unknown time of diagnosis). All disease cohorts, excluding HD, included at least one participant unable to work (14 in total), while other participants worked full- or part-time, or were retired. Some IBD participants even worked several part-time jobs. While most participants were Caucasian, four participants were of Asian ethnicity and belonged to IMID group. One participant was African American and belonged to the healthy group.

A flow diagram illustrating different stages of analyses and their participant sample size is shown in Figure 2. A subset of 91 participants were applicable for the analysis of association

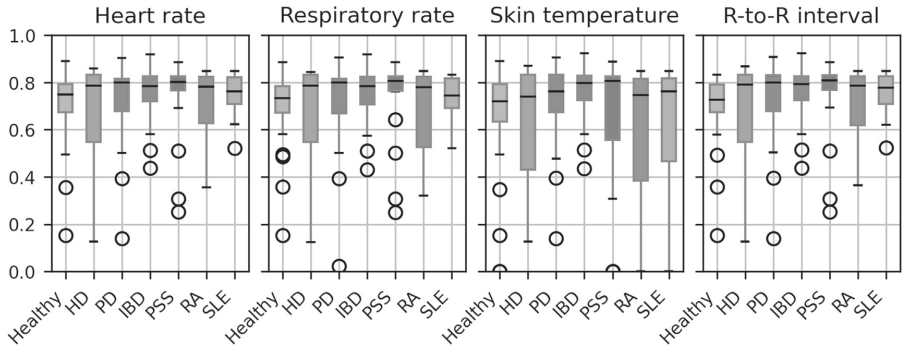

FIGURE 3
Average daily coverage of the cleaned digital measures across participants, presented by cohort.

TABLE 5 Mean (with 95% confidence interval) feature coverage (%) in the 2-h windows preceding PRO responses, presented by participant group. Confidence intervals are based on the empirical rule (2*SD).

	Heart rate	R-to-R interval	Respiratory rate	Skin temperature
Healthy	99.3 [91.6, 100]	99.4 [91.7, 100]	98.9 [87.0, 100]	89.0 [27.7, 100]
NDD	99.4 [93.7, 100]	99.6 [94.0, 100]	99.4 [93.3, 100]	99.6 [94.1, 100]
IMID	99.3 [93.3, 100]	99.5 [93.7, 100]	98.9 [86.7, 100]	92.3 [40.6, 100]
Total	99.3 [92.8, 100]	99.5 [93.1, 100]	99.0 [87.6, 100]	92.5 [41.3, 100]

between the digital measures and PROs. The concurrent measurements of digital measures and PRO data totalled 632 days, varying between 1 and 12 days per participant, and included 15 NDD patients (6 HD, 9 PD), 46 IMID patients (12 IBD, 13 PSS, 10 RA, 11 SLE), and 30 healthy controls.

All VitalPatch data were scanned for sequences applicable for heart rate recovery analysis. In total, 73 participants were included in the HRR analysis, comprising 19 NDD patients (9 HD, 10 PD), 33 IMID patients (11 IBD, 8 PSS, 8 RA, 6 SLE), and 21 healthy participants.

3.2 Data quality

In all VitalPatch data measured throughout the study, 2.3% of skin temperature data were range outliers, while only contextual outliers were identified for HR and RR (0.2% and 0.1%, respectively). For R-to-R intervals, less than 0.5% were outliers (0.1% invalid, 0.3% range and 0.1% contextual outliers). After outlier processing, the average daily coverage rates were 71.6% (16.3% SD) for HR, 71.7% (16.3% SD) for R-to-R interval, 70.9% (16.9% SD) for RR, and 65.5% (25.3% SD) for skin T.

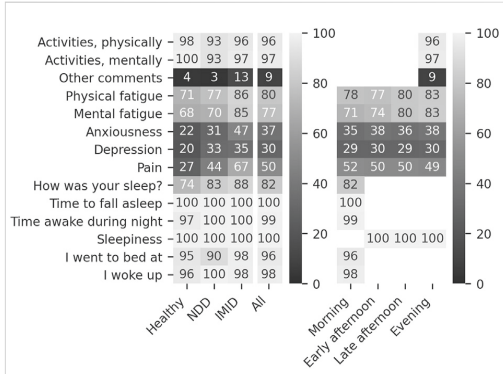

FIGURE 4
Coverage (%) of PROs per participant group (left) and questionnaire timing (right).

Moreover, the median daily coverage was 77% for skin T and about 78% for all other features. Hence, the sensor was typically worn for most of the day.

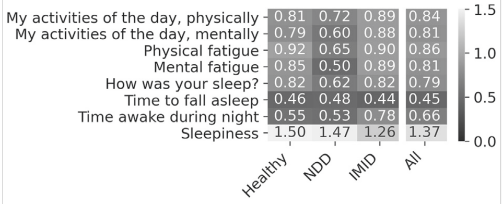

FIGURE 5
The SD of responses across participant groups. Here, the PROs correspond to the feature aggregates normalized with the participant-mean L5 window parameters.

Figure 3 illustrates the obtained coverage for each digital measure in each study cohort. While the smaller cohort groups (HD, RA) exhibit higher variation, the medians are comparable across cohorts. Notably, 11 participants stand out with zero coverage for skin T. However, further inspection revealed that these participants were recruited at the Newcastle site. The outliers could potentially indicate a need for improved device usage instructions at one site.

Table 5 presents the coverage of digital measures in the selected 2-h windows, analysed for association with PROs. Only skin temperature, which failed for 11 participants (including healthy, PSS, RA, and SLE participants), shows notable coverage differences across participant groups. The median coverage in the 2-h windows was 100% for all measures (10% percentile was 95.9% for skin temperature and above 99% for all other measures).

The coverage of PROs corresponding to the 2-h feature aggregates, analysed for association with digital measures, are presented in Figure 4. The analysis focuses on Likert item or drop-down menu PROs with overall coverage beyond 70%.

Figure 5 presents SD captured in the PRO responses. The median number of distinct responses received from a participant was 2 for the activity and sleep detail questions, 3 for the fatigue questions, and 4 for the sleepiness question. The PRO response distributions were similar from participant group to another. Because the drop-down menu PROs (time to fall asleep and time awake during night) exhibit low variability, they are excluded from further analyses.

Self-reported sleep times were obtained for 244 nights concurrent with the VitalPatch data, allowing a comparison between reported sleep times and the extracted L5 periods. Overall, 68.4% of the L5 periods were entirely within the reported sleep time (82.0% started and 86.5% ended within the reported sleep time), and 86.9% were within a 30-min threshold of the reported sleep time (92.6% started and 94.3% ended within the reported sleep time). All L5 windows overlapped with the reported sleep times to some degree: in the case of least overlap, the L5 window started 3 h and 19 min before the reported sleep time. We note that the comparison only covers 54.6% of the total 447 extracted L5 periods.

3.3 L5 features in participant groups

The mean resting time (L5) physiological measures for HR, RR, R-to-R interval, and skin T are compared across participant groups in Figure 6 (top). The average L5 mean HR observed for healthy participants was lower than that of either of the disease groups, and similarly mean R-to-R interval was higher. Additionally, a larger variety of L5 mean skin T was observed for the healthy group.

Selected L5 HRV parameters are similarly presented in Figure 6 (bottom). The frequency-domain features (LFnu, HFnu, and VLF) show some variations in the value distributions across groups. In accordance with the above-mentioned mean R-to-R interval distributions, pNN50 shows most R-to-R intervals exceeding 50 ms in the healthy group.

3.4 Feature aggregate association with PROs

The association analysis between the 2-h aggregated features and the PROs comprised a total of 1,646 (476 for healthy, 253 for NDD, and 917 for IMID group) comparable instances collected from 91 participants. The analysis revealed statistically significant correlations between PROs and several feature aggregates. Figure 7 depicts the correlation r values for each participant group when the feature aggregates were normalized using the L5 participant-mean parameters. The corresponding *p*-values, degrees of freedom, and 95% confidence intervals are presented in Supplementary Figures S1–S3, respectively. For the healthy and IMID patients, the most pronounced correlations are close to ± 0.3, most of them for sleepiness PRO. NDD group shows most of the statistically significant correlations with sleep quality ($|r| = 0.31–0.37$).

Figure 8 displays the correlation r to 2-h feature aggregates (see Supplementary Figures S4–S6 for the *p*-values, degrees of freedom, and 95% confidence intervals, respectively) normalized using the most recent previous L5 window parameters. It includes 1,319 (410, 209, and 700 for the healthy, NDD, and IMID group, respectively) PRO responses coupled with feature aggregates from a total of 84 participants. This is less than above because a normalization window with the set requirements was not always available. In this case, digital measure coverage shows significant correlation with mental daily activities. For the NDD group, the significant correlations are more spread over PROs. In the IMID group, features correlating with daily activity levels emerge.

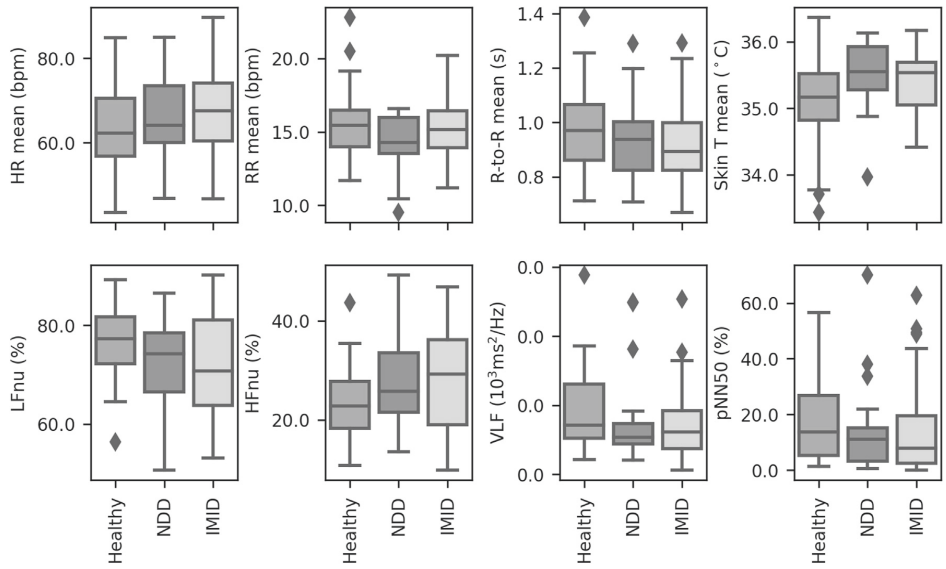

FIGURE 6
The participants' average L5 parameter, presented by participant group, for (top) the mean signal values and (bottom) selected frequency domain HRV features. The figure covers 31 healthy participants, and 21 NDD and 49 IMID patients.

3.5 Heart rate recovery

The full 1,297 days of VitalPatch data were scanned for sequences applicable for heart rate recovery analysis. A total of 274 applicable HRR resting periods were identified, covering 73 distinct participants, as detailed in Table 6. Each participant (among the 73) had 1–16 applicable periods (3.8 on average). HRR by participant group is presented in Figure 9, with the total walk durations of the accepted walks. Only one representative HRR value (the highest) is depicted for each participant.

ANCOVA showed a significant difference with an F statistic of 5.68 ($p < 0.006$) in HRR between participant groups while adjusting for age and gender. The partial η^2 implied a small effect, with 14% of the variance explained by the group (and 10% by age while the effect of gender was not significant). Post hoc analysis indicated that the healthy group differs significantly from both the NDD (T 3.95, $p = 0.001$) and IMID (T 2.51, $p < 0.038$) groups.

Among the 73 participants included in the HRR analysis, 65 had reported PROs during the full study period (3 participants among healthy and NDD patients and 2 among IMID patients had no response). The mean score for physical fatigue was 2.00 in healthy, 2.33 among NDD patients, and 2.39 in IMID patients. HRR's relation to fatigue is explored further in the Supplementary Material S1 (see Supplementary Figures S7, S8). Significant HRR differences between high and low fatigue groups were observed only within the healthy participants.

4 Discussion

Fatigue and sleep disturbances reduce the quality of life and the activities of daily living. Digital measures collected with wearable devices could improve the objectivity and sensitivity of fatigue and sleep assessment, ultimately providing additional support for disease assessment and evaluation of new therapies. Wearable technologies could facilitate continuous monitoring outside the clinical setting without requiring active interaction from the patient. Moreover, digital measures in free-living settings may enable assessment that is more meaningful to the patient's daily living. However, their potential for fatigue assessment have not been extensively studied, especially in the clinical context.

