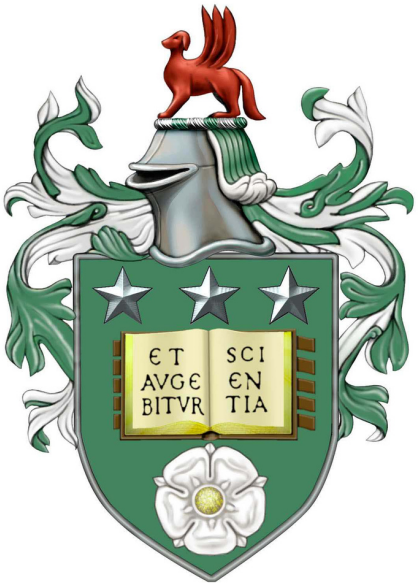
Cardiorespiratory Fitness Estimation utilising Wearable ECG Data

Aron Berger Syversen



A thesis submitted in accordance with the requirements
for the degree of Doctor of Philosophy in the

University of Leeds
School of Computer Science
UK

October 2025

Declaration of Authorship

I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. My contribution and the other authors to this work has been explicitly indicated below. I confirm that appropriate credit has been given within the thesis where reference has been made to the work of others.

- Chapter 2 is based on the review paper *Sensors as a Preoperative Assessment Tool: A Review* in Sensors MDPI by A. B. Syversen, A. Dosis, D. Jayne and Z. Zhang.
- Chapter 4 is based on the published paper *Assessment of ECG Signal Quality Index Algorithms Using Synthetic ECG Data* in Proceedings of Computing in Cardiology 2024 by A. B. Syversen, Z. Zhang, J Batty, M. Kaisti, D. Jayne and D. C. Wong.
- Chapter 5 is based on the published paper *Machine Learning for VO2max Predictions: A Comparison of Methods using Wearable Sensor Data* in Proceedings of IEEE EMBC 2025 by A. B. Syversen, A. Dosis, Z. Zhang, D. Jayne and D. C. Wong.
- Chapter 6 is based on the pre-print publication *Remote Prediction of Cardiorespiratory Fitness in a Preoperative Cohort: Exploring Short and Long-term Heart Rate Variability* published in White Rose Repository by A. B. Syversen, A. Dosis, Z. Zhang, D. Jayne and D. C. Wong. This manuscript is currently under review at *BMC Digital Health*.
- Chapter 7 is based on the published paper *A Framework for Task-Specific Signal Quality Assessment: A Case Study in Heart Rate Estimation* in Proceedings of Computing in Cardiology 2025 by A. B. Syversen, Z. Zhang, D. Jayne, A. Dosis, and D. C. Wong.

For each listed publication, I was the primary author, conducted the primary research, wrote the manuscript, and generated all tables/figures. Other authors contributed to project design, provided feedback on the manuscripts and manuscript revision where necessary. Chapters 5, 6 and 7 all use data from The Remote Monitoring for

Preoperative Risk Assessment for Major Abdominal Surgery (REMOTES) study which was collected by a clinical research team led by Alexios Dosis, who is a co-author on publications relating to these chapters.

Publications

This thesis was primarily based on one journal publication, one pre-print paper and three conference papers [1, 2, 3, 4, 5], with other relevant publications listed.

Journal Papers

- [1] **Syversen AB**, Dosis A, Jayne D, Zhang Z. Wearable sensors as a preoperative assessment tool: a review. Sensors. 2024 Jan 12;24(2):482.

Pre-print Papers

- [2] **Syversen AB**, Dosis A, Z Zhang, D Jayne, DC Wong. Remote Prediction of Cardiorespiratory Fitness in a Preoperative Cohort: Exploring Short and Long-term Heart Rate Variability. Research Square. 2025 July. This manuscript is currently under review at *BMC Digital Health*.A

Conference Papers

- [3] **Syversen AB**, Zhang Z, Batty JA, Kaisti M, Jayne D, Wong DC. Assessment of ECG signal quality index algorithms using synthetic ECG data. In Proceedings of 51st International Computing in Cardiology Conference 2024 Dec 20 (Vol. 51). Computing in Cardiology.

- [4] **Syversen AB**, Dosis A, Z Zhang, D Jayne, DC Wong. Machine Learning for VO2max Predictions: A Comparison of Methods using Wearable Sensor Data. Proceedings of IEEE EMBC. 2025 July. In press.

- [5] **Syversen AB**, Z Zhang, D Jayne, Dosis A, DC Wong. A Framework for Task-Specific Signal Quality Assessment: A Case Study in Heart Rate Estimation. In Proceedings of 52nd International Computing in Cardiology Conference 2025 Dec 20 (Vol. 52). Computing in Cardiology. In press.

Accepted Abstracts

- **Syversen AB**, Zhang Z, Dosis A, Jayne D and Wong DC. A wearable ECG pipeline for cardiorespiratory fitness estimation and HRV interpretation. Accepted for presentation at the 7th UK Mobile, Wearable and Ubiquitous Systems Research Symposium (MobiUK7), 2025.

Other Publications (Unrelated to this Thesis)

- [6] **Syversen AB**, Umney O, Howell L, Breen J, Briggs E, Hancox Z, Khan S, Mills O, Moglia V, Paterson M, Stephens R. “How can we involve Patients?”-Students’ perspectives on embedding PPIE into a doctoral training centre for AI in medical diagnosis and care. Research Involvement and Engagement. 2025 Jul 7;11(1):77.

- [7] Dosis A, Helliwell J, **Syversen AB**, Tiernan J, Zhang Z, Jayne D. Estimating postoperative mortality in colorectal surgery-a systematic review of risk prediction models. International Journal of Colorectal Disease. 2023 Jun 1;38(1):155.

- [8] Wang X, **Syversen AB**, Ding Z, Battye J, Ho SYS, Wong DC. Two-Stage Domain Adversarial Learning to Identify Chagas Disease from ECG and Patient Demographic Data. In Proceedings of 52nd International Computing in Cardiology Conference 2025 Dec 20 (Vol. 52). Computing in Cardiology. In press.

Acknowledgements

As I come to the end of my PhD, there are many people I want to thank. First and foremost, my supervisors Zhi-Qiang, David Jayne, and Dave Wong for their effort, guidance, and patience throughout. Dave may have joined a little later, but certainly not with any less input. Alexios also, thank you for all your hard work in collecting the data, being a constant source of insight and for entertaining any clinical questions with good humour.

Thanks to the CDT for AI in Leeds for taking a chance on a sports scientist who decided to take a leap into computer science. I'm especially grateful to the directors David, Owen, and Vania for supporting every endeavour. I also want to thank Richard, our patient representative, who became more than a board member, a friend, a source of motivation and humorous emails throughout this journey. On that note of course I cannot forget to give thanks to all the patients in Leeds who generously contributed their data: without you, none of this research would have been possible.

For the friends who started as colleagues and ended up friends for life, thank you for the chats, maths help, pints, the encouragement and meme sharing. I apologise for any distractions I brought into the office; I hope introducing juggling to the office didn't slow you down too much! Also to Joe for joining me on a new start in Leeds.

To Amy, thank you for your incredible support (especially this past year) and for putting up with everything — you're the best. Finally, to my family, who backed me through another four years at university after the first four: thank you for always being on my side. I'm aware I still haven't explained particularly well what any of this thesis is about, but you supported me anyway and that means everything.

If I've missed anyone by name, please know I'm deeply grateful.

Abstract

Cardiorespiratory fitness (CRF), expressed as maximal oxygen uptake (VO_2max), is a key indicator of perioperative risk and long-term health outcomes. However, gold-standard cardiopulmonary exercise testing is resource-intensive and not always practical in preoperative pathways. Wearable sensors offer a scalable alternative for estimating VO_2max, but challenges remain around the quality of raw signals, transparency of preprocessing, and integration of advanced features such as heart rate variability (HRV). This thesis aimed to develop and evaluate a methodological framework for predicting VO_2max from wearable sensor data in a preoperative cohort. It (i) evaluated existing signal quality indices (SQIs) for ECG, (ii) assessed the contribution of HRV features, and (iii) developing a task-specific SQI tailored to accurate heart rate estimation.

Data were collected from the REMOTES clinical study which recorded 72-hour wearable-ECG and accelerometer signals from patients scheduled for major abdominal surgery. Multiple open-source SQIs were first compared using annotated synthetic ECG data, and the best-performing approach was applied to the REMOTES dataset. From these SQI-filtered signals, wearable-derived features were used to train a range of machine learning models; regression-based models performed best and were selected for further analysis. HRV features were then incorporated to evaluate their added predictive value, and a new HR-specific SQI was developed using open-access datasets to assess its impact on HR extraction and predictive accuracy.

Integrating HRV features improved VO_2max prediction (R^2 = 0.47 vs 0.42). Implementing the HR-specific SQI further enhanced model performance (R^2 = 0.51; correlation = 0.73), confirming that aligning signal-quality assessment with analytical goals improves HR-derived features and downstream model accuracy. Findings highlight how successive optimisation of feature extraction (via HRV) and signal processing (via a task-specific SQI) can enhance predictive performance. This work presents an analytical pipeline from raw wearable ECG data to VO_2max prediction, providing a foundation for scalable, data-driven cardiorespiratory fitness assessment.

Contents

Chapter 1

Introduction

1.1 Motivation

Approximately 60% of the population in England will undergo some form of surgery during their lifetime [9]. This equates to 12–13 million surgical procedures performed every year in the UK, with global numbers rising year-on-year [10]. As surgical demand continues to grow, reducing the burden of surgery for both patients and healthcare systems is a critical challenge.