The results presented in this study suggest the feasibility of collecting reasonable quality physiological measures with a wearable biosensor on patients with chronic NDD and IMID diseases, as well as healthy controls. The median coverage was 77%–78% for all digital measures, with minimal variability across different cohorts. The coverage result implies high compliance to using the wearable biosensor. In contrast, only 91 among the

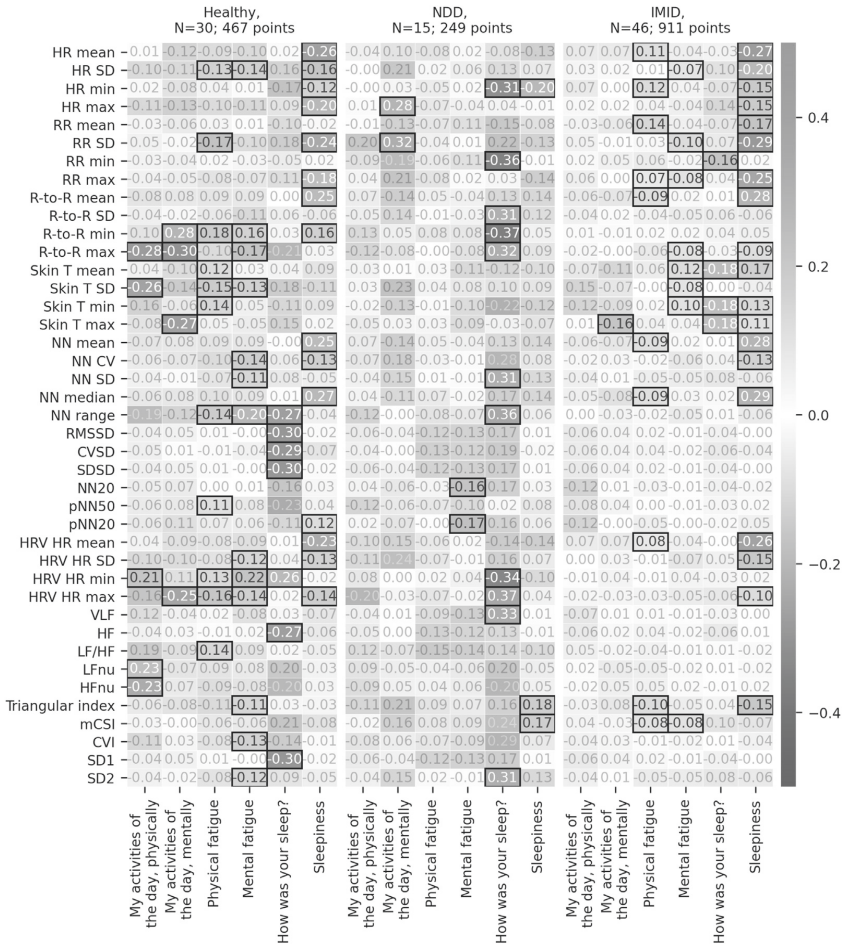

FIGURE 7
Repeated measures correlation r values between the 2-h feature aggregates and the corresponding PROs. Here, the 2-h feature aggregates have been normalized with the participant-mean L5 parameters. Statistically significant correlation results (p-value<0.05) are emphasized with black borders, other r values have faded annotation.

136 participants reported PROs at least three times during the study. Furthermore, in all the collected VitalPatch data, less than 0.5% of HR, RR, and R-to-R interval data and only 2.3% of skin temperature data needed to be cleaned out, indicating a sufficient data quality given the criteria used in this study.

To evaluate the association between the digital physiological measures and fatigue and sleep, we presented results of repeated measures correlation. We selected to evaluate the association between features aggregated over a 2-h window prior to a self-evaluation instance. Thus, the results represent the relationship between the 2-h physiological measures and the PRO at any time of the day. In the morning, the 2-h window may overlap with sleep. To account for the natural person-to-person variability in the digital measures, we further normalized the aggregated features with respect to the participant's average parameters at rest, representing their typical resting state. The L5 windows representing rest were identified utilizing the activity measures available from the same wearable sensor. The L5 periods were reasonably aligned with the self-reported sleep times. While no major differences were observed in the participant-mean

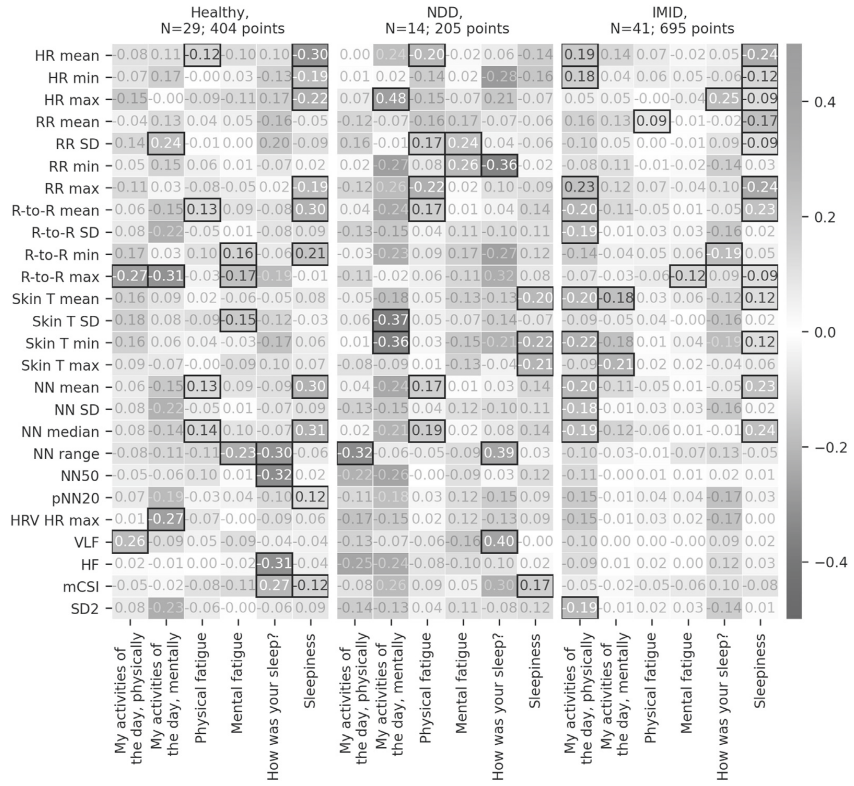

FIGURE 8
Repeated measures correlation r values between the 2-h feature aggregates and the corresponding PROs. Here, the 2-h feature aggregates have been normalized with the participants' latest L5 window parameters. That is, as compared to Figure 7, the normalization parameters are not averaged over the full study periods. Statistically significant correlation results (p-value<0.05) are emphasized with black borders, other r values have faded annotation.

TABLE 6 Number of participants and sequences applicable for HRR analysis, and the median duration of walks leading to the inspected resting period.

Group	Cohort	N	Sequences for HRR analysis	Median walk duration (min)
Healthy	Healthy	21	91	9
NDD	HD	9	21	9
	PD	10	27	8
IMID	IBD	11	32	11
	PSS	8	45	11
	RA	8	36	9
	SLE	6	22	8
Total	*7 cohorts*	*73*	*274*	*9*

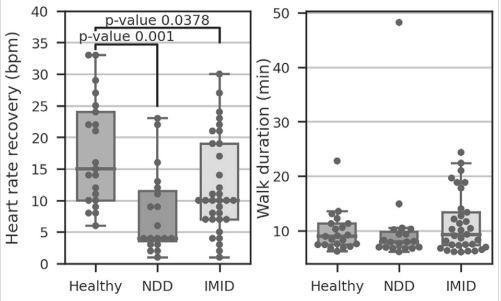

FIGURE 9
Heart rate recovery (left) in the three participant groups, including one representative value for 21 healthy participants, 18 NDD (9 HD, 10 PD), and 33 IMID patients (11 IBD, 8 PSS, 8 RA, 6 SLE). The adjusted *p*-values from post hoc analysis with Tukey's method indicate the significance of the differences. The total duration of each accepted walk leading to a 1-min resting period is presented on the right.

L5 parameters themselves across participant groups, some expected variations appeared. For instance, we observed lower mean HR and higher R-to-R intervals for the healthy, which is consistent with the presumption that increased fatigue is associated with reduced HRV (Escorihuela et al., 2020). Interestingly, the NDD group showed higher skin temperatures than others, with less variation, too. Although the group is small, this observation is in line with study by Eggenberger et al. (2021) where they found that cognitively healthy adults have lower skin temperatures than those with mild cognitive impairment. It is noted that the skin temperature may be affected by ambient temperatures.

The statistically significant correlations between the 2-h feature aggregates and the PROs varied from participant group to another. For NDD patients, most of the significant correlations associated with sleep quality. For the IMID patients, most correlations were found for sleepiness, whereas a reasonable number of correlations were also identified for both physical and mental fatigue. The same is true for the healthy participants, although there is some variance in the specific digital measures that correlate with the PROs.

We also proposed an alternative method for feature normalization, which uses the latest L5 parameters instead of the averaged ones. Using this method revealed correlations for physical and mental fatigue also in the NDD patient group. In the IMID group, significant correlations with the physical activities of the day emerged. This normalization approach may be better able to account for shifts in the daily baseline.

Inspecting the individual feature aggregates in Figures 7, 8 further imply the relevance of the digital measures. HR is relevantly associated with sleepiness, both in the healthy group and IMID patients. Interestingly, this association is not seen in the NDD group, suggesting that neurodegeneration breaks this association, e.g., by affecting the central autonomic nuclei and/or pathways. Significant associations between skin T and the dependent variables in the healthy and the IMID patients, but not in NDD patients in Figure 7, suggest a similar mechanism. These observations may be related to the circadian rhythm abnormalities in NDD patients reported in previous studies (Hood and Amir, 2017). The LF/HF ratio, which in controlled settings reflects the ratio between sympathetic nervous system and parasympathetic nervous system activity, was associated with daytime symptoms in the healthy, but not in the NDD and IMID patients, suggesting an affection of this balance in NDD and IMID in the daytime. It is also noteworthy that in NDD most of the significant results occur between the dependent variables and sleep quality, and in IMID between dependent variables and (daytime) sleepiness, which speaks for different mechanisms of vegetative control between the different types of diseases. Conversely, it is also interesting to observe that sleepiness and mental/physical fatigue obviously represent different concepts and mechanisms, since the distribution of the significances for the respective variables is very different. A detailed analysis of the clinical relevance of the findings is, however, out of the scope of this work and will be left for future research. The clinical implications of free-living heart rate variability details may require further examination (Hayano and Yuda, 2019, Hayano and Yuda, 2021). Since wearable devices often utilize a lower sampling frequency to reduce power consumption and prolong the battery life, careful consideration on the sampling rate should be made during experiment planning. Although prior work has demonstrated reliability and clinical utility of heart rate variability measures quantified from a sampling frequency of 125 Hz (Ellis et al., 2015; Nallathambi et al., 2020; Hirten et al., 2021; Lee et al., 2022), very low variability in R-to-R interval, such as those observed in heart failure patients, may require higher sampling frequencies for sufficient temporal resolution (Kuusela, 2013).

We note that most of the correlations are modest, and a larger group especially of NDD patients is required to validate the presented findings. More advanced features beyond the 2-h statistical aggregators and classical HRV features should be studied in the future to capture more complicated temporal patterns. Additionally, although repeated measure correlation was selected to account for participant-to-participant differences in PRO reporting, the subjectivity and limited sensitivity of the PROs could limit the possibilities to detect associations.

The PRO-association analysis was complemented by an explorative analysis of 1-min HRR during rest, after periods of sustained activity. We discovered that the NDD and IMID patients showed significantly (p = 0.001 and p = 0.0378, respectively) lower HRR values as compared to the healthy controls. This finding is consistent with previous research indicating deteriorated HRR in the HD, PD, IBD, RA, and SLE cohorts as compared to healthy controls after an exercise

test either on a treadmill or on a cycle ergometer (Dogdu et al., 2010; Sarli et al., 2016; Bienias et al., 2017; Pecąnha et al., 2018; Roberson et al., 2018; Steventon et al., 2018). The presented result suggests that the difference may also be observed in the context of daily walking activities using wearable technology in free living participants. While the 6-min walk test in controlled settings has been previously established as a valid test beside the more intense treadmill exercise, our results suggest that useful information can also be extracted from at-home continuous physiological measurements (Roberts et al., 2006). On the other hand, NDD are associated with disruption to blood flow, hypertension, and reduction in cerebral blood flow (Youwakim and Girouard, 2021). These factors may contribute to the variation in the HRR results in comparison to the healthy participants. For SLE, HRR deterioration has been suggested to associate with disease severity (Bienias et al., 2017). More research is required to assess the connection of HRR monitored by wearables to disease severity in the NDD and IMID patients.

The study did not show any significant association between HRR and fatigue in the patient groups, although on average the NDD and IMID patients reported higher fatigue than the healthy participants. However, the participant group correlated significantly to HRR and may act as a confounder whose effect dominates over that of fatigue. In the healthy group, in contrast, a significant difference in HRR was observed between high and low fatigue groups. Furthermore, because of the subjective differences in self-assessment of fatigue, the association between HRR in free-living settings and fatigue should be studied with repeated measures on a subject level, in a study covering a longer study period. A notably longer study period could also enable more advanced analysis, like evaluating the sensitivity of HRR to within-subject changes in fatigue.

Given the multifactorial nature of fatigue, future work will combine physiological measures studied here with multiple sensing modalities. For instance, acceleration signals could be utilized to investigate physiological responses in the context of specific activities, or physiological measures could be combined with the observed sleep stages to further investigate connections with sleep. The IDEA-FAST consortium intends to validate findings of this pilot study using multiple sensing modalities in a larger cohort of patient and healthy participants ($N = 2000$) and over a longer study period. The large-scale nature of this future study will enable further investigation on the sensitivity of HRR and other digital measures to changes in fatigue and sleep.

Data availability statement

The datasets presented in this article are not readily available because of the sensitive nature of the data. The study data will be made available upon request for validation purposes, subject to the signing of a suitable data sharing agreement. Validation can only take place on the IDEA-FAST data platform which supports all necessary statistical software tools. Requests to access the datasets should be directed to https://idea-fast.eu/contact/.