All surgical operations carry the risk of complications, which are a source of morbidity for the patient and additional costs for healthcare providers. While complication rates vary by procedure type, major abdominal surgery represents a particularly high-burden. For many cancers, abdominal surgery is considered first line treatment, and in advanced disease may offer the only potential cure when combined with neoadjuvant therapy [11, 12, 13]. However, postoperative complication rates following major abdominal surgery remain high, with 30–50% of patients experiencing at least one complication [14, 15]. Not only do these events have a lasting impact on patients' recovery and quality of life, but they are also strongly associated with prolonged hospital stays and frequent hospital readmissions, driving substantial healthcare costs and amplifying their economic burden [16].

To mitigate these risks, preoperative assessment plays a crucial role in perioperative care. By stratifying patients according to their likely surgical risk, clinicians can optimise a patient's pathway through the perioperative period. This could involve guiding intra-operative planning, and allocating appropriate postoperative resources for recovery.

1.2 Cardiorespiratory Fitness

Cardiorespiratory fitness (CRF), most commonly measured by maximal oxygen consumption (VO_2max), is an individual's capacity to deliver and utilise oxygen during

sustained physical activity [17]. It is a well-established indicator of functional capacity and is consistently associated with morbidity and mortality risk across a range of health conditions. It is widely used in the preoperative setting to evaluate surgical risk, particularly for thoraco-abdominal and gastrointestinal surgery [18]. VO_2max is most accurately assessed via cardiopulmonary exercise testing (CPET) and this is considered the gold standard. There is strong evidence that reduced aerobic capacity measured by CPET is linked to worse perioperative outcomes within abdominal surgeries [19, 20].

Despite its proven value, CPET has several important limitations. It requires specialist equipment and trained staff, incurs substantial cost, and is not universally available across healthcare systems [1]. In addition, the test has poor compliance in some of the highest-risk patients, limiting its applicability in the groups where improved risk stratification is most needed. These barriers restrict the broader adoption of CRF assessment in routine surgical care.

1.3 The Potential of Wearable Sensors

Wearable sensors provide a potential alternative to overcome many of these barriers. Widely available consumer devices, including chest patches, wristbands, and smartphone-based monitors, are increasingly familiar to patients and clinicians alike. More broadly, wearable devices and associated digital biomarkers are emerging as promising tools for diagnosis, monitoring, and treatment across perioperative care [21].

These tools are increasingly suggested as having the potential to support preoperative decision making at scale by providing continuous data capture. Despite this promise, the evidence base remains limited. While calls for further research into wearable sensors in surgical care are growing, there are relatively few studies that investigate how such data should be processed, which physiological markers are most informative, and how accurately they can predict clinical outcomes [22].

One promising application is the estimation of VO_2max from wearable data. Consumer devices already use in-built exercise testing and basic features to approximate fitness, but similar approaches have not been validated in surgical settings where accurate CRF assessment could be most valuable. Developing reliable wearable-based VO_2max prediction methods in the preoperative setting could therefore address many of the barriers to CPET and extend risk assessment to a broader group of patients.

1.4 Challenges with wearable-based CRF Estimation

Despite their promise, several challenges currently limit the use of wearable sensors for preoperative CRF assessment. Much of the current wearable literature in preoper-

ative risk assessment relies on proprietary devices and photoplethysmography (PPG) sensors, which limit transparency and constrain the physiological features that can be analysed. By contrast, raw ECG has the potential to provide more detailed measures such as heart rate variability (HRV), but this remains under-explored in the preoperative setting. ECG also brings additional challenges, particularly vulnerability to artefacts and noise in free-living conditions. These gaps highlight the need for methods that can process raw ECG, quantify signal quality, and determine whether less researched features (for example HRV) provide added value for predicting CRF. This motivates the focus of this thesis.

1.5 Thesis Aims

The overarching aim of this thesis is to investigate whether wearable sensors can provide reliable estimates of cardiorespiratory fitness in the preoperative setting. To address this, the following research questions are posed:

1. How should raw wearable ECG and accelerometer signals collected in free-living conditions be characterised, cleaned, and processed to ensure meaningful downstream analysis?
2. To what extent can VO_2max be predicted from multi-day wearable data in a preoperative surgical cohort?
3. What is the added value of heart rate variability (HRV) features, compared to simpler metrics such as step counts or resting heart rate?

1.6 Thesis Contributions

The main contributions of this thesis are as follows:

- A detailed characterisation of raw wearable ECG and accelerometer data in a preoperative surgical cohort, including analysis of missingness and signal quality.
- Benchmarking of open-source signal quality indices (SQIs) for ECG in free-living conditions, and development of a task-specific SQI framework for heart rate extraction.
- Development of machine learning models for VO_2max prediction, with the added evaluation of HRV features compared to common wearable features.

Following this introductory chapter, this thesis is organised into seven subsequent chapters, each addressing a different aspect of the research. Chapter 2 provides a

technical background, reviewing current preoperative assessment approaches and reviewing wearable sensor modalities, while highlighting key gaps in the literature. In particular, it assess previous attempts to predict cardiorespiratory fitness using wearable data providing an overview of their sensor modalities, processing pipelines, features extracted and modelling methods. Chapter 3 goes on to describe the REMOTES protocol detailing participant recruitment and device design, before characterising the cohort and the dataset extracted from the wearable device by assessing missingness and quality.

Chapters 4-7 are the primary research chapters, each focusing on a different form of signal processing or model development. Chapter 4 evaluates several open-source signal quality indices (SQIs) and identifies their strengths and limitations for free-living ECG data, before applying the selected SQI to process the REMOTES dataset. Chapter 5 applies these insights to develop machine learning models for VO_2max prediction, using conventional physiological features. Chapter 6 focuses on the added value of HRV features in VO_2max prediction, specifically comparing short- and long-term metrics. Chapter 7 proposes a task-specific SQI framework for heart rate extraction, and applies this tool back to the REMOTES dataset to examine its impact on predictive modelling. Finally, Chapter 8 summarises the key findings, limitations of the work, and avenues for future research.

Each of the experimental chapters builds directly upon published or submitted work. Chapters 4, 5, 6, and 7 are based on conference papers and journal submissions, while Chapter 2 is based on a published review and Chapter 3 includes original analyses and a dataset description. Together, they establish a systematic pipeline for processing raw wearable ECG for CRF estimations in the preoperative setting, from data collection and cleaning to feature extraction and predictive modelling.

Chapter 2

Technical Background

2.1 Introduction

This chapter provides a technical background to the thesis by describing current approaches to preoperative assessment, reviewing wearable sensor modalities and identifying gaps in the existing literature. It builds on a literature review originally conducted in late 2023 and published in early 2024, with more recent studies incorporated narratively where directly relevant to the thesis aims.

Demand for general surgery is expected to increase in line with population ageing [23]. A common elective example is abdominal surgery, often performed as treatment for bowel or colorectal cancer, which is the third most common cancer worldwide [24]. Survival outcomes have improved in recent decades, largely due to advances in surgical and perioperative care and mortality rates have more than halved [25, 26, 27, 28, 29].

In contrast, postoperative complication rates have not shown the same decline and remain high, particularly following major abdominal surgery. Common complications include surgical site infection, cardiorespiratory events, gastrointestinal (GI) problems, and anastomotic leak [30, 31, 32]. These events are associated with prolonged hospital stay, ICU readmission, and higher healthcare costs. For example, one large cohort review reported that 23.3% of patients undergoing colorectal surgery were readmitted to hospital, although this number has been known to vary [33]. With an ageing population, the economic burden of such complications is expected to rise further [34, 35].

The identification of high-risk patients therefore remains a central challenge in perioperative care. Preoperative assessment tools aim to stratify risk and guide both intraoperative management and postoperative support, however, none currently provide a precise assessment that is accessible to all patients.

2.1.1 Preoperative Assessment

Preoperative assessment occurs during the first stage of the perioperative period, as seen in Figure 2.1. Here, preoperative measurements are recorded from patients and used to stratify them into risk groups. The major goal of this is to identify patients at highest risk of perioperative morbidity and mortality [36]. As explained previously, it can also support the healthcare provider in resource allocation by estimating the support a patient may require across the perioperative pathway. However, as outlined by NICE guidelines, excessive preoperative testing is related to patient anxiety and significant delays to treatment [37]. Therefore, the benefits of testing should be carefully considered before implementation. The most common preoperative assessment tools are outlined here (see Figure 2.2).

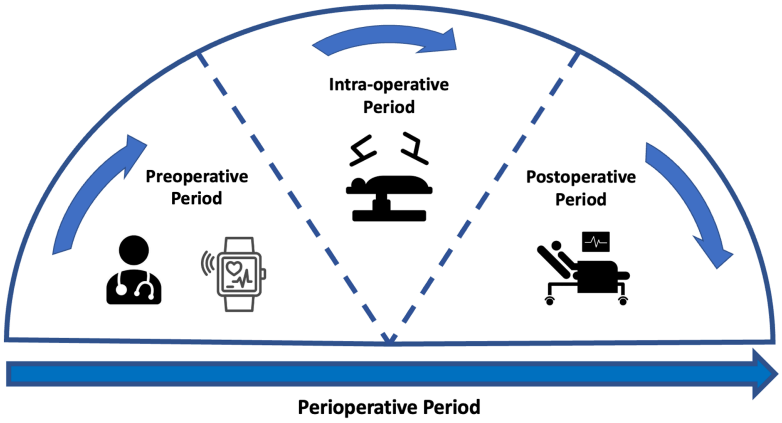


Figure 2.1: Stages across the perioperative period. The perioperative pathway refers to the period that spans from the first point at which surgery is considered as a treatment option up until the full recovery [38]. This pathway has several sub-stages [39]. The preoperative period represents the period prior to surgery where any preoperative assessment takes place. The intra-operative period is representative of the period whilst the patient is undergoing treatment. The postoperative period relates to any period immediately following the operation and can continue after patient discharge.

Physical examinations build on an assessment of the patient's medical history. These pre-anaesthesia examinations include a physical assessment of the lungs, heart function and possible evaluation of the main vital signs using a variety of tests [36, 37]. NICE guidelines provide a breakdown of recommendations for testing that vary depending on the severity of the surgical treatment and health status of the patient [37]. For example, an ECG is a common tool that has been shown to optimise risk stratification of cardiovascular complications for non-cardiac surgery [40]. However, for minor/intermediate treatments in young or patients considered healthy, a resting ECG

is not recommended as part of routine preoperative assessment [37]. These physical assessments are routinely performed at a preoperative clinic appointment in a resting state.

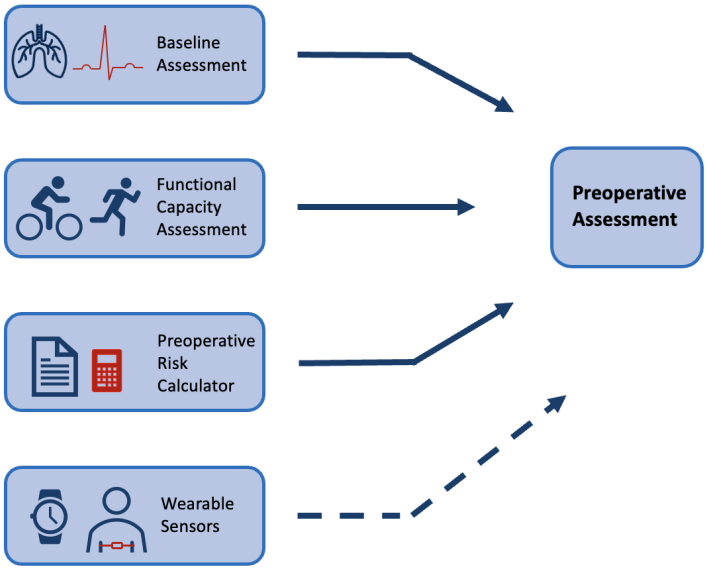


Figure 2.2: Common preoperative assessment tools used in practice. The top three boxes present common forms of preoperative assessment that are regularly used in practice (see Section 2.1.1), whilst the last box with a dashed arrow is included to show the potential for wearable sensors to be used alongside common methods in this context.

Multiple preoperative assessment tools exist that calculate patients' risk of adverse outcomes from routinely collected data. These tools are widely recommended by medical societies to be employed as a preoperative assessment tool [41]. The ASA system (American Society of Anaethesiology), the APACHE II (Acute Physiology and Chronic Health Evaluation) and the POSSUM (Physiological and Operative Severity Score forenUmeration of Mortality and morbidity) scores have been shown to have predictive value [42, 43, 44]. A review of common preoperative tools has shown that they have comparable predictive performance to machine learning (ML) techniques [45]. However, these tools are not consistently employed in practice. Lack of time and trust in the accuracy of measurements has frequently been reported by clinicians as a barrier to the implementation of risk calculators [46]. They present issues in that they can be open to subjectivity and sometimes require the input of variables that are not available in the preoperative period [47, 44]. Further, the majority were originally developed with evidence that predates the last three decades of research [42, 43, 44].

Functional capacity assessment is a measure that aims to quantify the ability of a patient to undertake activities from a free-living environment that need 'sustained aerobic metabolism' [48]. Much research has identified the association between a higher functional capacity and a reduction in postoperative complications [49, 50, 51]. The 6-minute walk test is a common exercise tolerance test but lacks accuracy [52]. In

comparison, cardiopulmonary exercise testing (CPET) is considered the current gold standard for preoperative assessment. CPET is a non-invasive clinical tool that evaluates cardio-respiratory function to measure exercise capacity [53]. In the clinic, patients undergo an exercise test following a ramp protocol on a cycle-ergometer or a treadmill whilst ventilation and respiratory gas parameters are measured [54]. Although submaximal tests are still used in some clinical contexts, maximal tests (i.e. exercising to the limit of tolerance) are generally recommended [18]. Multiple studies have been shown to support its use as a tool to identify patients at increased risk of developing postoperative complications following general surgery [55, 56]. Although CPET is a proven tool for risk stratification and is routinely implemented, there are several barriers to CPET being accessible and precise. CPET requires trained specialists to complete testing with ready access to dedicated facilities; in 2018, only 53% of Trusts in the UK offered the service [57]. Further, CPET is an expensive test with costs estimated at £289 per unit of testing in 2018/2019 (NHS Improvement, 2019). Although the test measures direct oxygen consumption, there can be considerable subjectivity with one study reporting possible miss-classification of outcomes in up to 60% of tests completed [58]. Finally, CPET might be contraindicated meaning that patients who are at high risk of complications are not always able to achieve a representative score [59]. Wearable technology has been proposed as a tool that can overcome some of the barriers that are common in these preoperative assessment tools.

2.1.2 Wearable Sensors

The development of technologies in wearable sensors (WS) in the last decade has led to a significant increase in consumer uptake [49]. As a result, there is a vast quantity of data relating to individuals' health whilst in free-living environments. There are also many examples of WS being implemented in a clinical setting as a cost-effective tool to measure physiological signals [60]. There is evidence to suggest that these devices could hold a vast volume of data that can give clinicians a quantitative representation of patients' health in their day-to-day environment [61].

WS have multiple attributes that make them a suitable tool for preoperative assessment. Physiological signals have inherent biological variation and therefore, recording these signals over a longer time period may allow detection of abnormalities that present at irregular time periods [62, 63]. A further advantage of collecting data over a longer time period is that the data may be more representative of normal routines. When collecting physical activity data, an increase in the number of days recorded is associated with a more reliable weekly estimate [64, 65]. WS are usually autonomous devices that can record signals away from the clinic. This can provide a simpler alternative for clinicians who, under time constraints, cannot always complete preoperative physical assessments [46, 66]. In some cases, measurements recorded away from

the clinical environment may be more accurate; the 'White-coat' effect describes the increases in physiological measurements that are only seen when taking measurements in the clinical environment [67]. These attributes have led to multiple successful implementations of WS for diagnostics.

WS have high efficacy for continuous monitoring of numerous physiological variables [68]. Subsequently, this has been shown to be applicable to support the diagnosis of several diseases including Parkinson's, kidney failure and viral infections [69, 70, 71]. Particularly in the case of cardiovascular disease monitoring, WS can provide live monitoring capabilities of patients' medical status that can be used to alert clinicians [72, 73]. In the postoperative period, wearable technology has shown consistent uptake and to be a particularly useful tool for monitoring recovery [74]. WS have been used to identify post-surgical cancer patients who are recovering slower than their predicted profile, allowing for earlier detection of complications and appropriate discharge planning to prevent re-admissions [75]. A wide range of wearable sensing technology has been shown to be useful for clinicians in the postoperative setting including chest patches and wrist-based fitness sensors [76, 77, 78, 79]. In the preoperative period, initial research has reported similar successful applications of WS.

Some research reports utilising WS as a method to measure adherence to prehabilitation programmes rather than as a preoperative risk assessment tool [80, 81]. In other cases, WS have been utilised specifically for preoperative assessment with an exclusive focus on accelerometer data [82, 83]. More recently, there have been instances where research has combined Heart Rate (HR) data with accelerometer data to approximate outcomes of cardiovascular fitness testing [59, 84]. These papers highlight the variation in sensor modalities and analysis methods that exist across the field, suggesting that the field would benefit from a review of these factors.

2.1.3 Aims of Review

Wearable sensors are increasingly investigated across medicine and exercise science, offering objective measurements of health in both clinical and free-living settings. In the postoperative period, multiple reviews have summarised their use in inpatient and outpatient care [74, 85]. However, prior to this work, there had been no comprehensive review of their role in the *preoperative* setting.

The primary aim of this review was therefore to investigate how wearable sensors have been applied in free-living environments to support preoperative risk assessment. To achieve this comprehensively, we considered research that sought to predict either (i) preoperative clinical measurements (including cardiorespiratory fitness, a central focus of this thesis) or (ii) postoperative outcomes, which remain closely linked to preoperative risk stratification. Including both groups of studies allowed us to capture the range of data modalities, pre-processing pipelines, feature extraction techniques, and

modelling strategies that have been trialled in this context that otherwise might have been left out.

Although the experimental chapters of this thesis focus specifically on VO_2max prediction from wearable sensors, this broader review remains directly relevant: it provides a field-wide perspective on available sensor modalities, analytic methods, and common challenges, all of which inform the methodological choices made in subsequent chapters. Research investigating emergency surgery will not be included as preoperative evaluation for emergency surgery does not allow analysis of free-living data. Additionally, the review will focus on major abdominal surgery; research completed in a cardiac or orthopaedic surgical setting will be excluded as these procedures are associated with different complications to general surgical procedures [86]. Although this is not a systematic review, search terms were employed to identify research from selected databases and a narrative synthesis was used to summarise findings. The findings of this review are presented across four sections: Sensor Modalities, Pre-processing, Feature Extraction and Predictive Models.

2.1.4 Literature Search

Table 2.1: Combination of search terms used for the review of the literature. These searches were combined with Boolean operators and entered into the databases MEDLINE, Web of Science and Google Scholar in equal format. Initial investigation of search terms was completed to find the combination of terms that returned optimal results. A narrative review of results is completed in this paper.

Surgery	**Preoperative Assessment**	**Wearable Sensor**
major surgery general surgery abdominal surgery elective surgery	preoperative pre-surg* preoperative evaluation	wearable technology wearable activity monitor heart rate monitor accelerometer fitness tracker wearable fitness*

*The asterisk is used for truncation in the search. The asterisk is added to end of a term to allow the databases to search for all forms of the word to broaden the search.

Search Methodology

To identify key literature and ensure valuable research was not missed, search terms were identified and used to search major databases. MEDLINE (Ovid) and Web of Science were searched as well as the first 200 returns in Google Scholar [87]. Relevant papers from ARXiv were also included. Search terms can be seen in Table 2.1. Further articles were identified through backward chaining. The literature search presented here was originally conducted in late 2023 and subsequently published in early 2024.

More recent contributions to the field have been incorporated in this chapter where they are directly relevant to the research aims.