Ethics statement

The studies involving human participants were reviewed and approved by Ethical Committee of the Medical Faculty, Kiel University, Kiel, Germany, Health Research Authority (HRA) and Health and Care Research Wales (HCRW), United Kingdom, Ethik-Kommission der Ärztekammer Westfalen-Lippe und der Westfälischen Wilhelms-Universität, Münster, Germany, and the Medical Ethics Review Committee, Erasmus MC, Netherlands. The patients/participants provided their written informed consent to participate in this study.

Author contributions

First author and analysis lead: EA analysis planning and results interpretation: EA, MC, TA, and MG. Data analysis: EA, HN, MC, DB, and RZR. Co-ordination and conception of study: W-FN, WM, NM, IP, PD, GG, and FB. Design of experiments: RR, CJW, WM, W-FN, IP, TA, and MC. Data collection: JK, VM, KD, HH, KE, RR, CJW, WM, and W-FN. Drafting of manuscript: EA, HN, JK, and MC. Content feedback, manuscript revision and approval: All authors.

Funding

The presented work was funded *via* the IDEA-FAST project, which has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 853981. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and associated partners.

Acknowledgments

The authors thank the IDEA-FAST project consortium members (https://idea-fast.eu/the-idea-fast-investigators/) for their contributions to the project, facilitating this study.

Conflict of interest

MC, IP, and NM are employees of Janssen Research & Development, LLC and may hold company stocks/stock options. FB is a former Janssen Research & Development, LLC employee and a current employee of Bristol Meyers Squibb and holds company stocks/stock options in both. EA and TA were employed by the VTT

Technical Research Centre of Finland Ltd., HN was employed by Sanofi and GVG was employed by Takeda Belgium.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2022.968185/full#supplementary-material

References

Altini, M., and Kinnunen, H. (2021). The promise of sleep: A multi-sensor approach for accurate sleep stage detection using the oura ring. *Sensors* 21, 4302. *Page 4302* 21. doi:10.3390/S21134302

Altini, M., Penders, J., Vullers, R., and Amft, O. (2014). Personalizing energy expenditure estimation using physiological signals normalization during activities of daily living. *Physiol. Meas.* 35, 1797–1811. doi:10.1088/0967-3334/35/9/1797

Areia, C. M., Santos, M., Vollam, S., Pimentel, M., Young, L., Roman, C., et al. (2021). A chest patch for continuous vital sign monitoring: Clinical validation study during movement and controlled hypoxia. *J. Med. Internet Res.* 23 (9), e27547. doi:10.2196/27547.2196/27547

Aziz, N. A., Anguelova, G. V., Marinus, J., Lammers, G. J., and Roos, R. A. C. (2010). Sleep and circadian rhythm alterations correlate with depression and cognitive impairment in Huntington's disease. *Park. Relat. Disord.* 16, 345–350. doi:10.1016/J.PARKRELDIS.2010.02.009

Bakdash, J. Z., and Marusich, L. R. (2017). Repeated measures correlation. *Front. Psychol.* 8, 456. doi:10.3389/FPSYG.2017.00456

Bangerter, A., Chatterjee, M., Manfredonia, J., Manyakov, N. V., Ness, S., Boice, M. A., et al. (2020a2020). Automated recognition of spontaneous facial expression in individuals with autism spectrum disorder: Parsing response variability. *Mol. Autism* 11, 31–15. doi:10.1186/S13229-020-00327-4

Bangerter, A., Chatterjee, M., Manyakov, N. V., Ness, S., Lewin, D., Skalkin, A., et al. (2020b). Relationship between sleep and behavior in autism spectrum disorder: Exploring the impact of sleep variability. *Front. Neurosci.* 14, 211. doi:10.3389/FNINS.2020.00211

Bellet, R. N., Adams, L., and Morris, N. R. (2012). The 6-minute walk test in outpatient cardiac rehabilitation: Validity, reliability and responsiveness—a systematic review. *Physiotherapy* 98, 277–286. doi:10.1016/J.PHYSIO.2011.11.003

Bienias, P., Ciurzyński, M., Chrzanowska, A., Dudzik-Niewiadomska, I., Irzyk, K., Oleszek, K., et al. (2017). Attenuated post-exercise heart rate recovery in patients with systemic lupus erythematosus: The role of disease severity and beta-blocker treatment. *Lupus* 27, 217–224. doi:10.1177/0961203317716318

Boissoneault, J., Letzen, J., Robinson, M., and Staud, R. (2019). Cerebral blood flow and heart rate variability predict fatigue severity in patients with chronic fatigue syndrome. *Brain Imaging Behav.* 13, 789–797. doi:10.1007/S11682-018-9897-X

Center for Disease Control and Prevention (2000). *Measuring healthy days population assessment of health-related quality of life*. Atlanta, Georgia: Center for Disease Control and Prevention.

Champseix, R. (2021). Aura-healthcare/hrv-analysis: Package for heart rate variability analysis in Python. Available at: https://github.com/Aura-healthcare/hrv-analysis [Accessed January 21, 2022].

Chavarría, C., Casanova, M. J., Chaparro, M., Barreiro-De Acosta, M., Ezquiaga, E., Bujanda, L., et al. (2019). Prevalence and factors associated with fatigue in patients with inflammatory Bowel disease: A multicentre study. *J. Crohns Colitis* 13, 996–1002. doi:10.1093/ecco-jcc/jjz024

Chen, L., Ma, X., Chatterjee, M., Kortelainen, J. M., Ahmaniemi, T., Maetzler, W., et al. (2022). Fatigue and sleep assessment using digital sleep trackers: Insights from a multi-device pilot study. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2022, 1133–1136. doi:10.1109/EMBC48229.2022.9870923

Cretikos, M. A., Bellomo, R., Hillman, K., Chen, J., Finfer, S., and Flabouris, A. (2008). Respiratory rate: The neglected vital sign. *Med. J. Aust.* 188, 657–659. doi:10.5694/J.1326-5377.2008.TB01825.X

Davies, K., Dures, E., and Ng, W. F. (2021). Fatigue in inflammatory rheumatic diseases: Current knowledge and areas for future research. *Nat. Rev. Rheumatol.* 17, 651–664. doi:10.1038/s41584-021-00692-1

Del Din, S., Godfrey, A., Galna, B., Lord, S., and Rochester, L. (2016). Free-living gait characteristics in ageing and Parkinson's disease: Impact of environment and ambulatory bout length. *J. Neuroeng. Rehabil.* 13, 46–12. doi:10.1186/S12984-016-0154-5

Djaoui, L., Haddad, M., Chamari, K., and Dellal, A. (2017). Monitoring training load and fatigue in soccer players with physiological markers. *Physiol. Behav.* 181, 86–94. doi:10.1016/J.PHYSBEH.2017.09.004

Dogdu, O., Yarlioglues, M., Kaya, M. G., Ardic, I., Oguzhan, N., Akpek, M., et al. (2010). Deterioration of heart rate recovery index in patients with systemic lupus erythematosus. *J. Rheumatol.* 37, 2511–2515. doi:10.3899/JRHEUM.100163

Eggenberger, P., Bürgisser, M., Rossi, R. M., and Annaheim, S. (2021). Body temperature is associated with cognitive performance in older adults with and without mild cognitive impairment: A cross-sectional analysis. *Front. Aging Neurosci.* 13, 585904. doi:10.3389/FNAGI.2021.585904

Ellis, R. J., Zhu, B., Koenig, J., Thayer, J. F., and Wang, Y. (2015). A careful look at ECG sampling frequency and R-peak interpolation on short-term measures of heart rate variability. *Physiol. Meas.* 36, 1827–1852. doi:10.1088/0967-3334/36/9/1827

Escorihuela, R. M., Capdevila, L., Castro, J. R., Zaragozà, M. C., Maurel, S., Alegre, J., et al. (2020). Reduced heart rate variability predicts fatigue severity in individuals with chronic fatigue syndrome/myalgic encephalomyelitis. *J. Transl. Med.* 18, 4–12. doi:10.1186/S12967-019-02184-Z

Garavaglia, L., Gulich, D., Defeo, M. M., Mailland, J. T., and Irurzun, I. M. (2021). The effect of age on the heart rate variability of healthy subjects. *PLoS One* 16, e0255894. doi:10.1371/JOURNAL.PONE.0255894

Gutierrez, G., Williams, J., Alrehaili, G. A., McLean, A., Pirouz, R., Amdur, R., et al. (2016). Respiratory rate variability in sleeping adults without obstructive sleep apnea. *Physiol. Rep.* 4, e12949. doi:10.14814/PHY2.12949

Hayano, J., and Yuda, E. (2021). Assessment of autonomic function by long-term heart rate variability: Beyond the classical framework of LF and HF measurements. *J. Physiol. Anthropol.* 40, 21–15. doi:10.1186/S40101-021-00272-Y

Hayano, J., and Yuda, E. (2019). Pitfalls of assessment of autonomic function by heart rate variability. *J. Physiol. Anthropol.* 38, 3–8. doi:10.1186/s40101-019-0193-2

Hewlett, S., Chalder, T., Choy, E., Cramp, F., Davis, B., Dures, E., et al. (2011). Fatigue in rheumatoid arthritis: Time for a conceptual model. *Rheumatol. Oxf.* 50, 1004–1006. doi:10.1093/rheumatology/keq282

Hirten, R. P., Danieletto, M., Scheel, R., Shervey, M., Ji, J., Hu, L., et al. (2021). Longitudinal autonomic nervous system measures correlate with stress and ulcerative colitis disease activity and predict flare. *Inflamm. Bowel Dis.* 27, 1576–1584. doi:10.1093/IBD/IZAA323

Hood, S., and Amir, S. (2017). Neurodegeneration and the circadian clock. *Front. Aging Neurosci.* 9, 170. doi:10.3389/FNAGI.2017.00170

Karlsson, M., Hörnsten, R., Rydberg, A., and Wiklund, U. (2012). Automatic filtering of outliers in RR intervals before analysis of heart rate variability in holter recordings: A comparison with carefully edited data. *Biomed. Eng. Online* 11, 2. doi:10.1186/1475-925X-11-2

Kluger, B. M., Krupp, L. B., and Enoka, R. M. (2013). Fatigue and fatigability in neurologic illnesses: Proposal for a unified taxonomy. *Neurology* 80, 409–416. doi:10.1212/WNL.0B013E31827F07BE

Kuusela, T. (2013). "Methodological aspects of heart rate variability analysis," in *Heart rate variability (HRV) signal analysis*. Editors M. V. Kamath, M. Watanabe, and A. Upton (Boca Raton, FL: CRC Press), 9–42.

Lamberts, R. P., Swart, J., Capostagno, B., Noakes, T. D., and Lambert, M. I. (2009). Heart rate recovery as a guide to monitor fatigue and predict changes in performance parameters. *Scand. J. Med. Sci. Sports* 20, 449–457. doi:10.1111/j.1600-0838.2009.00977.x

Lee, K. F. A., Chan, E., Car, J., Gan, W.-S., Christopoulos, G., Fye, K., et al. (2022). Lowering the sampling rate: Heart rate response during cognitive fatigue. *Biosensors* 202212, 315. doi:10.3390/BIOS12050315

Legge, A., Blanchard, C., and Hanly, J. G. (2017). Physical activity and sedentary behavior in patients with systemic lupus erythematosus and rheumatoid arthritis. *Open Access Rheumatol.* 9, 191–200. doi:10.2147/OARRR.S148376

Lendrem, D., Mitchell, S., McMeekin, P., Bowman, S., Price, E., T Pease, C., et al. (2014). Health-related utility values of patients with primary Sjögren's syndrome and its predictors. *Ann. Rheum. Dis.* 73, 1362–1368. doi:10.1136/annrheumdis-2012-202863

Luo, H., Lee, P. A., Clay, I., Jaggi, M., and De Luca, V. (2020). Assessment of fatigue using wearable sensors: A pilot study. *Digit. Biomark.* 4, 59–72. doi:10.1159/000512166

Malik, M., John Camm, A., Thomas Bigger, J., Breithardt, G., Cerutti, S., Cohen, R. J., et al. (1996). Heart rate variability. *Circulation* 93, 1043–1065. doi:10.1161/01.CIR.93.5.1043

Martinez-Nicolas, A., Meyer, M., Hunkler, S., Madrid, J. A., Rol, M. A., Meyer, A. H., et al. (2015). Daytime variation in ambient temperature affects skin temperatures and blood pressure: Ambulatory winter/summer comparison in healthy young women. *Physiol. Behav.* 149, 203–211. doi:10.1016/J.PHYSBEH.2015.06.014

Nallathambi, G., Selvaraj, N., and Rajbhandary, P. L. (2020). An innovative hybrid approach for detection of pacemaker pulses at low sampling frequency. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2020, 5012–5015. doi:10.1109/EMBC44109.2020.9176390

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi:10.1111/J.1532-5415.2005.53221.X

Nicolò, A., Marcora, S. M., Bazzucchi, I., and Sacchetti, M. (2017). Differential control of respiratory frequency and tidal volume during high-intensity interval training. *Exp. Physiol.* 102, 934–949. doi:10.1113/EP086352

Nicolò, A., Massaroni, C., Schena, E., and Sacchetti, M. (2020). The importance of respiratory rate monitoring: From healthcare to sport and exercise. *Sensors* 202020, 6396. *Page 6396* 20. doi:10.3390/S20216396

Pecąnha, T., Rodrigues, R., Pinto, A. J., Sá-Pinto, A. L., Guedes, L., Bonfiglioli, K., et al. (2018). Chronotropic incompetence and reduced heart rate recovery in rheumatoid arthritis. *J. Clin. Rheumatol.* 24, 375–380. doi:10.1097/RHU.0000000000000745