Study Cohort

Searches brought back a broad range of papers from which the most relevant were selected. Several inclusion criteria were outlined for the search. Research should analyse free-living data collected from WS in research. Data should be used to investigate the association of these signals with outcomes related to either clinical variables routinely collected preoperatively or postoperative outcomes. Clinical variables collected preoperatively varied from CPET outcomes to cardiovascular responses. A subsection of these papers was selected for in-depth analysis whilst multiple papers outside of this subset are referenced throughout this review. A summary of the sample sizes and participant demographics can be seen in the appendix A (A.1). A compilation of the key features extracted from these papers selected for in-depth analysis can also be found in the appendix A (A.2). Sample sizes varied greatly across studies and had a range of 16 to 80,137. For research that had a sample size of under 1000, the mean sample size was 48.9 indicating that the majority of research in this field has a relatively small number of participants. This is likely due to having to provide hardware to each participant included in the research study rather than having access to pre-compiled data sets. There was also a broad range in the sex split across research; there was a slightly higher prevalence of male participants with an average of 57% males but this was not significantly imbalanced. Ethnicity was rarely reported with under 30% reporting the ethnic breakdown of participants. When ethnicity was reported there were significant imbalances with largely white participants.

For research collecting data from patients waiting to undergo surgery, participants were commonly approached immediately after being enlisted for surgery or at the preoperative evaluation clinic [88, 89, 90, 91, 92, 93]. It was also common for patients to wear the device right up until the date of operation [94, 95, 96, 97, 88, 89]. For all research apart from two papers, the data collection from the WS device took place within 33 days prior to surgical treatment. In two cases, it was unclear how far in advance of treatment patients wore their WS [82, 98].

Applications of Wearable Sensors

Across the reviewed literature, wearable sensors have been applied to a range of related research goals in the preoperative context. Many studies focused on characterising preoperative physical activity behaviour, using wearable-derived measures such as step count or activity intensity to examine associations with postoperative complications, length of hospital stay or readmissions. Other studies aimed to compare activity

or physiological profiles between patient subgroups, for example comparing complication rates between in-active and active groups rather than direct outcome prediction. These approaches reflect an interest in understanding how preoperative functional status relates directly to surgical risk.

A different subset of studies investigated the relationship between wearable-derived features and functional capacity. Although several proxy measures were used including walking tests, functional capacity was most commonly measured and reported using VO_2max. While some of these studies were conducted directly in preoperative cohorts, related work in non-preoperative populations was also included in this review. This was motivated by the clinical importance of cardiorespiratory fitness as a reference measure for surgical risk stratification and the relative rarity of preoperative wearable studies addressing this outcome directly. Taken together, the reviewed papers span multiple application domains and sensor modalities. This heterogeneity motivates the structured methodological approach adopted in this literature review.

2.2 Hardware/Sensing Technologies

This review identified a range of different devices that have been utilised in the research in the preoperative setting. This paper does not give an analysis of each commercial device but presents an overview of the different sensor modalities that they employ to collect data. The accelerometer, ECG and PPG sensors were widely implemented; Figure 2.3a shows the breakdown of these sensor modalities across research. A full breakdown of the WS devices and models along with their sensor modalities can be found in the appendix A (A.2). The following sections outline the functionality of these sensors and their outputs.

2.2.1 Accelerometry

Raw accelerometer data is recorded as a signal of acceleration measured across axes; this can be analysed to detect regular patterns that represent movement [99]. This is most commonly measured across three axes in a tri-axis accelerometer: X, Y and Z (see Figure 2.4a). An accelerometer will measure linear acceleration across these three axes and by combining this with time, the signal can be used to quantify human movement and physical activity [99]. Specifically, movements including steps can be identified to track a step count whilst the general movement of the sensor can also be categorised into intensities of movement. The association between features extracted from an accelerometer with several health metrics has consistently been evidenced in research. Step count is one of the most commonly extracted features from accelerometer data (see Section 2.4.1) and has been linked to various health outcomes. A sys-

tematic review of over 13,000 adults identified that an increase in step count of 1000 daily steps above baseline was associated with a lower risk of all-cause mortality and cardiovascular disease (CVD) [100]. Similarly, there is a strong relationship between any recorded physical activity and a reduced risk of all-cause mortality [101]. Further, accumulated time spent in moderate to vigorous physical activity (MVPA) is linked to significant reductions in mortality risk [102]. This suggests that the collection of these variables prior to surgery may have a strong association with a patient's health and could be useful for predicting complications.

The majority of research in the preoperative period utilises WS from commercial providers that have their own internal proprietary systems. As a result, research rarely analyses raw accelerometer data in the preoperative period. Multiple projects utilised a Fitbit wearable from which they could extract step count and energy expenditure [103, 83, 88, 92, 89]. Similarly, multiple studies utilised Garmin-derived metrics including step count, distance travelled and energy expenditure [104, 94, 59]. Data generated from a combination of mobile devices and WS that had a common platform (Apple Health; Achievement) extracted similar variables [105, 106]. Particularly amongst large-scale research with many participants, these pre-calculated features were utilised for analysis rather than raw data [106, 107]. Some research reported utilising pedometers as their WS to collect patient data; these pedometers all included accelerometers rather than a traditional step counter allowing them to represent the intensity of movement [108].

Although uncommon, some researchers did utilise the raw accelerometer data from the devices. Three studies utilised raw accelerometer data to stratify the signals into different intensities of activity [82, 90, 109]. This allows researchers to quantify the intensity of movement and compare time spent in sedentary behaviour against higher intensities including MVPA. Figure 2.3b presents a breakdown of the location at which sensors were worn. The majority of all sensor devices, and therefore accelerometer sensors, were worn on the wrist in commercial devices (Fitbit/Garmin) [94, 104, 98, 89]. However, many of the research papers that only included an accelerometer sensor were worn on the participant's hip, see Figure 2.3b [97, 95, 96, 82].

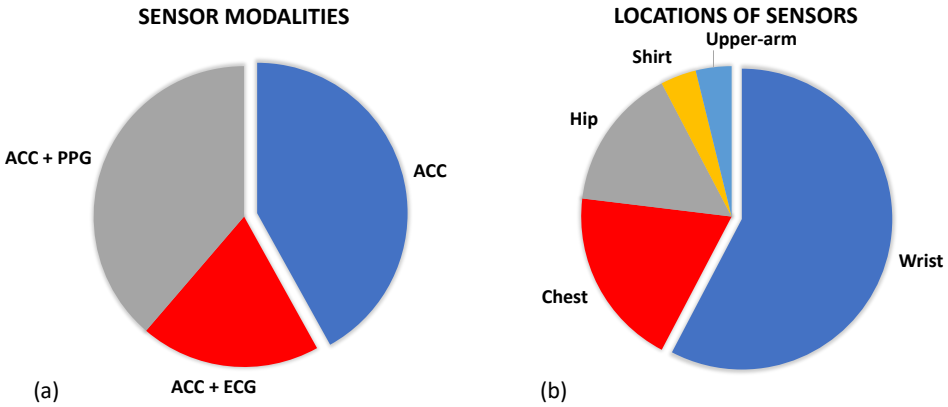


Figure 2.3: Sensor modalities and their locations. (**a**) The percentages of sensor modalities used across research. All research employed accelerometer sensors but a further subsection combine this with either ECG or PPG sensors. (**b**) Variation in locations of sensor types. The common locations for sensors used in research applicable to the preoperative period are outlined.

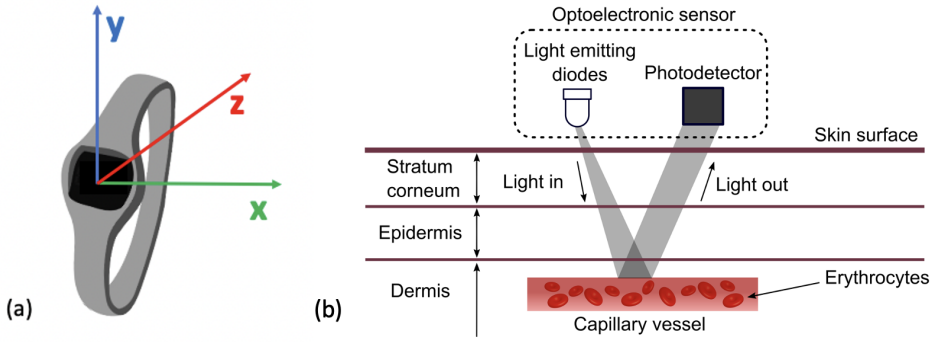


Figure 2.4: (**a**) Reference axes in a Tri−axes accelerometer. Presents the axes along which acceleration of movement can be measured across x, y and z. (**b**) The mechanism for HR detection in a PPG sensor by reflection. The LED can be seen emitting light which is reflected and then detected by the photo-detector and converted into a HR signal. This figure was taken from Moraes et al. (2018) with no changes made, Creative Commons Attribution International 4.0 License [110, 111].

2.2.2 Photoplethysmography

Photoplethysmography (PPG), sometimes referenced as optical heart rate monitoring, is a sensor that can estimate an individual's heart rate (HR). It utilises an optical emitter to give out light emitting diodes (LED) onto the skin that is attenuated from the pulse in the artery [112]. The reflection of the LED is captured by a photo-diode. The digital signal processor located in the device then translates this into heart rate data, as seen in Figure 2.4b. HR is one of the most commonly measured vital signs across medicine. It holds significant prognostic value for predicting general health and mortality. A lower HR is generally associated with a lower risk of cardiovascular mortality as well as a

lower risk of all-cause mortality [113, 114]. Research has also identified that a lower preoperative HR is associated with a lower risk of postoperative myocardial injury in patients undergoing non-cardiac surgery [115]. Therefore, cardiac assessments are common practice prior to non-cardiac surgery [116].

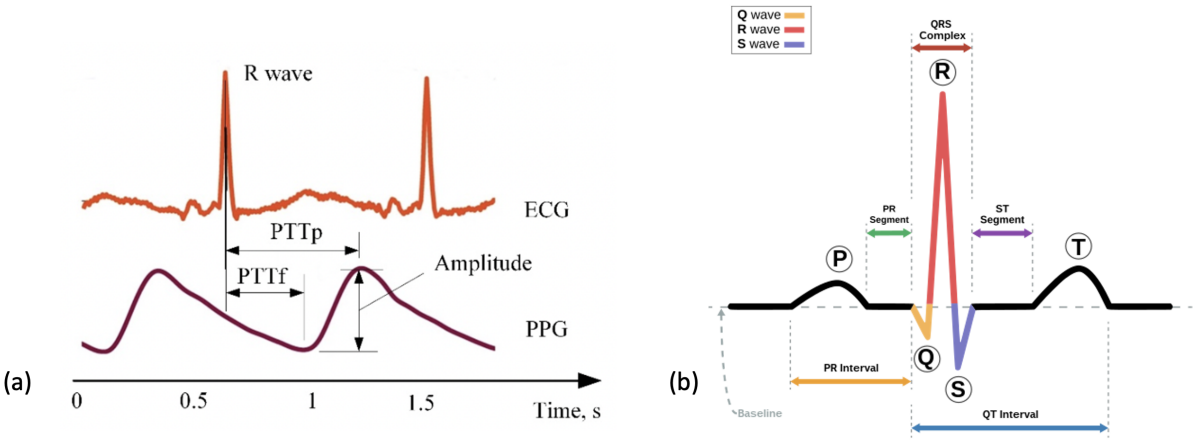


Figure 2.5: Example recordings produced from ECG and PPG recordings. (**a**) A comparison of the cardiac signals produced from a PPG versus ECG sensor over a period of 2 s. This figure was produced by Elgendi et al. (2019) and was taken from a larger figure with no changes made as part of the Creative Commons Attribution International 4.0 License [117, 111]. (**b**) A segment of an ECG graph that has been portioned to show the stages in a normal cardiac cycle including the P wave, the QRS complex and the T-wave.

Previous research has investigated whether PPG produces comparable outputs to ECG. PPG sensors require both the LED and photo-detector to have contact with the surface of the skin and as a result, can be heavily impacted by movement or distance between the LED and skin resulting in optical noise [118]. A comparison of PPG to ECG signals can be seen in Figure 2.5a [117]. A large-scale study evidenced that HR estimates collected from PPG sensors correlate strongly with those from ECG signals [119]. Other research has concluded that at rest and at low HR levels, PPG has shown to have high accuracy but that this decreased with intensity of activity [120]. One publication reported the threshold for a reduction in accuracy of HR estimation to occur between 155-160 beats per minute [112]. This high threshold is notably above estimated HR maximum values for elderly populations indicating potential suitability for this population and for monitoring low-intensity exercise [121]. Most WS that include accelerometry at the wrist also include a PPG sensor [88, 92, 83, 89, 103, 93, 106, 122, 59]. One popular device that included a PPG sensor reported a data storage limit of 7 days and a charge limit of 10 days within the device [91]. PPG is a widely utilised technology in the preoperative period due to its convenience as a tool to measure HR. Wrist-based wearables have few requirements and are a practical tool for researchers as they require little input from users and have long periods of storage.

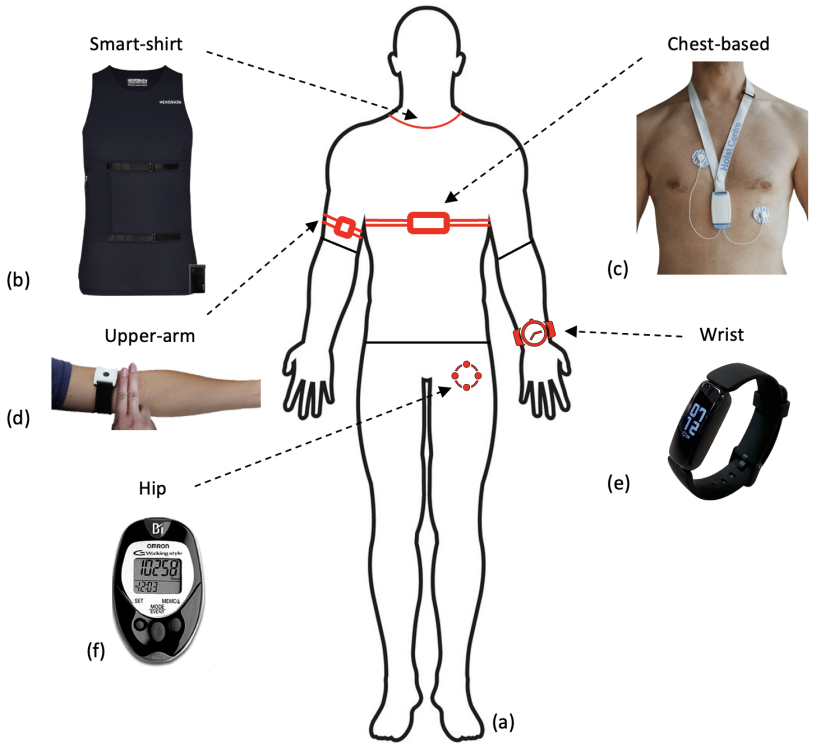


Figure 2.6: Wearable sensor devices used in research across the body (**a**) and where these are located (**b**–**f**). (**b**) The Hexoskin smart shirt that collects both ECG and activity data, used with permission from Hexoskin [123]. (**c**) An ECG wearable device that collects recordings from a single-lead ECG device and 3D-accelerometer data, used with permission from [124]. (**d**) An upper-arm PPG sensor utilising reflective PPG detection, similar to that used in preoperative monitoring research [125]. The figure is taken as part of a larger figure from Wang et al. (2023), Creative Commons Attribution International 4.0 License [120, 111]. (**e**) The Fitbit Inspire collects a combination of accelerometer and PPG data from the user and is commonly used in preoperative research. The figure is taken from Li et al. (2023), Creative Commons Attribution International 4.0 License [126, 111]. (**f**) The OMROM walking style pedometer that utilises a tri-axis accelerometer to collect step data, used in predicting VO_2max [96]. This figure is taken from Bartlett et al. (2017) as part of a larger figure, Creative Commons Attribution International 4.0 License [127, 111].

2.2.3 Electrocardiography

An ECG directly measures the electrical activity of the heart through electrodes placed upon the skin [128]. The electrodes measure electrical impulses from the heart that are then converted into an ECG graph. An ECG graph segment with annotations can be seen in Figure 2.5b. These annotations represent the detection of different stages in the cardiac cycle: the p-wave representing depolarization of the atria, the QRS complex representing the electrical impulse spreading to ventricular depolarization and the T-wave representing the re-polarization of the ventricles following contraction. The location and morphology of these annotations are used to extract several features including heart rate (HR) and heart rate variability (HRV) [129]. Using these features amongst

others, an abnormal cardiac cycle can be identified from a signal and categorised. Heart disease is a leading cause of death worldwide making detection of cardiac abnormalities from the PQRST complex a key tool for preliminary diagnoses [130]. In clinical practice, a 12-lead ECG is common practice but this can be impractical for WS in a free-living environment. In this review, reduced lead ECG devices are considered wearable devices [131]. By analysing the ECG graph, HR can be calculated at any given time point. Additionally, using the RR interval (time portion between each R peak) other variables including heart rate variability (HRV) can be calculated.

An ECG sensor has several advantages as a prognostic tool compared to PPG sensors. The ECG is considered the 'gold standard' tool for measuring HR; the accurate identification of a heartbeat allows the calculation of HRV from which further inferences about health can be made [132]. HRV has been highlighted as a tool that has promising value for predicting complications during and after surgery [133]. Additionally, an ECG allows for further detection of potential abnormalities including atrial fibrillation [134]; preoperative atrial fibrillation has shown to be predictive of complications in patients undergoing non-cardiac surgery [135, 136].

One large cohort study used the Actiheart wearable ECG, which places two leads on the sternum from which three papers analysed the HR data [107, 84, 137]. Other research utilised an ECG 'necklace', which involved placing electrodes in the II lead configuration on the chest (see Figure 2.6c) whilst a further project included an ECG sensor that was integrated into a smart shirt (Hexoskin), see Figure 2.6b [138, 139]. Although not utilising all 12-leads, wearable ECG devices have been shown to have high accuracy for heartbeat detection [128].

2.3 Pre-Processing of Signals

Pre-processing involves all changes to data that are made in order to prepare the data for analysis. Pre-processing can be the most vital stage in data processing and has a large impact on the inferences that can be made from a data set. Wearable data, even when collected in a controlled clinical environment, often requires heavy pre-processing due to the nature of the data. A wide range of pre-processing methods were implemented across the key papers and the most important techniques for each are outlined in the appendix A (A.2). There are two main challenges in WS data that pre-processing aims to overcome: missing data and noise.

2.3.1 Missing Data

Missing data is a frequently reported problem across research involving WS, particularly when using data from free-living environments [140]. Poor electrode placement,

poor contact with skin or removal of device might lead to significant portions of poor quality or missing data. Often, the underlying reasons for periods of missing data are unknown. The prevalence of missing data in WS used in the preoperative period is outlined.

There is a wide variety of missing data reported across studies using wearable sensors in the preoperative period. Missing data was frequently reported at ranges of up to 25% from WS in this context [125, 88, 89]. One study reported that across all accumulative days of collected data, only 0.25% of days had complete HR data [89]. For the majority of research, the reporting of missing data refers to HR rate, rather than movement data (see Section 3.1.1). The reporting of missing data differed between research; some researchers report the overall percentages of data that were missing whilst others report the number of participants excluded due to missing data [59]. In both of these cases, research rarely goes into detail as to the causes of missing periods and how to categorise these. Missing data can generally be classified into three separate categories: missing completely at random (MCAR) where no systematic relationship is present between values that are missing and existing values; missing at random (MAR) where missing data is systematically related to existing data that has been observed but not unobserved data and missing not at random (MNAR) where missing data is systematically related to unobserved data [141]. Depending on the category of missing data that is assumed for the data, different methods may be better suited for minimising the potential bias that may be introduced [142].