Peltola, M. A. (2012). Role of editing of R-R intervals in the analysis of heart rate variability. *Front. Physiol.* 3, 148. doi:10.3389/FPHYS.2012.00148

Qiu, S., Cai, X., Sun, Z., Li, L., Zuegel, M., Steinacker, J. M., et al. (2017). Heart rate recovery and risk of cardiovascular events and all-cause mortality: A meta-analysis of prospective cohort studies. *J. Am. Heart Assoc.* 6, e005505. doi:10.1161/JAHA.117.005505

Rajbhandary, P. L., and Nallathambi, G. (2020). Feasibility of continuous monitoring of core body temperature using chest-worn patch sensor. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2020, 4652–4655. doi:10.1109/EMBC44109.2020.9175579

Roberson, K. B., Signorile, J. F., Singer, C., Jacobs, K. A., Eltoukhy, M., Ruta, N., et al. (2018). Hemodynamic responses to an exercise stress test in Parkinson's disease patients without orthostatic hypotension. *Appl. Physiology, Nutr. Metabolism* 44, 751–758. doi:10.1139/APNM-2018-0638

Roberts, E., Li, F. K. W., and Sykes, K. (2006). Validity of the 6-minute walk test for assessing heart rate recovery after an exercise-based cardiac rehabilitation programme. *Physiotherapy* 92, 116–121. doi:10.1016/J.PHYSIO.2005.06.005

Sarli, B., Dogan, Y., Poyrazoglu, O., Baktir, A. O., Eyvaz, A., Altinkaya, E., et al. (2016). Heart rate recovery is impaired in patients with inflammatory Bowel diseases. *Med. Princ. Pract.* 25, 363–367. doi:10.1159/000446318

Selvaraj, N., Nallathambi, G., Moghadam, R., and Aga, A. (2018). Fully disposable wireless patch sensor for continuous remote patient monitoring. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2018, 1632–1635. doi:10.1109/EMBC.2018.8512569

Shaffer, F., and Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Front. Public Health* 5, 258. doi:10.3389/FPUBH.2017.00258

Siciliano, M., Trojano, L., Santangelo, G., De Micco, R., Tedeschi, G., and Tessitore, A. (2018). Fatigue in Parkinson's disease: A systematic review and meta-analysis. *Mov. Disord.* 33, 1712–1723. doi:10.1002/mds.27461

Sokas, D., Paliakaitė, B., Rapalis, A., Marozas, V., Bailón, R., and Petrėnas, A. (2021). Detection of walk tests in free-living activities using a wrist-worn device. *Front. Physiol.* 12, 706545. doi:10.3389/FPHYS.2021.706545

Somers, V. K., Dyken, M. E., Mark, A. L., and Abboud, F. M. (1993). Sympathetic-nerve activity during sleep in normal subjects. *N. Engl. J. Med.* 328, 303–307. doi:10.1056/NEJM199302043280502

Stefani, A., and Högl, B. (2019). Sleep in Parkinson's disease. *Neuropsychopharmacology* 45, 121–128. doi:10.1038/s41386-019-0448-y

Steventon, J. J., Collett, J., Furby, H., Hamana, K., Foster, C., O'Callaghan, P., et al. (2018). Alterations in the metabolic and cardiorespiratory response to exercise in Huntington's Disease. *Park. Relat. Disord.* 54, 56–61. doi:10.1016/J.PARKRELDIS.2018.04.014

Stone, A. A., Shiffman, S., Schwartz, J. E., Broderick, J. E., and Hufford, M. R. (2002). Patient non-compliance with paper diaries. *Br. Med. J.* 324, 1193–1194. doi:10.1136/bmj.324.7347.1193

Swanson, G. R., and Burgess, H. J. (2017). Sleep and circadian hygiene and inflammatory Bowel disease. *Gastroenterol. Clin. North Am.* 46, 881–893. doi:10.1016/J.GTC.2017.08.014

Tanaka, H., Monahan, K. D., and Seals, D. R. (2001). Age-predicted maximal heart rate revisited. *J. Am. Coll. Cardiol.* 37, 153–156. doi:10.1016/S0735-1097(00)01054-8

The IDEA-FAST project consortium (2020). *D2.1: First study subject approvals package of the Feasibility Study (FS)*. Grant Agreement No. 853981.

Vildjiounaite, E., Kallio, J., Kyllönen, V., Nieminen, M., Määttänen, I., Lindholm, M., et al. (2018). Unobtrusive stress detection on the basis of smartphone usage data. *Pers. Ubiquitous Comput.* 22, 671–688. doi:10.1007/S00779-017-1108-Z

Voss, A., Schroeder, R., Heitmann, A., Peters, A., and Perz, S. (2015). Short-Term heart rate variability—influence of gender and age in healthy subjects. *PLoS One* 10, e0118308. doi:10.1371/JOURNAL.PONE.0118308

Wijsman, J., Grundlehner, B., Liu, H., Hermens, H., and Penders, J. (2011). "Towards mental stress detection using wearable physiological sensors," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Boston, MA, USA, 30 August 2011 - 03 September 2011. doi:10.1109/IEMBS.2011.6090512

Witting, W., Kwa, I. H., Eikelenboom, P., Mirmiran, M., and Swaab, D. F. (1990). Alterations in the circadian rest-activity rhythm in aging and Alzheimer's disease. *Biol. Psychiatry* 27, 563–572. doi:10.1016/0006-3223(90)90523-5

Xiao, M., Yan, H., Song, J., Yang, Y., and Yang, X. (2013). Sleep stages classification based on heart rate variability and random forest. *Biomed. Signal Process. Control* 8, 624–633. doi:10.1016/J.BSPC.2013.06.001

Youwakim, J., and Girouard, H. (2021). Inflammation: A mediator between hypertension and neurodegenerative diseases. *Am. J. Hypertens.* 34, 1014–1030. doi:10.1093/AJH/HPAB094

Zhai, B., Perez-Pozuelo, I., Clifton, E. A. D., Palotti, J., and Guan, Y. (2020). Making sense of sleep. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1–33. doi:10.1145/3397325

Zhang, S., Sun, J., and Gao, X. (2020). The effect of fatigue on brain connectivity networks. *Brain Sci. Adv.* 6, 120–131. doi:10.26599/bsa.2020.9050008

Zielinski, M. R., Systrom, D. M., and Rose, N. R. (2019). Fatigue, sleep, and autoimmune and related disorders. *Front. Immunol.* 10, 1827–1926. doi:10.3389/fimmu.2019.01827

PUBLICATION III

Predicting daytime sleepiness from electrocardiography based respiratory rate using deep learning

E. Antikainen, R. Z. U. Rehman, T. Ahmaniemi, and M. Chatterjee

In *2022 Computing in Cardiology (CinC)*, 2022
DOI: 10.22489/CinC.2022.100

Predicting Daytime Sleepiness from Electrocardiography Based Respiratory Rate Using Deep Learning

Emmi Antikainen[1], Rana Zia Ur Rehman[2], Teemu Ahmaniemi[1], Meenakshi Chatterjee[3]

[1] VTT Technical Research Centre of Finland Ltd., Tampere, Finland
[2] Newcastle University, Newcastle Upon Tyne, United Kingdom
[3] Janssen Research & Development, Cambridge, MA, USA

Abstract

Daytime sleepiness impairs the activities of daily living, especially in chronic disease patients. Typically, daytime sleepiness is measured with subjective patient reported outcomes (PROs), which could be prone to recall bias. Objective measures of daytime sleepiness, which are sensitive to change, would benefit the assessment of disease states and novel therapies that impact the quality of life. The presented study aimed to predict daytime sleepiness from two hours of continuously measured respiratory rate using a 1-dimensional convolutional neural network. A wearable biosensor was used to continuously measure electrocardiography (ECG) based respiratory rate, while the participants (N=82) were asked to fill in Karolinska Sleepiness Scale three times a day. Considering the need for a sleepiness measure for chronic diseases, neurodegenerative disease (NDD, N=14) patients, immune-mediated inflammatory disease (IMID, N=42) patients, as well as healthy participants (N=26) were included in the study. The disease-agnostic model achieved an accuracy of 63% between non-sleepy and sleepy states. The result demonstrates the potential of using respiratory rate with deep learning for an objective measure of daytime sleepiness.

1. Introduction

Chronic disease patients commonly experience sleep disturbances and fatigue, which contribute to daytime sleepiness. Daytime sleepiness, in return, interferes with the activities of daily living, ultimately deteriorating the quality of life [1]. It affects cognitive functionalities, increasing the risk of falls resulting in injuries and increased healthcare costs [1, 2]. For instance, among Parkinson's Disease (PD) patients, over 35 % experience excessive daytime sleepiness [3]. Objective measurement of daytime sleepiness is important for both assessing new therapies and evaluating the effect of interventions.

Currently, daytime sleepiness is assessed with subjective patient reported outcomes (PROs), such as the Karolinska Sleepiness Scale (KSS) or the Epworth Sleepiness Scale. However, such subjective measures suffer from recall bias [4]. Objective measures of the physiological signs of sleepiness could provide better accuracy, reliability, and continuous assessment. Electrocardiography (ECG) based wearable sensors can facilitate the continuous monitoring of chronic disease patients in free-living settings and may capture how the disease affects the patient's daily-living.

Previous studies on patients' daytime sleepiness prediction have utilized clinical data or laboratory measurements [5, 6]. To our knowledge, sleepiness prediction with wearable sensors in free-living settings has not been studied extensively: Igasaki et al. predicted sleepiness from respiratory signals with support vector machines during a simulated drive, achieving an 89 % accuracy, whereas Bao et al. used wearable body temperature sensing to assess sleepiness over two days [7, 8]. Both studies only included a small sample (6-7) of healthy adults measured for a short time in simulated or restricted free-living settings.

One intuitive manifestation of daytime sleepiness in respiration is yawning. Previous studies have established that yawning frequency increases with sleepiness [9, 10]. This study uses deep learning (DL) to predict patient reported daytime sleepiness from continuously measured respiratory rate, which is often readily measured by modern wearable sensors and may offer an easily accessible continuous measure for sleepiness. The proposed disease-agnostic model uses a 1-dimensional convolutional neural network (1D CNN) and builds on a longitudinal multi-site data set, covering several days of respiratory rate and KSS responses (three times a day) collected from 82 volunteers, including neurodegenerative disease (NDD) patients (N=14), immune-mediated inflammatory disease (IMID) patients (N=42), and healthy participants (N=26).

2. Material and Methods

The study is based on the data collected in the IDEA-FAST feasibility study [11, 12].

2.1. Study participants

The study data comprised a total of 82 volunteered adults, including NDD patients (N=14), IMID patients (N=42), and healthy volunteers (N=26). The NDD group comprised patients with PD (N=8) and Huntington's Disease (HD, N=6). The IMID group included Inflammatory Bowel Disease (IBD, N=10), Primary Sjögren's Syndrome (PSS, N=13), Rheumatoid Arthritis (RA, N=7), and Systemic Lupus Erythematosus (SLE, N=12).

The participants were recruited at four sites, and the ethical approvals were granted (in June to November 2020) by the research ethics committees of each site: the ethical committee of the Medical Faculty of Kiel University (K) (D491/20), Newcastle upon Tyne Hospitals National Health Service Foundation/Newcastle University (N), Erasmus University Medical Centre in Rotterdam (E), and George-Huntington-Institute in Muenster (G). The study was registered in the German Clinical Trial Registry under DRKS00021693.

2.2. Study protocol

The participants wore a patch sensor, VitalPatch, on their chest [13]. It adheres to the skin, and its battery lasts up to seven days. VitalPatch uses a single lead ECG (and partially a tri-axial accelerometer) to derive respiratory rate readings at 0.25 Hz. It is a class IIa medical device with FDA clearance. The participants wore the biosensor for five consecutive days at a time in free-living settings. The wear period was repeated up to four times during their enrollment and was always followed by at least two rest days.

The PRO for daytime sleepiness was collected using the KSS, which was prompted three times a day (at 13:00, 17:00, and 21:00 local time) via a smartphone application, the VTT Stress Monitor App [14]. The KSS was available for a response for 3 hours in the early and late afternoon and 2.5 hours in the evening. The response was selected from a drop-down menu list of ten options, ranging from "extremely alert" to "extremely sleepy".

2.3. Data pre-processing

The respiratory rate data were pre-processed by sorting the timestamps into monotonically increasing order while removing duplicates and by removing (a) manufacturer-defined invalid values, (b) values beyond the range of 4–60 breaths per minute (bpm), and (c) contextual outliers [15]. For (c), each value was compared to the mean of the surrounding ± 1.5 minutes of data and removed if the inspected value differed from the mean by more than 50%.

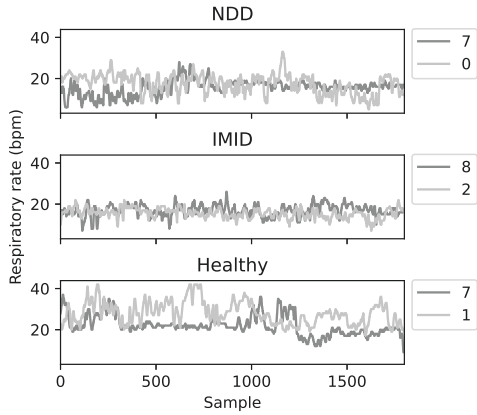

Figure 1. Samples of 2h respiratory rate signals associated with sleepy (blue) and non-sleepy (orange) states. A row shows two samples from an individual randomly chosen from the NDD (top), IMID (middle), and healthy (bottom) groups. The KSS levels (0–9) are shown in the legend.