Across research applicable to the preoperative period, three different strategies for handling missing periods of data were employed. As seen in Figure 2.7, missing periods of data were either deleted entirely, tolerated or imputed. These techniques have been previously identified in research using WS and are common solutions for missing data across fields [141, 143]. To build on this, several techniques were identified in the present review that involve overlap between categories. To differentiate between data that has missing portions but is still usable versus data that should be deleted, an extraction threshold can be identified. Further, some research employs imputation on only short-term segments of data (see Section 2.3.1).

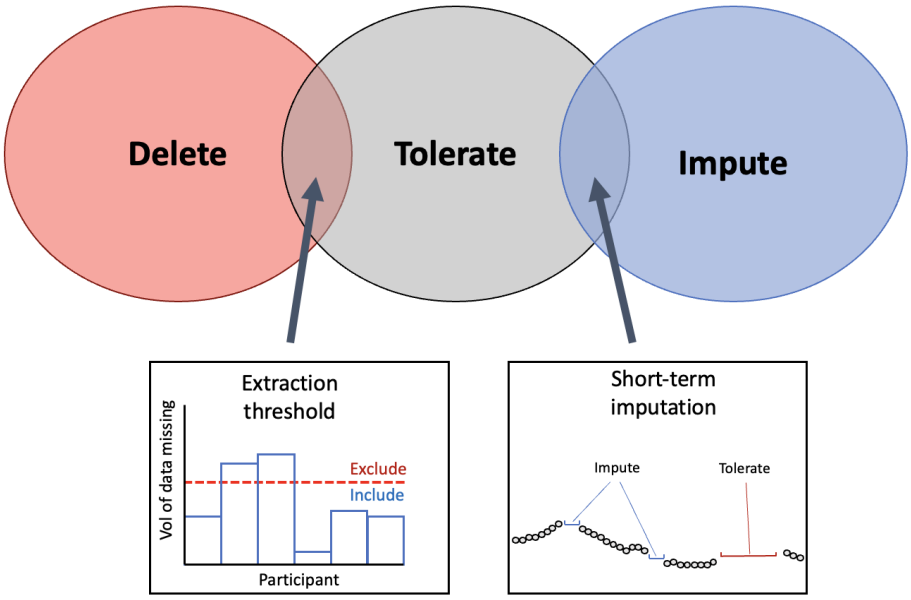


Figure 2.7: Venn-diagram to present the common methods for handling missing data from WS. At the intersection between 'delete' and 'tolerate' the implementation of an extraction threshold was identified to delete data below the threshold and tolerate missing data above the threshold. At the intersection between 'tolerate' and 'impute', imputation on short-term segments of missing periods was identified as a solution that employs that imputation on select segments.

Extraction Threshold

The extraction threshold identifies a point at which a subject's data will be included in final analyses or is abandoned/processed further. This extraction threshold is usually only applied to wearable devices that measure HR in some format. For research that only utilised the pre-extracted step count, it is not possible to assess the exact volume of missing data. Step count data can appear as null values and still represent viable recordings indicating sedentary periods and so it is not always obvious to know whether this is as a result of non-wear, device malfunction or sedentary behaviour [144]. This is particularly true when step-count is only reported at the daily level [94, 97, 98].

These extraction thresholds differ widely between research. One study set a daily yield extraction threshold at a minimum of 8 h of collected data for that day to be included in analysis [89]. Other research set their extraction threshold at 10%, defining that any day with data of a daily yield above 10% would be included for analysis [88]. These studies indicate that the daily extraction threshold can be set at a relatively low value to allow for a high level of missingness in data and prevent this data from being abandoned. Research that used large data sets could set their extraction thresholds at a higher level; one study with over 80,000 participants only selected participants that had a minimum daily yield of 20 h [84]. However, this study did not report what percentage of participants had to be excluded as a result. Large data sets may have

more flexibility in their extraction thresholds whilst a small research study may have to accept a higher level of missingness in order to prevent excluding a large portion of their sample.

A total yield extraction threshold can also be applied to the number of days in the monitoring period that have data [125]. This can be employed by only including participants that have above x number of days of data, with x indicating the threshold. The employment of a daily extraction threshold (i.e., 8 h of data) versus an extraction threshold for total data yield (i.e., 3 days of data needed) will depend on whether the data are subdivided into daily segments or kept as a total per participant.

Selecting an Extraction Threshold

Extraction thresholds should not be randomly selected. To investigate the influence of the extraction threshold on the predictive performance of analysis, one study varied their extraction threshold from 1 to 10 h and identified that between 8 and 10 h achieves the best performance [89]. This highlights the importance of identifying an optimal extraction threshold. Setting a high threshold for inclusion will result in less data available for analysis; a low threshold has the potential to allow data from days with large missing periods into analysis. If this is the case then the underlying reasons for missing periods should be assessed to prevent bias in the data (see Section 3.1).

Aside from reported extraction thresholds within their own data sets, very little research has focused on quantifying the volume of data that is needed to obtain reliable preoperative baseline measurements. It has been reported that when using a PPG sensor combined with an accelerometer, a minimum of three days of monitoring should be completed; however, an extraction threshold within each day was not specified and whether the location and type of sensors have an impact on this threshold was not discussed [145]. An appropriate extraction threshold is a useful tool for selecting data for analysis but does not provide a solution to missing periods. To overcome the missing periods of data that remain, several imputation methods can be employed.

Imputation

Data imputation in WS data is a complex process. Individuals will often have varying levels of missingness between them due to compliance with wearing the device. Further, there may be missing portions caused by technical issues in a device. Therefore, imputation techniques in WS data should generalise to both the participants' behaviour and the device's patterns [146]. Often in research collecting data in the preoperative period, data that was identified as missing was abandoned. Little effort is made to impute the missing values and why they are missing; this could reduce the size of the data sample and in some cases may introduce bias [146]. The imputation techniques that were employed in the research are outlined in the sections below.

The simplest method to replace missing data points in HR signals is with the mean HR values of activities at waking periods [106]. However, if the mechanism for missing portions is known to occur during sedentary periods or periods of vigorous activity then this may lead to under or over-estimations of daily HR. One particular study substituted missing HR values with HR recorded during a hospital visit [91]. This technique was not common across research, likely due to inconsistencies that may be present between free-living data and data collected in-clinic (see Section 2.1.2). For the studies that utilised the temporal aspects of data to employ imputation, they both did this using a two-layered pipeline [89, 88].

The k-nearest neighbours (KNN) technique was shown to be a common method to impute missing HR values [89, 88]. KNN has previously been implemented as a technique to address missing data in a range of applications [147]. The KNN algorithm is implemented as a 'sliding window' that allows missing HR data to be calculated from a combination of recent step count and HR data [88]. This method is rationalised by explaining that imputation is useful for short-term missing segments where previous values of step count and HR have a high correlation with future values. One technique employed a k-nearest neighbours (KNN) algorithm for an entire day of data if the daily yield for the relevant day was above 10% [88]. A different study utilised the same technique but for all portions of missing data that were shorter than 10 min long, regardless of daily yield. If the segments of missing data were less than 10 min, a KNN sliding window (length of 5) utilised recent HR and step counts to predict HR values [89], see Figure 2.8. Using this method, the feature vector is imputed to the KNN algorithm where hr_t and $step_t$ represent HR and step data at time t.

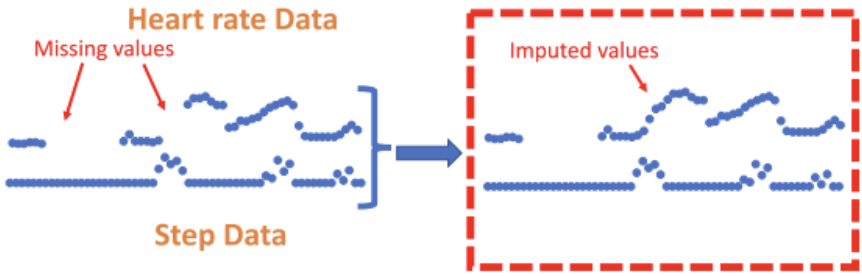


Figure 2.8: Imputation using K-nearest-neighbours. Zhang et al. (2023) utilise the KNN technique to impute on short-term segments of missing data under 10 min in length by utilising previous values from both the step count and heart rate signals to calculate missing values. This figure was produced by Zhang et al. (2023) and was taken from a larger figure but had no changes made, taken as part of the Creative Commons Attribution International 4.0 License [89, 111].

Feature Level Imputation

In instances where data has been abandoned due to significant periods of missing data, this can be imputed by employing feature-level imputation techniques. After aban-

doning days with a daily yield below the extraction threshold, a feature-level imputation technique can be employed to compute the features that represent the days with high portions of missing periods. In one case, researchers again utilise the KNN method to impute statistical and semantic features based on the neighbouring features that are available for that participant, rather than attempting to impute the missing values in the signal [88]. A further technique to deal with missing data that fell below the daily yield of 8 h was to employ imputation on high-level features that have been calculated from daily features [89].

It is important to note that research only employed imputation for HR data, this was not performed for other features extracted from accelerometer data. For research where HR signals are collected alongside step count data, the proportion of missing step count data can be extrapolated from the time periods with missing data points of the HR signal. Step count data is reported as being significantly less correlated and so less predictable than HR data [89]. Instead, these data were normalised by dividing the step count by the daily yield so as to prevent the step count from being drastically increased just for those patients with more data accumulated.