2.4. Deep learning approach

A 1D CNN was employed to learn respiratory rate patterns and classify the samples into two target classes: (1) non-sleepy and (2) sleepy. The non-sleepy class was defined as KSS levels from "extremely alert" to "rather alert" (indexed 0 to 3), whereas the sleepy class was represented by KSS levels from "neither alert nor sleepy" to "extremely sleepy" (indexed 4 to 9).

The respiratory rates over the 2 hours preceding a KSS response was selected as the prediction input. Thus, the input samples were 1800-value time series. Any missing values were padded with zeros; however, a respiratory rate coverage of at least 90% was required in the 2h window. The coverage was evaluated as the number of observations compared to the observations expected per the sampling frequency. Additionally, for each participant, at least six eligible 2h windows with a corresponding KSS score were required, to capture variation within subject. Figure 1 depicts some eligible samples coupled with the KSS scores.

The 1D CNN model was built from two convolutional layers coupled with max pooling and dense layers. Rectified linear units were used for activation. The training set samples' average respiratory rate and standard deviation were used to standardize the prediction input. The final layers consisted of a dropout layer and a dense layer with softmax activation. The weighted sparse categorical cross-entropy loss function was used together with the Adam optimizer. Model performance was measured via classifica-

tion accuracy, sensitivity, and specificity.

The dataset was grouped by the participant and split randomly into training and test sets, with 20 % of the subjects held out for testing. The training set was further split by 5-fold cross-validation (CV), and the cross-validation was utilized in hyperparameter selection. The final model was trained on the full training data and tested on the held-out 20 % test set. During training, 10% of training data were used to estimate validation metrics and over-fitting.

3. Results

Table 1 shows the participants' demographics. The mean age was 50.9 (±16.1) years, time since diagnosis 11.1 (±9.2) years, and body-mass index 23.8 (±7.2) kg/m^2.

Each participant wore the patch-like sensor for 3–12 days while responding to KSS questionnaires. The total number of 2h respiratory rate samples coupled with a patient reported sleepiness score was 1255, comprising 187 samples for NDD patients, 708 for IMID patients, and 360 for the healthy group. Whilst a minimum number of 6 samples was required, the mean number of obtained samples was 15 (maximum was 30).

The architecture of the 1D CNN is summarized in Figure 2. The hyperparameter optimization yielded a batch size of 30 samples, a learning rate of 0.0001, 50 % dropout rate, and kernel size 5 in the first and 3 in the second convolutional layer. Additionally, the second convolutional layer included an L2 regularization factor of 0.1.

In the cross-validation, the 1D CNN model achieved an average accuracy of 58.3%, sensitivity of 62.5%, and specificity of 52.6%, as detailed in Table 2. Over-fitting was monitored via the training and validation loss curves, and the training was limited to 25 epochs.

The final model achieved 62.6 % accuracy, 57.2 % sensitivity, and 69.2 % specificity in a held-out test set. Based on observations during CV, early stopping was applied after the training and validation loss diverged beyond 0.05 from each other for three consecutive epochs. The test set comprised 10 IMID (5 SLE, 3 RA, 2 PSS) and 2 NDD (both PD) patients and 5 healthy participants, with 7–30 samples per participant. On average, 52% of the participant's samples represented the sleepy class.

Table 1. Participant demographics by participant group.

	NDD	IMID	Healthy
Study sites	K, G	E, K, N	E, G, K, N
Number	14	42	26
Female	7	34	12
Male	7	8	14
Age (mean±SD)	53.6 (±12.6)	53.1 (±16.0)	46.1 (±17.4)
Years since diagnosis (mean±SD)	4.3 (±2.5)	13.5 (±9.4)	–
Body-mass index (mean±SD)	24.0 (±2.4)	22.5 (±9.2)	25.7 (±4.4)

Table 2. Cross-validation results.

	Accuracy	Sensitivity	Specificity
1	0.5226	0.6484	0.4167
2	0.6061	0.6458	0.5686
3	0.5707	0.4884	0.6334
4	0.6396	0.7344	0.4638
5	0.5758	0.6061	0.5455
Average	0.5829	0.6246	0.5257

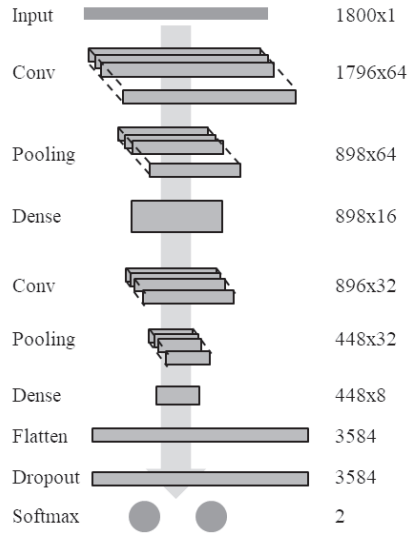

Figure 2. The model architecture comprises 10,426 trainable parameters.

4. Discussion

This study presented a 1D CNN model using ECG-based respiratory rate data to predict daytime sleepiness at the end of a 2-hour monitoring sequence. The final model achieved a 63% accuracy between non-sleepy and sleepy states, together with 57% sensitivity and 69% specificity. The training data included participants from six chronic disease cohorts and healthy participants, capturing several days and several times of the day, from afternoon to evening. We note that notable variations in the performance metrics were revealed in cross-validation within the

training set. However, the average accuracy of 58% was reasonably close to the final test accuracy. Overall, our results suggest that respiratory rate may have potential as a disease-independent predictor of daytime sleepiness.

A well performing objective digital measure can be useful for the clinical assessment of daytime sleepiness in chronic patients. Continuous measures may capture long-term temporal trends more easily than a PRO and enable assessing patient state in the free-living environment, comprehensively describing the effect of sleepiness on the patient's day-to-day quality of life. However, the PROs act as a reference in prediction modelling. This complicates the development of a model that can generalize to new subjects since the scoring may differ significantly from person to person due to individuals' subjective experiences. Thus, personalized models may achieve improved results.

Future studies may explore more specific respiratory patterns from the respiratory signal, and combine with other modalities, e.g, activity measures. Moreover, previous studies have reported that age can affect yawning frequencies [10]. We note that the presented study did not incorporate any demographic information in the model. Personalized prediction models may overcome this and perform better for individual patients. Future studies should eventually focus on developing more complex models that can indicate the level of sleepiness, especially capturing clinically relevant changes in daytime sleepiness.

Acknowledgments

The study was funded via the IDEA-FAST project, which has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 853981. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and associated partners. The authors express their gratitude to the IDEA-FAST project members (https://idea-fast.eu/the-idea-fast-investigators/) for their work facilitating the presented study.

References

[1] Colten HR, Altevogt B (eds.). Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem. National Academies Press (US), 2006.

[2] Ohayon MM, Vecchierini MF. Daytime Sleepiness and Cognitive Impairment in the Elderly Population. Archives of Internal Medicine 2002;162(2):201–208. Doi: 10.1001/archinte.162.2.201.

[3] Feng F, Cai Y, Hou Y, Ou R, Jiang Z, Shang H. Excessive daytime sleepiness in parkinson's disease: A systematic review and meta-analysis. Parkinsonism Related Disorders 2021;85:133–140. Doi: 10.1016/j.parkreldis.2021.02.016.

[4] Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient non-compliance with paper diaries. BMJ May 2002;324:1193–1194. Doi: 10.1136/bmj.324.7347.1193.

[5] Laberge L, Gallais B, Auclair J, Dauvilliers Y, Mathieu J, Gagnon C. Predicting daytime sleepiness and fatigue: a 9-year prospective study in myotonic dystrophy type 1. Journal of Neurology 2020;267:461–468.

[6] Chervin RD, Burns JW, Ruzicka DL. Electroencephalographic changes during respiratory cycles predict sleepiness in sleep apnea. American Journal of Respiratory and Critical Care Medicine 2004;171(6).

[7] Igasaki T, Nagasawa K, Akbar IA, Kubo N. Sleepiness classification by thoracic respiration using support vector machine. In 2016 9th Biomedical Engineering International Conference (BMEiCON). 2016; 1–5. Doi: 10.1109/BMEiCON.2016.7859630.

[8] Bao J, Han J, Kato A, Kunze K. Sleepy watch: Towards predicting daytime sleepiness based on body temperature. In Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers, UbiComp-ISWC '20. 2020; 9–12. Doi: 10.1145/3410530.3414415.

[9] Guggisberg AG, Mathis J, Schnider A, Hess CW. Why do we yawn? Neuroscience Biobehavioral Reviews 2010; 34(8):1267–1276. Doi: 10.1016/j.neubiorev.2010.03.008.

[10] Zilli I, Giganti F, Uga V. Yawning and subjective sleepiness in the elderly. Journal of Sleep Research 2008;17(3):303–308. Doi: 10.1111/j.1365-2869.2008.00666.x.

[11] The IDEA-FAST project consortium. D2.1: First study subject approvals package of the feasibility study (FS), 2020. https://idea-fast.eu/results-and-publications/. Accessed: 2022-08-20.

[12] Chen L, Ma X, Chatterjee M, Kortelainen JM, Ahmaniemi T, Maetzler W, Wang P, Zhang D. Fatigue and sleep assessment using digital sleep trackers: Insights from a multi-device pilot study. In 2022 44th Annual International Conference of the IEEE Engineering and Medicine Society (EMBC). 2022; Doi: 10.1109/EMBC48229.2022.9870923.

[13] Morgado Areia C, Santos M, Vollam S, Pimentel M, Young L, Roman C, Ede J, Piper P, King E, Gustafson O, Harford M, Shah A, Tarassenko L, Watkinson P. A chest patch for continuous vital sign monitoring: Clinical validation study during movement and controlled hypoxia. J Med Internet Res 2021;23(9). Doi: 10.2196/27547.

[14] Vildjiounaite E, Kallio J, Kyllönen V, Nieminen M, Määttänen I, Lindholm M, Mäntyjärvi J, Gimel'farb G. Unobtrusive stress detection on the basis of smartphone usage data. Personal and Ubiquitous Computing 2018;22:671–688. Doi: 10.1007/s00779-017-1108-z.

[15] Nicolò A, Massaroni C, Schena E, Sacchetti M. The importance of respiratory rate monitoring: From healthcare to sport and exercise. Sensors 2020;20(21). Doi: 10.3390/s20216396.

Address for correspondence:

Emmi Antikainen

Visiokatu 4, P.O. Box 1300, 33101 Tampere, Finland

emmi.antikainen@gmail.com

PUBLICATION IV

Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records

E. Antikainen, J. Linnosmaa, A. Umer, N. Oksala, M. Eskola, M. van Gils, J. Hernesniemi, and M. Gabbouj

Scientific Reports, vol. 13, no. 3517
DOI: 10.1038/s41598-023-30657-1

scientific reports

Check for updates

OPEN

Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records

Emmi Antikainen[1]✉, Joonas Linnosmaa[1], Adil Umer[1], Niku Oksala[2,3,4], Markku Eskola[3,5], Mark van Gils[2], Jussi Hernesniemi[2,3,5,7] & Moncef Gabbouj[6,7]

With over 17 million annual deaths, cardiovascular diseases (CVDs) dominate the cause of death statistics. CVDs can deteriorate the quality of life drastically and even cause sudden death, all the while inducing massive healthcare costs. This work studied state-of-the-art deep learning techniques to predict increased risk of death in CVD patients, building on the electronic health records (EHR) of over 23,000 cardiac patients. Taking into account the usefulness of the prediction for chronic disease patients, a prediction period of six months was selected. Two major transformer models that rely on learning bidirectional dependencies in sequential data, BERT and XLNet, were trained and compared. To our knowledge, the presented work is the first to apply XLNet on EHR data to predict mortality. The patient histories were formulated as time series consisting of varying types of clinical events, thus enabling the model to learn increasingly complex temporal dependencies. BERT and XLNet achieved an average area under the receiver operating characteristic curve (AUC) of 75.5% and 76.0%, respectively. XLNet surpassed BERT in recall by 9.8%, suggesting that it captures more positive cases than BERT, which is the main focus of recent research on EHRs and transformers.

Electronic health records (EHRs) encompass evidence of patient care paths and outcomes. Different EHR models have been widely adopted by healthcare facilities and continue to accumulate increasing amounts of data with potential to discover new medical knowledge and to support decision making to improve outcomes for new patients[1]. Although EHRs offer large volumes of longitudinal real-life data for improved machine learning (ML), they still challenge the methodology with their heterogeneous, sparse, often incomplete and even erroneous data[2]. Moreover, due to the sensitive nature of the data, privacy issues and regulations will further complicate model development and deployment in the future[3]. Some regulations may require database anonymization to protect data privacy but this may result in decreased data quality due to additional noise and gaps.

Cardiovascular diseases (CVDs) have held their ranking as the leading cause of death worldwide for years and continue to impose an increasing challenge to the global health. In 2017, CVDs alone caused 17.8 million deaths globally, showing an alarming 21.2 % increase in the yearly CVD death count since 2007[4]. Furthermore, CVDs can be a risk factor in relation to other diseases and increase the demand for hospital care. For instance, they have been linked with poor prognosis in the context of COVID-19, which threatened health care capacity all over the world[5]. The problem of CVDs has not been sufficiently addressed. While the risk could be efficiently reduced with lifestyle changes towards physically active lives and healthier diets, the reports of the aging population and overwhelming obesity rates indicate enduring prosperity for CVDs[4]. Predictive models may help identify high-risk patients and patient deterioration and may be used to focus healthcare resources efficiently to improve patient outcomes and manage the increasing CVD counts. Data-driven approaches are expected to renew the clinical cardiology practice, ascertain their place in the clinician's toolbox, and to reform our understanding of the causes of CVDs[6,7].