2.3.2 Noise

Aside from missing data in the signal, noise can also prevent meaningful features from being extracted. Accelerometry data can be plagued by white noise, altered by human motion or vibration whilst both ECG and PPG signals can be corrupted by motion artefact, baseline wander and electromyography (EMG) noise [148, 149]. Few papers utilised raw signal data (see Section 2.2) meaning filtering of signals was not commonly reported. When techniques were employed to filter noise, this was performed on the accelerometer and HR data separately.

In the present review, many WS devices only report pre-calculated HR values from internal algorithms meaning processing of raw ECG or PPG signals was not reported. In certain cases, some removal of noise from HR signals was completed. Previously, Gaussian process robust regression has shown to be successful for noisy HR data and was implemented by one study to utilise prior knowledge of the HR data to reduce noise [84, 150]. In comparison, a simpler method to limit the noise in HR data was to average the HR extracted from R-R intervals over a set time period, this varied from between 15 s to 15 min [138, 137]. When employing this technique, all inaccurate HR values were identified and removed from the data where consecutive HR values varied by more than 20% [138]. Cardiac signals were very rarely passed through a low-pass filter; however, one study resampled the HR to 1 Hz before passing HR through a 0.01 Hz low-pass filter to remove high-frequencies affected by non-linearities introduced from circulatory distortions [139].

For accelerometry, to convert the raw signal data into magnitude of acceleration the

Euclidean norm minus one and high-passed filtered vector magnitude were used [84]. Altini et al. (2016) employed a different filter technique where a low-pass filter (1Hz) was used to isolate the static component in the signal due to gravity and a band-pass filter (0.1 Hz, 10 Hz) was used to isolate dynamic components due to body noise [138]. As mentioned above, Beltrame et al. (2017) used a similar low-pass filter at 0.01 Hz for accelerometer as well as HR data [139]. The only implementation of a fast Fourier transformation (FFT) was to integrate the frequency in accelerometer data between 1 Hz and 10 Hz [109]. One particular paper reports a method for outlier detection within data by removing values that are greater than 3 standard deviations from the mean [122]. Building on the techniques commonly employed on HR data by averaging values over a short period, research employing a pedometer categorised each accelerometer period of 10 s into either lying, stationary or active periods [109].

For research using large-scale cohort data, after normalising their data through standard scaling with unit variance, researchers applied Principle Component Analysis (PCA) onto the original training data set that retained the components that explained 99.9% of variance [107]. To prevent any information leakage across the data sets, the fitted PCA scaler was applied individually to the test set. In a different project utilising the same large dataset, researchers attempted to reduce the noise that is seen in the labelling of HR data from the ECG wearable using deep learning [137]. The authors propose UDAMA (Unsupervised Domain Adaptation and Multi-discriminator Adversarial) training network [84]. However, these techniques utilising deep learning require a large pool of data and may not be suitable for smaller single-centre studies. In order to reduce the noise that is present in daily features, one research study utilised singular spectrum analysis to further extract high-level features from daily features [89]. This allows trends to be extracted from the noisy data with missing portions by computing the mean, variance and slope from each time series of daily features.

2.4 Feature Extraction

Feature extraction is an important step in signal processing to convert the signal data into numerical features that can be processed in a model [151]. It can also be useful to reduce the dimensions of the data when a large amount of data is collected [152]. Most research performed feature extraction from each signal separately but where possible features were extracted from multiple signals. The resulting features are outlined, and a full breakdown of the relevant features extracted from selected papers can be found in A.2 in the appendix.

2.4.1 Features Extracted from Accelerometer Signals

Step Count

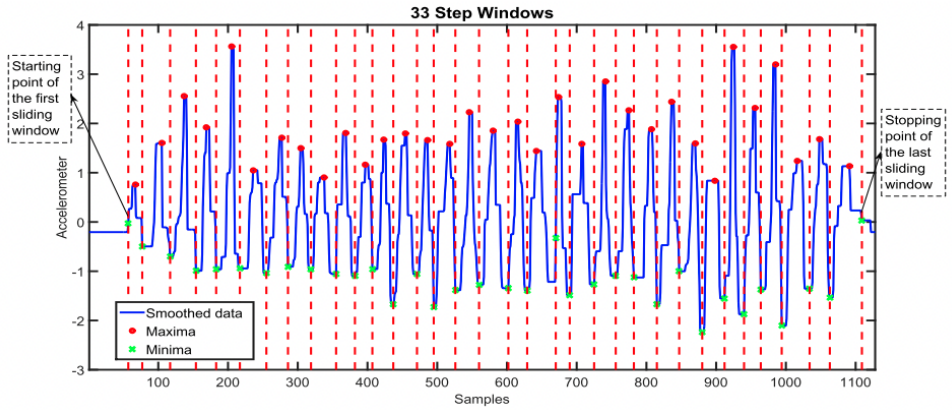


Figure 2.9: Implementation of a maxima and minima step-counting algorithm that counts step number based on the number of steps windows detected. Each red line indicates the stopping point of each step window and the start of the next corresponding window, the length of time between each red vertical line indicates the step window size [153]. This figure was produced by Ho N et al. (2016) and was taken with no changes made, used as part of the Creative Commons Attribution International 4.0 License [153, 111].

Step count is a frequently reported feature that is used in WS research. Extracting step count from raw accelerometer signals involves implementing an algorithm to detect a pattern in the data, the choice of algorithm might depend on the computational complexity and resources that are available. Common algorithms include the peak detection algorithm, which identifies the local maxima and minima (see Figure 2.9) or simpler thresholding algorithms that identify optimal thresholds to detect steps [154, 155]. Machine learning techniques also have shown good application but may require labelled data to train models, which are not often available [156, 157].

Much of the research that utilised commercial devices employed internal proprietary resulting in pre-calculated features. If the raw signal is automatically converted into step count then any further features relating to the intensity of movement are not able to be extracted. Frequently, steps are reported as an average measure across all days meaning any useful temporal aspect of this data will be lost [94, 104, 98, 95].

Patients can be further classified into groups based on their average step counts to then produce features labelling individuals as active/inactive. Often research used pre-determined thresholds to classify participants into these groups. Threshold values range from 2500 to 5000 daily steps and there is no consensus on the threshold that should identify an individual as ‘active’ [98, 95, 91]. Another technique is to define these thresholds based on the variance of step count within the research cohort so that the split of participants in each group is even [97]. It is not conclusive which of

these methods to stratify patients into activity groups is most conclusive but this should be relevant to the research cohort and context of the research [158].

Movement Intensity

Where raw signal data is available, the intensity of movement can be calculated from the acceleration and as a result, the time spent in sedentary behaviour (SB), moderate physical activity (MPA), moderate to vigorous physical activity (MVPA) and vigorous physical activity (VPA) can be reported [82, 90]. This is calculated from time spent at activity counts above a specified threshold. Activity counts are calculated from ActiGraph's proprietary algorithm; this algorithm has been widely used across research employing accelerometers and has been published as open access software [159]. However, the selection of cut-off points for different intensities of physical activity results in significant differences in total MVPA between research and there is no standardised cut-off threshold [160]. Therefore, care should be taken when selecting a cut-off that is suitable for the research population.

Distance Covered

By utilising the features that can be extracted directly from the accelerometer signal, further features can be inferred. The distance that is covered by a participant can be calculated from a combination of the number of steps times that are taken in a day multiplied by the stride length of the participant [83, 96]. It must be noted that stride length should be adjusted for using further participant information including sex and height information.

2.4.2 Features Extracted from Cardiac Signals

Heart Beat Detection

HR signals are complex and vary depending on sensor modality. As outlined previously, PPG and ECG signals are the most commonly collected cardiac signals that are used. To calculate HR from these signals, a process of data cleaning and beat detection is employed. The Pan-Tompkins algorithm (PT) is the most widely used beat detection algorithm which has been shown to be capable of detecting the location of a QRS complex in the signal across both clean and noisy data [161]. The PT algorithm employs a band-pass filter to isolate the relevant frequency before using a combination of thresholding and dynamic adjustment to identify the R-peaks. Other popular beat detection algorithms include wavelet-transform-based methods that analyse wavelet coefficients or simple algorithms that look for peaks in local maxima [162]. However, recent research has shown that the Neurokit and New South Wales (NSW) algorithm

have shown to outperform other open-source ECG beat detectors, with the Neurokit algorithm showing considerably faster processing times out of the two [163].

Heart Rate

By using the locations for the QRS complex that have been detected, see Figure 2.5b, HR at any given time can be calculated. As previously mentioned, the majority of research that is applicable to the preoperative period utilises commercial devices where HR values are often pre-calculated from the detected beats using an internal proprietary algorithm in the WS. As a result, it is rare that research in the preoperative period has to detect the location of a QRS complex. Instead, HR values are given at a varying frequency. The update period of the heart rate signal will dictate how regularly the heart rate is updated, commonly this is updated every beat. As a result, HR signals from WS are often extended signals that require further processing to extract meaningful features.

Resting HR

Resting HR (RHR) is a term that does not have a consistent definition but generally refers to the HR of an individual when they are inactive [164]. RHR is widely considered an important bio-marker of physical health and has been shown to be associated with both mortality and morbidity after non-cardiac surgery [165, 166, 115]. Although RHR is widely accepted as an important biomarker, there are also no set guidelines in medical literature for calculating RHR. Recent literature has suggested that when employing WS in research, a minimum four-minute rest time is required for a reliable RHR measurement, and that this should be measured between the hours of 0300 and 0700 [164]. One research paper reported calculating resting HR over a 24-hour period but made a distinction between resting HR recorded during the night [91].

HR Changes

From the 24-hour period, time spent in different HR zones can also be extracted as an indicator of activity throughout the daily period [91]. Other research utilised HR signals to create a new variable by assessing the difference between a current HR value and a previous value at a 1 s lag to represent 'dynamic changes' in HR and cardiac activity [139]. Other research also employed a two-level feature extraction pipeline where first-order statistical features like skewness and kurtosis of the HR are extracted and high-level features are then taken from the daily level features including the slope, mean and variance [89].