[1]VTT Technical Research Centre of Finland Ltd., 33101 Tampere, Finland. [2]Faculty of Medicine and Health Technology, Tampere University, 33720 Tampere, Finland. [3]Finnish Cardiovascular Research Center Tampere, 33520 Tampere, Finland. [4]Vascular Centre, Tampere University Hospital, 33520 Tampere, Finland. [5]Tays Heart Hospital, Tampere University Hospital, 33521 Tampere, Finland. [6]Faculty of Information Technology and Communication Sciences, Tampere University, 33720 Tampere, Finland. [7]These authors contributed equally: Jussi Hernesniemi and Moncef Gabbouj. ✉email: emmi.antikainen@gmail.com

Transformer neural networks are the state-of-the-art machine learning methods for sequential data modelling. Developed for natural language processing, their built-in properties respond to many needs that arise when using EHR data. Thanks to them combining attention and positional encoding, transformers can be applied to learn bidirectional temporal dependencies despite the sparsity and possible errors in the large volumes of EHR data. Their design to handle textual input does not exclude numerical input and may thus be useful for heterogeneous input types. In the context of EHRs, they have been applied mainly to clinical notes or diagnoses[8–11]. Yet, their capabilities to capture more complex dependencies in heterogeneous databases have received little attention[12]. Furthermore, prior studies have focused on one transformer variant; bidirectional encoder representations from transformers (BERT)[13]. A newer model, XLNet, has surpassed BERT in many baseline natural language processing tasks[14]. This work uses an anonymous cardiac patient EHR database to compare the learning capabilities of BERT and XLNet in the important application of mortality risk prediction. Here, the transformer models are applied to multi-modal heterogeneous patient event time series, comprising both textual and numerical attributes.

Prior to transformers, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) achieved encouraging results in, e.g., arrhythmia detection from electrocardiograms (ECG), diagnostic decision support using cardiovascular images, and diagnosis prediction from EHR data[15–18]. The introduction of attention mechanisms provoked countless new studies reporting improved results[19–23]. Importantly for clinical applications, attention gave interpretability to the model outcomes, thus offering one solution to the primarily criticized shortcoming of deep learning (DL) methods[19]. For example, Choi et al. presented the RETAIN model which coupled attention with recurrent neural networks (RNNs) to predict heart failure from EHR data. They presented a method for prediction interpretation while reporting an 87% area under the receiver operating characteristics curve (AUC)[21]. Another relevant study was conducted by Rajkomar et al. who used an ensemble of three DL models, one of which was attention-based, and tested their system on EHR data from two hospitals. They achieved 93–95% AUC for in-patient mortality prediction at 24 h after admission[22].

The original Transformer relied exclusively on attention mechanisms[24]. The Transformer and its variants surpassed RNNs by allowing parallelized computing and by learning bidirectional dependencies. They learned longer-range dependencies at improved training time, which is crucial with long input sequences like EHR histories[24]. The first studies applying transformers directly on EHRs were built on BERT, which bases its learning strategy on masking the input[13]. Shang et al. combined BERT with ontology embeddings from a graph neural network creating G-BERT for medication recommendation[10]. They reported a 1% increase in precision-recall AUC as compared to RETAIN. BEHRT by Li et al. applied BERT directly for disease prediction by using sequences of diagnoses available in the EHRs[9]. They reported a patient-averaged AUC of 95–96% for varying prediction windows extending up to 12 months. Thirdly, Rasmy et al. reported up to 2% improvement in disease prediction with their Med-BERT as compared to RETAIN[11]. They evaluated Med-BERT for heart failure prediction in diabetic patients and pancreatic cancer onset prediction. Some studies have additionally proposed somewhat modified transformers for EHR representation learning[25–27].

In this study, we apply the ground-breaking transformer models on patient time series to predict 6-month mortality in cardiac patients. The 6 months prediction period may offer actionable predictions for many chronic conditions. Unlike BEHRT and Med-BERT which were trained on sequences of diagnostic codes, we incorporate over a dozen different event types each described by multiple attributes to capture a more complete depiction of the patient history. By feeding the transformers sequences of patient events with timestamps based on age, the models may learn how the interplay between different events and their outcomes, as well as temporal dependencies, affect the patient outcome. With this approach, the patterns learned by the model may unveil unforeseen associations between different events. Moreover, we study both BERT and XLNet. Unlike BERT, XLNet is an auto-regressive transformer variant that avoids corrupting the input[14]. We exploit the anonymous EHRs of over 23,000 cardiac patients who were treated at Tays Heart Hospital in Finland and report our findings on using privacy-preserving anonymous data in model development, an increasingly common starting point for future EHR studies. A previous machine learning study on the same database considered a subset of 9066 consecutive acute coronary syndrome patients and achieved an AUC of 89% for 6-month mortality using conventional, non-deep learning methods[28]. This study takes up a more complex challenge of predicting mortality in all available CVD patients, comprising a more heterogeneous patient population.

Methods

Study data. The longitudinal study data comprised three Finnish data sources: (1) the EHR by the Pirkanmaa Hospital District (PHD), (2) the KARDIO registry by the Tampere Heart Hospital, and (3) the Finnish mortality registry by Statistics Finland. The PHD EHR extend back to the 1990's and the date of death from the mortality registry was included for the matching period. The PHD EHR data include hospital discharge diagnoses, which record every diagnosis recorded for the patient in ICD-10 format (and previously in ICD-9 and ICD-8 formats). This data is equally reported in every hospital nationally and the validity of the registry is high for many significant cardiovascular conditions such as strokes, coronary heart disease and heart failure[29–31]. The KARDIO registry is the most recent of the three; its first entries date back to the early 2000's. The original database was automatically collected from the three registries until January 2020 as a part of a retrospective registry study, MADDEC (Mass Data in Detection and Prevention of Serious Adverse Events in Cardiovascular Disease)[32].

The study was approved by the Pirkanmaa Hospital District Institutional Review Board's scientific steering committee. Informed consent is waived since the retrospective nature of the study by the Pirkanmaa Hospital District Institutional Review Board's scientific steering committee. The study was conducted according to the declaration of Helsinki as applicable and the study data was processed in accordance with the Finnish legislation.

An anonymous version of the database was used, comprising 72,680 patients (9172 deceased patients within six months of their last visit, i.e., 12.6%) all treated at the Tays Heart Hospital for different cardiac conditions.

Input sequence formulation. The patient records were extracted from the event-oriented database, pre-processed, and finally ordered on temporal attributes to formulate a single time series of events for each patient. The resulting time series were further processed into appropriate input for the transformer neural networks, as summarized in Fig. 1. In the anonymous database, the temporal attributes were age-based (on a daily level, i.e., days since birth) and the real dates and times were unknown.

Data pre-processing included replacing Roman numerals with Arabic numerals, filling in event start or end times when only one of the two was missing, and unifying notations. Units and measurement names, anesthesia types, and urgency classes were translated from Finnish to English. Additionally, some body-mass index (BMI) values below 0.02 were presumed to use centimeters instead of meters for height, and thus, multiplied by 10^4 to restore correct units. For events where the ending time preceded the recorded event start time, the timestamp order was presumed a typographical error and the timestamps were switched. In the end, only the pre-processed event start time was included as the event durations were generally error-prone.

Any events occurring before the patient became of age were excluded. Additionally, any events with missing event timing or overlap with the date of death were excluded. The latter consists of events extending to the date of death, e.g., resuscitation or procedures, or beyond, such as lab values or diagnoses.

The data sources contained 14 distinct types of events each described by a different set of attributes, as presented in Table 1. Here, we took the liberty of excluding any attributes that were never present for an event type. For angiography, percutaneous coronary intervention (PCI), coronary care unit (CCU), transcatheter valve implantation (TAVI), and resuscitation events, the attributes were limited to the nine most available attributes (out of tens of attributes) to control input sequence length and to fit multiple events in the input sequence. Each event was constructed into a sequence simply by listing the event type and the corresponding attributes in one sequence. Thus, each event type was represented by a specific "sentence structure" mimicking natural language. Any missing attribute values were filled in with 'None'. All event representations started with the event type name, the event starting time, and residence (among Finnish counties) when available. Residence was available in 67%, 74%, and 79% of CCU, resuscitation, and hospital ward events respectively, while it was missing completely for TAVI and in 59–94% for other event types. Event type, start time, all operation attributes, times repeated, ward, sex, stenosis, imaging type, dialysis, temporary pacemaker, primary vasoactive medication, fluoroscopy time, and glomerular filtration rate attributes were all fully available for the relevant event types. The remaining attributes in labs were available in 72–75% of lab events (except textual values only in 1%). Diagnosis code and priority were available in 36% and 41% of diagnosis events respectively, and anesthesia type, ASA class, and urgency in 56%, 61%, and 93% of procedures. All measurement event attributes were available in 91–100% of measurement events. The remaining attributes in angiography, PCI, CCU, and TAVI were available in 98-100% of the respective events, whereas the other attributes for resuscitation events were available in 89% of resuscitation events.

The individual pre-processed events were combined in order of occurrence into one sequence per patient, forming the patient event timeline. Until this point, the events were linked via patient and event pseudo-identifiers. The pseudo-identifiers were removed and the date of death was isolated and transformed into a binary class: positive (1) when the date of death occurred within 182 days of the last event, and negative (0) otherwise. The date of death was comprehensively obtained from the Finnish mortality registry. Importantly, in the real-life

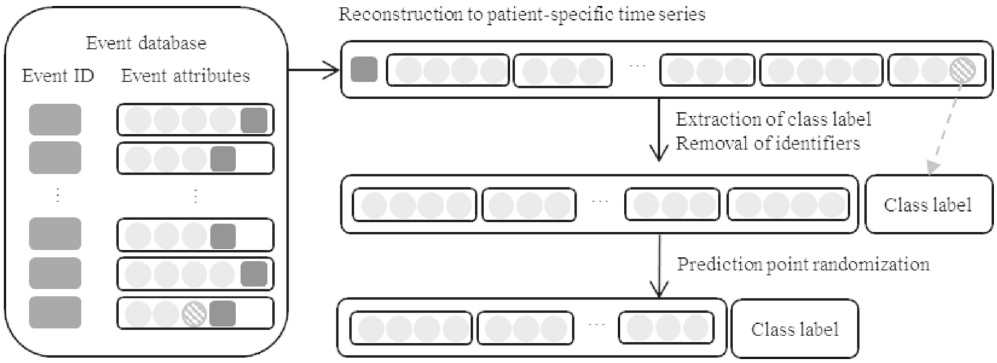

Figure 1. A schematic example of how a patient's events were formulated into a time series. The records in the event-oriented database contain the event type specific attributes (yellow circles). First, the events related to the same patient ID (magenta square) were combined to a sequence sorted according to their temporal attributes. Hereafter, the patient ID was no longer necessary. Next, the class (prediction target, i.e., death within six months or alive) was computed using the death-related attribute (green striped circle) and the time between that and the previous event. Finally, the point of prediction was randomized. If the class was positive, the time to death from the final remaining event was maintained within the selected six month period.

Event type	Attributes
Labs	Event type, start time, residence, lab test value (num), lab test value (char), lab test name, lab test unit
Diagnosis	Event type, start time, residence, diagnosis code[a], diagnosis priority
Medication	Event type, start time, residence, ATC code[b], daily dosage, dose unit, administration method
Operation	Event type, start time, residence, sequence number, code ID, code[c]
Procedure	Event type, start time, residence, anesthesia type, ASA class[d], operation urgency
Measurement	Event type, start time, residence, measurement value (num), measurement context*, measurement name, measurement unit, measurement context code**
Hospital visit	Event type, start time, residence
Hospital ward	Event type, start time, residence, times repeated, ward
Angiography	Event type, start time, residence, times repeated, ward, primary angiography findings, sex, stenosis (boolean), primary puncture places
Percutaneous coronary intervention (PCI)	Event type, start time, residence, times repeated, ward, complications, sex, indication, urgency
Imaging	Event type, start time, residence, imaging type
Coronary care unit (CCU)	Event type, start time, residence, times repeated, ward, dialysis, sex, temporary pacemaker, primary vasoactive medication
Transcatheter aortic valve implantation (TAVI)	Event type, start time, times repeated, ward, dyslipidemia, fluoroscopy time, sex, glomerular filtration rate, hypertension
Resuscitation	Event type, start time, residence, times repeated, ward, family history***, sex, hypertension, smoking

Table 1. Event specific attributes *Measurement context in text format, for example a suspected diagnosis or type of the visit. **Measurement context related code (e.g. an ICD-10 diagnostic code). ***Family history (for early coronary artery disease) was positive if at least one of the patient's first degree relatives had suffered a myocardial infarction or underwent coronary revascularization (PCI or coronary artery bypass surgery) at an early age (< 55 and < 65 years in men and women, respectively). [a]International Statistical Classification of Diseases and Related Health Problems, the 10th revision (ICD-10). [b]Anatomical Therapeutic Chemical (ATC) code. [c]Nordic Classification of Surgical Procedures (NCSP). [d]American Society of Anesthesiologists (ASA) classification of physical status.

clinical use-case the model could be used to produce predictions at any time of a patient timeline. Therefore, to produce realistic evaluation of model performance and avoid bias due to the retrospective nature of the data, a random number of events at the end of a patient's timeline were erased. The number of erased events was selected randomly between zero and a patient-specific maximum number such that at least five events remained for the patient, and the death for any positive case would still occur within the selected cutoff from the final remaining event.

Finally, the input sequences were tokenized and the special tokens for class (CLS) and sentence separation (SEP) were added once to each patient timeline according to their expected position in BERT and XLNet[33]. Any numerical input was transformed into string-type integers for tokenization. The age in days was transformed into full years.

Hyperparameter optimization. The model hyperparameters were optimized using Population Based Training (PBT)[34,35]. PBT is an evolutionary algorithm, which trains several networks with varying hyperparameters in parallel. During the training process, each network can explore hyperparameters randomly in a predefined space or exploit another better performing parallel model by copying its parameters and continuing to explore new hyperparameters with the partially trained model, without restarting the training from scratch.

PBT was applied to optimize the learning rate, dropout fraction, and model dimensions including the number of heads and layers, as well as layer size. Due to memory limitations, only batch sizes 16 and 24 were tested. PBT was run for both BERT and XLNet for 30 epochs on 12 trials with the perturbation interval of ten epochs. Similarly to the original transformers, Gaussian Error Linear Unit (GELU) was used for activation.

Model evaluation. Eighty percent of the study data was used for model development, while 20% was held out as a test set. The development data was further split into training and validation data, comprising 80% and 20% of the development set, respectively. Stratified splits were used to maintain a similar distribution of positive and negative cases in each set. The data sets were further balanced by taking a random sample of negative cases to match the number of positive cases (see details in Implementation). Model performance was assessed with AUC, precision (positive predictive value), and recall (sensitivity)[36].

PBT was performed on the development set. The top-performing BERT and XLNet models were validated using stratified fivefold cross validation with the development data. Subsequently, the final BERT and XLNet models were trained with the selected hyperparameters on the full development data and evaluated on the held out test data set.

The final model training was repeated five times to account for the effect of random initialization. Early stopping was applied when the training loss failed to improve at least by 0.0045 over 5 epochs (`min_delta` and `patience` in keras EarlyStopping, selected based on the previously observed cross-validation losses). To interpret what the final models had learned, the models were fed example time series from the test set and the attention weights were visualized using BertViz[37]. Min–max normalization was applied to the attention layers prior to the visualization to properly highlight where attention was at its highest and lowest.

Implementation. The data were tokenized using pretrained tokenizers (bert-base-cased, xlnet-base-cased) available in the Hugging Face model database[13,14]. The transformer models were implemented in Python by using the Hugging Face Transformers library together with Tensorflow[33,38]. The Ray Tune package (function API) was used for hyperparameter optimization[35]. The data split for model evaluation was obtained using scikit-learn[39]. The final models were trained using an Adam optimizer with an epsilon of 10^{-8}. The sequence length was restricted to 512 tokens such that the latest information in the patient history was included. Overlength sequences were truncated and under-length sequences padded using the tokenizer-specific padding token.

Class imbalance was managed by (1) down-sampling the negative examples in the training and validation sets and (2) using a weighted binary cross-entropy loss function. To ensure that each limited-size batch had a reasonable chance of including some positive cases, the negative samples were randomly down-sampled so that 25% of the samples in both training and validation set were positive. By limiting the extent of down-sampling, the related data loss was also limited. The remaining imbalance was counteracted via the loss function using balanced class weights; each class was weighted by its inverse prevalence in the development set, further divided by the number of classes (two).

A 32 gigabyte Tesla V100-DGXS graphics processing unit (GPU) was used in hyperparameter optimization and training the models.

Results

Implementing the exclusion criteria reduced the study data from 72,680 patients to 57,377 adult patients, including 3771 (6.57%) positive cases. The demographic details are described in Table 2. The average age of patients was 65 years (79 for positive cases). The sex of the patient was only available for 35.7%, most of which (61.2%) were male. The notably large portion of sex information was lost upon anonymization as the national identification numbers were removed.

After down-sampling the development sets to counteract class imbalance (as detailed in "Methods"), the resulting training and validation sets contained 9640 and 2420 patients (12,060 in total with 3015 positive cases). The test set comprised 11,482 patients with the number of positive cases, 756 (6.58%) corresponding approximately to the prevalence in the full pre-processed data. Thus, the study involved 23,542 individuals (including all 3771 positive cases). The patient flow is summarized in Fig. 2.

The hyperparameters optimized using PBT are presented in Table 3. BERT performed best on learning rates around 5×10^{-7} to 1×10^{-6}, whereas rates an order of magnitude larger (5×10^{-6} to 1×10^{-5}) worked best

	N	Female	Male	Age range	Mean years of data (SD)	Mean no. of events (SD)
Positive	3771	691 (18.3%)	1183 (31.4%)	18–102	6.5 (3.4)	1755 (2364)
Negative	53,606	7249 (13.5%)	11,365 (21.2%)	18–105	4.2 (3.8)	553 (1091)
Total	57,377	7940 (13.8%)	12,548 (21.9%)	18–105	4.4 (3.9)	632 (1255)

Table 2. Pre-processed study data. The percentages depict the proportion of the (known) sex with respect to the full number of patients on the same row. *SD* standard deviation.

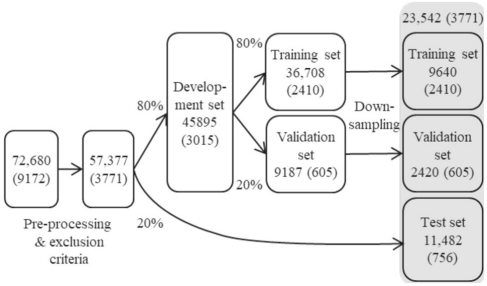

Figure 2. Patient flow diagram. The total number of patients is indicated for each step and the number of positive cases is denoted in brackets. The final data used in model evaluation comprised 23,542 patients and is depicted on a gray background.

Hyperparameter	BERT	XLNet
Hidden size	144	144
Number of layers	12	6
Number of attention heads	12	6
Feed-forward layer hidden size	128	128
Learning rate	1×10^{-6}	5×10^{-6}
Batch size	16	16
Dropout	0.5	0.4

Table 3. Hyperparameters optimized via population based training.

for XLNet. The selected configurations comprised 108,312,578 trainable parameters for BERT and 5,482,130 parameters for XLNet.

The models with optimized hyperparameters were cross-validated using 5-fold validation to assess their sensitivity to the selection of training instances. The validation results are presented in Table 4. The models achieved similar average AUC. BERT achieved slightly higher precision but the variance between folds was also higher. However, less than half of the predicted cases were true positive cases. Finally, XLNet reached a notably higher average recall, with low variance between folds. Thus, the optimized XLNet was more sensitive to detect positive cases than BERT.

The final model training was repeated five times to examine the effect of random initialization. The test results obtained on the held-out test set are presented in Table 5. The corresponding mean specificity scores were 78% and 69% for BERT and XLNet, respectively. The test set results support the observations from cross-validation. The slight improvement in AUC and recall were likely due to early stopping, which stopped the training already before 50 epochs in all cases. This prevented over-fitting, which occurred remarkably early for this data and models. The drop in precision is explained by the increased class imbalance in the test set but also underlines that both models produced mostly false positives, despite capturing 73–83% of the positive cases on average. In comparison to BERT, the improvement in XLNet's recall exceeds the drop in precision and, thus, the XLNet model may be more useful.

The final BERT and XLNet models exhibited very similar metrics (run number 5 in Table 5) and were fed an example time series from the test set for interpretation. The model attention for the 50 tokens nearest to the classification token in an example time series are depicted in Fig. 3a,b for XLNet and BERT, respectively. The selected (full sequence) example was correctly labeled positive by XLNet and mislabeled negative by BERT. The

	BERT			XLNet		
Fold	AUC	Precision	Recall	AUC	Precision	Recall
1	0.7452	0.4567	0.8126	0.7438	0.4592	0.8027
2	0.7366	0.5244	0.6783	0.7570	0.4740	0.8159
3	0.7703	0.4711	0.8640	0.7692	0.4873	0.8292
4	0.7432	0.5350	0.6849	0.7454	0.4496	0.8292
5	0.7689	0.5047	0.7993	0.7557	0.4815	0.7977
Mean	0.7528	**0.4984**	0.7678	**0.7542**	0.4703	**0.8149**

Table 4. 5-fold cross-validation of optimized models. Performance metrics in the validation set, after 50 epochs. The best mean score for each metric (AUC, precision, recall) is in bold.

	BERT			XLNet		
Run	AUC	Precision	Recall	AUC	Precision	Recall
1	0.7398	0.2248	0.6336	0.7556	0.1533	0.8373
2	0.7612	0.1923	0.7421	0.7602	0.1574	0.8360
3	0.7586	0.1919	0.7355	0.7654	0.1665	0.8201
4	0.7547	0.1937	0.7209	0.7609	0.1601	0.8280
5	0.7591	0.1571	0.8333	0.7586	0.1563	0.8347
Mean	0.7546	**0.1919**	0.7330	**0.7602**	0.1587	**0.8312**

Table 5. Blind test results on five different initializations. The best mean score for each metric (AUC, precision, recall) is in bold.

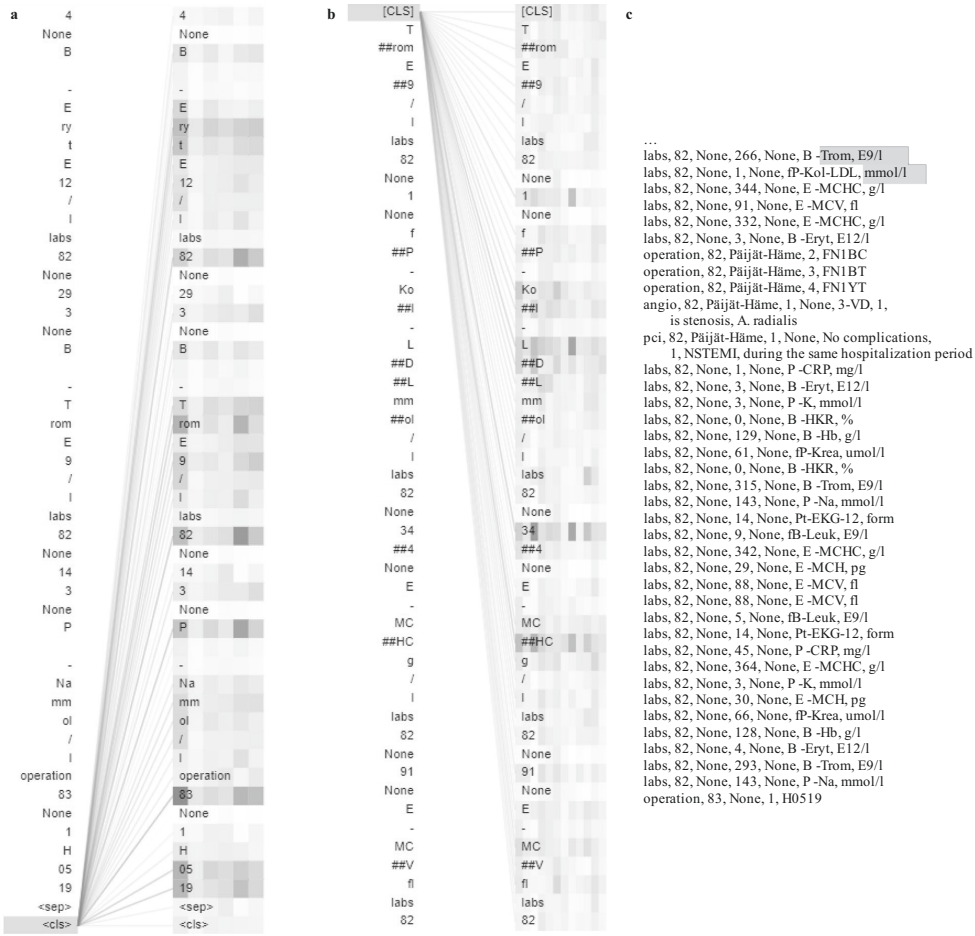

Figure 3. Attention (**a**) in the fifth attention layer in XLNet at the end of an example subsequence near the <cls> token, and (**b**) in the final attention layer in BERT at the start of an example subsequence near the [CLS] token. The different colours represent the (**a**) six and (**b**) 12 attention heads; the more opaque the colour, the heavier the attention. The final events of the example time series are presented in a human readable format in (**c**), where the first information visible to BERT is highlighted with purple and with green for XLNet. Figures (**a,b**) were produced using BertViz[37].

corresponding pre-processed example input before tokenization is depicted in Fig. 3c. The full patient history comprised 111 events, whereas the models could only consume input from up to 38 events.

The 83 years old patient's latest event was an operation encoded as H0519, which stands for a simulation film, possibly related to radiation therapy planning. Their history also showed, e.g., an angiography of the heart and/or coronary artery, a percutaneous transluminal coronary angioplasty, and an intraventricular stent placement to enlarge the coronary artery, all within the past year. As seen in Fig. 3a at the end of the sequence, XLNet attends especially to the age (three instances visualized) and to the operation code. Most other layers also attend to age and the operation code at the <cls> classification token, while exhibiting varying attention to the other inputs. In contrast, BERT's attention at the [CLS] classification token in Fig. 3b does not exhibit special attention to the patient's age (not the primary focus of attention in any layer) but attends to some of the lab results. It is noted that a tokenizer specialized in EHR data might not only make the interpretation easier but also improve attention results.

Discussion

This work explored and compared the potential of two popular transformers, BERT and XLNet, in the task of predicting 6-month mortality in cardiac patients at randomly chosen events recorded in their EHR. The heterogeneous electronic health record data were constructed into semi-structured multi-event time series to exploit the temporal information. We achieved a higher recall with XLNet, suggesting that it captures more positive cases than BERT. It has been argued that the learning strategy implemented in XLNet is better capable of capturing long-term dependencies in sequences[14]. To our knowledge, this is the first study exploring XLNet for mortality prediction from electronic health records.

Previous studies often set their focus on in-patient mortality within 24 h of admission, which can be especially beneficial for applications at intensive care units[2]. In contrast, patients with long-term conditions may profit from earlier predictions. The 6-month prediction period selected in this study allows time for clinicians to re-evaluate the patient's needs and make their care more effective to decrease their risk of death. It provides time for any additional tests and diagnostics, as well as a realistic possibility for interventions to take effect. Six months was considered a suitable period to explore model performance in such a heterogeneous cardiac patient population.

As compared to a prior study using extreme gradient boosting (XGBoost) on the same database and prediction target, the presented results fall short of the previous AUC result[28]. This may, however, be expected because the prior study focused on a specific homogeneous patient group (with acute coronary syndrome) whereas the current work with a larger portion of the database included a wide heterogeneous spectrum of CVD patients. Moreover, the more refined and smaller subset of data in the previous study allowed for features selected by expert clinicians, which may have further facilitated good performance but also increased manual work. Additionally, this study used the anonymous database, which lead to more noise and gaps in the training data and only offered dates relative to a patient's birth instead of real dates. Hence, the importance of the concurrent planning of the analysis and anonymization is underscored. In this study, because the collection of study data was terminated on a specified date without any follow-up, the data contained patients that were still in care or did not have a full six months since their last event. These examples could not be filtered from the anonymous data as the real dates were no longer available and, thus, they may cause the model to be too optimistic about patient survival. The missing real dates also prevented the analysis of time-dependent differences between patient timelines which might exist due to, e.g., updates in care guidelines. Moreover, the sex of patient was largely missing although it is an important clinical factor affecting patient outcomes.

Even though some transformers such as XLNet are in principle able to consume sequences of any length, the models are still limited by the memory resources of the hardware used for training and visual output interpretation. This poses a challenge for incorporating all different event types and their attributes from the patient history. Here, the 512 tokens representing the most recent events of the patient were used while the captured time period varied. Formulating the EHR data as multi-event time series may facilitate the extraction of new knowledge concerning the role and relationships of different types of events. Future research may explore longer input sequences with XLNet or alternative ways to incorporate multi-event information. For instance, replacing code based attributes with full text descriptions may improve performance but would require longer input sequences to feed the model the equivalent portion of patient history. In the future, harmonization of hospital information management systems may additionally yield better grounds regarding the selection of attributes as they are inherited from the hospital's original system. Further improvement may be achieved by using tokenizers specially trained on clinical data or pre-trained transformers. Here, due to the lack of such resources for XLNet, both models were trained from scratch to facilitate a fair comparison. Notably, our results show that the standard English tokenizers can produce promising learning results.

As demonstrated in this work, transformers provide a means to interpret individual outputs and the predictions may therefore become a valuable part of the clinical workflow and answer to the requirements set for ML models in CVD predictions[40]. Nevertheless, intuitive and user-friendly output interpretation interfaces for clinicians need further development so that this capability can be properly harnessed. The resulting tools may be efficiently integrated to the EHR system itself, although additional computing resources are likely required.

Conclusion

Using transformers to learn bi-directional dependencies in EHRs shows promise in mortality prediction, despite the sparsity of the data. We compared BERT and XLNet for CVD patient mortality risk prediction from EHR data. While prior research has focused on BERT for EHR applications, the results of this study suggest that future studies may achieve improved results using XLNet. Similar models with actionable outputs, as presented here, could improve patient outcomes with chronic diseases, such as CVDs, and be directly integrated to the EHR systems for everyday clinical use.

We also observed that transformers may perform better in more refined patient groups. The wide spectrum of CVD patients in this study added complexity to the prediction problem, producing weaker performance as compared to conventional machine learning in a more refined patient group. Furthermore, more concise representations have reached better learning results, whereas the multi-attribute multi-event representation faces computational restrictions. Hence, in the future, improved results may be obtained via more sophisticated representations, transfer learning from pre-trained models, or via improved computational power. As in the presented study, anonymous data will become an increasingly common basis for model development. In such cases, the performance of data-driven models may benefit from an improved anonymization process.

Data availibility

The anonymized data is available for scientific purpose upon reasonable request to J.H. (jussi.hernesniemi@sydansairaala.fi) pending the approval of the MADDEC study steering committee.

Received: 7 July 2022; Accepted: 27 February 2023
Published online: 02 March 2023

References

1. Kruse, C. S., Stein, A., Thomas, H. & Kaur, H. The use of electronic health records to support population health: A systematic review of the literature. *J. Med. Syst.* **42**, 214. https://doi.org/10.1007/s10916-018-1075-6 (2018).
2. Si, Y. *et al.* Deep representation learning of patient data from electronic health records (EHR): A systematic review. *J. Biomed. Inform.* **115**, 103671. https://doi.org/10.1016/j.jbi.2020.103671 (2021).
3. Lähteenmäki, J., Pajula, J. & Antikainen, E. Development of medical applications based on AI models and register data-regulatory considerations. in *Proceedings of the 18th Scandinavian Conference on Health Informatics*. https://doi.org/10.3384/ecp187024 (2022).
4. Roth, G. A. *et al.* Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the global burden of disease study 2017. *Lancet* **392**, 1736–1788. https://doi.org/10.1016/S0140-6736(18)32203-7 (2018).
5. Zhao, M. *et al.* Advances in the relationship between coronavirus infection and cardiovascular diseases. *Biomed. Pharmacother.* **127**, 110230. https://doi.org/10.1016/j.biopha.2020.110230 (2020).
6. Quer, G., Arnaout, R., Henne, M. & Arnaout, R. Machine learning and the future of cardiovascular care. *J. Am. Coll. Cardiol.* **77**, 300–313. https://doi.org/10.1016/j.jacc.2020.11.030 (2021).
7. Hemingway, H. *et al.* Big data from electronic health records for early and late translational cardiovascular research: Challenges and potential. *Eur. Heart J.* **39**, 1481–1495. https://doi.org/10.1093/eurheartj/ehx487 (2017).
8. Gao, S. *et al.* Limitations of transformers on clinical text classification. *IEEE J. Biomed. Health Inform.* **25**, 3596–3607. https://doi.org/10.1109/JBHI.2021.3062322 (2021).
9. Li, Y. *et al.* BEHRT: Transformer for electronic health records. *Sci. Rep.* **10**, 7155. https://doi.org/10.1038/s41598-020-62922-y (2020).
10. Shang, J., Ma, T., Xiao, C. & Sun, J. Pre-training of graph augmented transformers for medication recommendation. in *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019*, IJCAI International Joint Conference on Artificial Intelligence. 5953–5959. https://doi.org/10.24963/ijcai.2019/825 (2019).
11. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* **4**, 86. https://doi.org/10.1038/s41746-021-00455-y (2021).
12. Meng, Y., Speier, W., Ong, M. K. & Arnold, C. W. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE J. Biomed. Health Inform.* **25**, 3121–3129. https://doi.org/10.1109/JBHI.2021.3063721 (2021).
13. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. in *Proceedings of NAACL-HLT 2019*. 4171–4186. https://doi.org/10.48550/arXiv.1810.04805 (2019).
14. Yang, Z. et al. XLNet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.*. https://doi.org/10.48550/arXiv.1906.08237 (2019).
15. Kiranyaz, S., Ince, T. & Gabbouj, M. Personalized monitoring and advance warning system for cardiac arrhythmias. *Sci. Rep.* **7**, 9270. https://doi.org/10.1038/s41598-017-09544-z (2017).
16. Oh, S. L., Ng, E. Y., Tan, R. S. & Acharya, U. R. Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. *Comput. Biol. Med.* **102**, 278–287. https://doi.org/10.1016/j.compbiomed.2018.06.002 (2018).
17. Litjens, G. *et al.* State-of-the-art deep learning in cardiovascular image analysis. *JACC Cardiovasc. Imaging* **12**, 1549–1565. https://doi.org/10.1016/j.jcmg.2019.06.009 (2019).
18. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F. & Sun, J. Doctor AI: Predicting clinical events via recurrent neural networks. in *Proceedings of the 1st Machine Learning for Healthcare Conference*. 301–318. https://doi.org/10.48550/arXiv.1511.05942 (PMLR, Northeastern University, 2016).
19. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. in *Proceedings of the 3rd International Conference on Learning Representations, ICLR*. https://doi.org/10.48550/arXiv.1409.0473 (2015).
20. Ayala Solares, J. R. *et al.* Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J. Biomed. Inform.* **101**, 103337. https://doi.org/10.1016/j.jbi.2019.103337 (2020).
21. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F. & Sun, J. RETAIN: Interpretable predictive model in healthcare using reverse time attention mechanism. *CoRR abs/1608.05745*. https://doi.org/10.48550/arXiv.1608.05745 (2016).
22. Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* **1**, 18. https://doi.org/10.1038/s41746-018-0029-1 (2018).
23. Pham, T., Tran, T., Phung, D. & Venkatesh, S. Predicting healthcare trajectories from medical records: A deep learning approach. *J. Biomed. Inform.* **69**, 218–229. https://doi.org/10.1016/j.jbi.2017.04.001 (2017).
24. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* 5998–6008. https://doi.org/10.48550/arXiv.1706.03762 (2017).
25. Choi, E. *et al.* Learning the graphical structure of electronic health records with graph convolutional transformer. in *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 606–613. https://doi.org/10.48550/arXiv.1906.04716 (2020).
26. Ren, H., Wang, J., Zhao, W. X. & Wu, N. *RAPT: Pre-Training of Time-Aware Transformer for Learning Robust Healthcare Representation*. 3503–3511. https://doi.org/10.1145/3447548.3467069 (Association for Computing Machinery, 2021).
27. Kodialam, R. *et al.* Deep contextual clinical prediction with reverse distillation. in *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*. https://doi.org/10.48550/arXiv.2007.05611 (2021).
28. Hernesniemi, J. A. *et al.* Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome—The MADDEC study. *Ann. Med.* **51**, 156–163. https://doi.org/10.1080/07853890.2019.1596302 (2019).
29. Tolonen, H. *et al.* The validation of the Finnish hospital discharge register and causes of death register data on stroke diagnoses. *Eur. J. Cardiovasc. Prevent. Rehabil.* **14**, 380–385. https://doi.org/10.1097/01.hjr.0000239466.26132.f2 (2007).
30. Pajunen, P. *et al.* The validity of the Finnish hospital discharge register and causes of death register data on coronary heart disease. *Eur. J. Cardiovasc. Prevent. Rehabil.* **12**, 132–137. https://doi.org/10.1097/00149831-200504000-00007 (2005).
31. Vuori, M. A. *et al.* The validity of heart failure diagnoses in the Finnish hospital discharge register. *Scand. J. Public Health* **48**, 20–28. https://doi.org/10.1177/1403494819847051 (2020).
32. Hernesniemi, J. A. *et al.* Cohort description for MADDEC—Mass data in detection and prevention of serious adverse events in cardiovascular disease. in *EMBEC & NBC 2017* (Eskola, H., Väisänen, O., Viik, J. & Hyttinen, J. eds.). 1113–1116. https://doi.org/10.1007/978-981-10-5122-7_278 (Springer Singapore, 2018).
33. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6. (Association for Computational Linguistics, 2020).
34. Jaderberg, M. *et al.* Population based training of neural networks. *CoRR abs/1711.09846*. https://doi.org/10.48550/arXiv.1711.09846 (2017).

35. Liaw, R. *et al.* Tune: A research platform for distributed model selection and training. arXiv preprint arXiv:1807.05118. https://doi.org/10.48550/arXiv.1807.05118 (2018).
36. Tohka, J. & van Gils, M. Evaluation of machine learning algorithms for health and wellness applications: A tutorial. *Comput. Biol. Med.* **132**, 104324. https://doi.org/10.1016/j.compbiomed.2021.104324 (2021).
37. Vig, J. A multiscale visualization of attention in the transformer model. in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 37–42 . https://doi.org/10.18653/v1/P19-3007 (Association for Computational Linguistics, 2019).
38. Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467. Software available from tensorflow.org. https://doi.org/10.48550/arXiv.1603.04467 (2016).
39. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
40. van Smeden, M. *et al.* Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. *Eur. Heart J.*https://doi.org/10.1093/eurheartj/ehac238 *(2022).*

Acknowledgements

This work was supported by Ministry of Social Affairs and Health, Finland, via the *Data-driven identification of elderly individuals with future need for multi-sectoral services* (MAITE) project and by Competitive State Research Financing of the Expert Responsibility Area of Tampere University Hospital, and the Tampere University Hospital support association. The MADDEC project to which the research data is based on was funded by the Business Finland research funding (Grant 4197/31/2015).

Author contributions

E.A. and U.A. prepared the pre-acquired data for the study. E.A. and J.L. built, trained, and tested the models. E.A. performed model hyperparameter optimization. E.A., J.H., M.E., N.O., M.v.G., and M.G. designed the study and the data collection protocol. J.H. and M.G. contributed equally. E.A. drafted and all authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to E.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.