HR Variability

HR variability (HRV) is a metric that is calculated from the variations in intervals between detected heartbeats. It is an accepted tool for measuring the function of the autonomic nervous system, a vital factor in cardiovascular health [167]. Recent research has confirmed that preoperative HRV can be a useful predictor of postoperative outcomes [133]. However, in the preoperative setting, HRV has traditionally been calculated in a clinical setting using only ECG signals from a short recording period (e.g. five minutes) rather than employing WS [168, 169, 170]. The only paper of note to incorporate HRV measures into CRF predictions used a Holter-ECG device and reported that the only measure to be independently associated with VO_2max was the Standard Deviation of N-N intervals over 24 hours (SDNN24), however this measure is yet to be investigated in the preoperative setting [171].

Recent WS research has shown promising efficacy in calculating HRV from noisy signals in both PPG and ECG signals [172, 173]. This highlights the potential for HRV to be used as a preoperative tool calculated from WS. One study did calculate HRV from participants' ECG data by differentiating the second-shortest and the second-longest inter-beat intervals [84]. There are variations in how HRV is calculated depending on whether they are time or frequency domain features. A common time domain frequency measure is the square root of the mean of the sum of the square of differences between NN intervals (RMSSD); both Garmin and Fitbit devices employ RMSSD to measure HRV in their commercial devices [174, 175]. Although these devices may be able to report HRV, it is not always available to be extracted for use in research; one particular study reported that HRV, although measured by the sensor, was not able to be extracted from the device for use in research [91]. RMSSD is presented in Equation (3) where NN_i is the length of time of the *i*th interval and N is the total number of NN intervals in the dataset.

$$RMSSD = \sqrt{\frac{\sum_{i=1}^{N-1}(NN_i - NN_{i+1})^2}{N-1}} \tag{2.1}$$

Frequency domain features are calculated using estimation of power spectral density and include a range of features such as low frequency (lf) and high frequency (hf) [176]. Further non-linear properties of HRV can be analysed and extracted using Poincare Plot and Sample Entropy; however, these are less frequently calculated.

2.4.3 Multi-Modal Sensor Feature Extraction

HR Recovery

Where possible, a combination of signals can be used to form features. HR recovery was calculated by measuring the decrease in HR one minute after exercise cessation

with accelerometer data used to identify when activity ends [83]. HR recovery has been shown to predict cardiovascular health; however, this measure typically requires either a controlled exercise test or clearly defined cessation of activity, which can be difficult to identify in noisy, free-living data [177].

Respiration Rate

Respiration rate is sometimes collected by wearable sensors, most commonly estimated indirectly from PPG or ECG signals, or using chest/torso accelerometers [178, 179]. Although respiration rate has been identified as a potential marker of health status in other clinical settings, its use in preoperative wearable research remains rare. In the small number of studies that have attempted to estimate RR, methods vary widely and are often highly sensitive to noise, particularly in free-living data [180]. As a result, respiration rate has not yet been established as a reliable or widely adopted feature for preoperative risk assessment.

2.5 Data Analysis

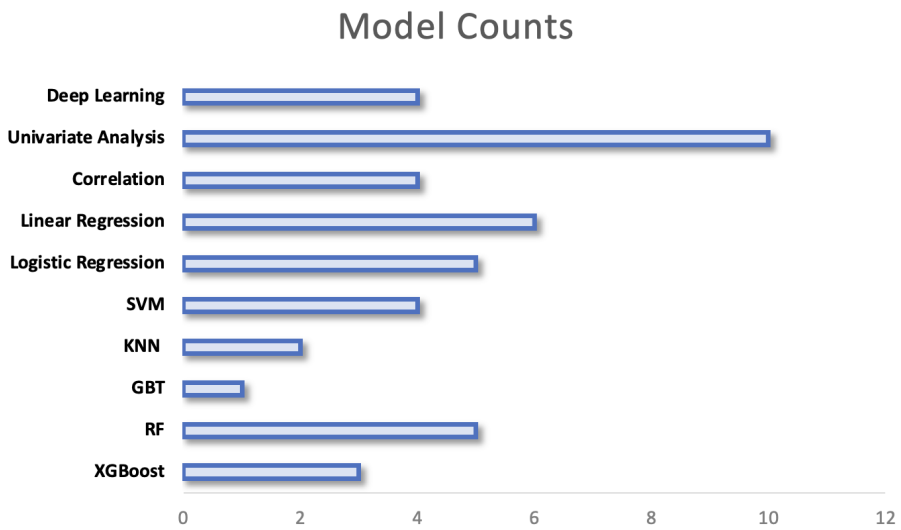


Figure 2.10: Prevalence of each model for analysis across research. In research where multiple models are compared, all models are counted.

Research using data from individuals' behaviour and vital signs from WS to make predictions use a large variety of computational techniques, this can be seen in Figure 2.10. The chosen models range from simple statistical analyses to complex models with high computational requirements. The variation in models may be explained by several factors including the outcome that is being predicted, the size of the data set and the nature of the features that are extracted from signals. The models employed

are separated by their complexity in the following sections, a breakdown of the methods employed within key papers can be found under the appendix (A.2):

2.5.1 Feature Selection

Free-living WS datasets can yield many potential features, some of which may be redundant or weakly related to outcomes. Feature selection is therefore a key first stage in model development, helping to improve performance, reduce dimensionality, and enhance interpretability while limiting the risk of overfitting.

The most common approach was univariate filtering, where each feature is tested independently for its association with the target outcome. The features found to have a strong relationship with the outcome are selected for further modelling. For example, Pearson correlation was used to test relationships between physical activity features and cardiorespiratory measurements, with variables below a set significance threshold (e.g. $p < 0.05$) excluded from further modelling [105]. To address multicollinearity, features with covariance above 0.7 were then compared, and the feature with the stronger correlation with the outcome retained. Similar univariate filtering has been applied to link WS data to postoperative complications [90]

In other studies, categorical grouping of participants (e.g. by thresholds in WS data or by postoperative outcome) led to the use of chi-squared, Fisher's exact, or non-parametric tests such as Kruskal–Wallis and Wilcoxon rank-sum [97, 95, 103]. Only features with significant group differences were then entered into multivariate models, although the significance level for inclusion varied ($p < 0.05$–$p < 0.1$). More advanced studies employed multivariable or model-based selection, such as backward stepwise regression guided by the Akaike Information Criterion (AIC), to retain only the most informative predictors [98, 181].

Overall, feature selection was inconsistently applied across the reviewed literature, and thresholds for inclusion varied. This variability may partly explain differences in model performance reported between studies.

2.5.2 Univariate and Exploratory Analysis

Early studies investigating relationships between preoperative wearable sensor data and outcomes primarily employed simple univariate methods. Univariate analysis refers to statistical tests that examine the relationship between a single predictor variable and an outcome variable, considered one at a time. These approaches are useful as exploratory tools to test whether specific features show associations with clinical outcomes or highlighting potential high-risk groups. However, by design they consider each feature in isolation meaning they can't capture interactions between variables or

provide predictions for individual patients' risk, limiting their clinical utility when used alone.

Group comparisons

In some cases, studies applied univariate techniques as the main analysis to explore the association between preoperative physical activity levels and surgical outcomes. The earliest example of this stratified patients based on their preoperative physical activity (PA) levels and compared complication rates between groups [82]. Independent t-tests were performed between those who did or did not develop a complication. More recent work has applied Chi-squared tests to categorical variables and non-parametric tests (e.g. Wilcoxon sum-rank) for continuous variables [83]. These analyses are the simplest way to test whether wearable-derived features can differentiate between patient subgroups.

Correlation Analysis

Correlation analysis examines the strength and direction of the linear relationship between two continuous variables, typically quantified by a correlation coefficient (e.g. Pearson's or Spearman's r) [182]. In the wearable sensor literature, these analyses are often used as an exploratory step to assess whether sensor-derived features are associated with clinical outcomes.

Greco et al. (2023) examined the relationship between preoperative daily step counts and six-minute walk test outcomes, finding strong associations [83]. Jones et al. (2021) extended this approach by correlating wearable features with fitness measurements and then incorporating them into regression models [59]. Similarly, Sun et al. (2022) assessed whether preoperative step counts correlated with postoperative complication rates [94]. While such analyses cannot provide predictive performance on their own, they are useful for identifying potentially informative features for subsequent modelling.

2.5.3 Machine Learning

Amongst papers investigating the link between preoperative WS data and outcomes, machine learning (ML) techniques are widely implemented. ML methods are well suited to non-linear, large-scale data. Broadly, ML can be defined as a range of models/algorithms that aim to learn patterns in data and make predictions, ranging from simple regression approaches to more complex non-linear modes such as neural networks[183]. In this review, we include both regression-based approaches, for example logistic regression, and more advanced ML techniques under this umbrella term since both have been applied in the wearable sensor literature.

Logistic Regression

Logistic regression is a statistical modelling technique used when the outcome variable is binary. It estimates the probability of the outcome as a function of one or more predictor variables, making it widely used in clinical risk prediction [184]. In this case, multiple papers investigated the ability of preoperative PA to predict postoperative readmission as a binary outcome [104, 103, 97]. To improve model generalisability, internal validation techniques were reported by Rossi et al. (2021) using four-fold patient cross-validation and regularisation in the model to prevent the model learning noise and overfitting [104]. In comparison, another study employed leave-one-patient-out cross-validation (LOPO) to allow a change in the distribution of the data and prevent the model from over-fitting to the training data [93]. This technique is a process where one patient is left out as the validation set and the validation is performed k-number of times, where k is the number of patients [185]. In cases where multiple outcomes were modelled, separate logistic regression models were fitted for each. For evaluating multivariable models, the C-index (Harrell's concordance index) was reported as a useful measure of discriminative performance [98].

While logistic regression is useful for binary classification problems, it is less applicable to continuous outcomes such as VO_2max. In principle, VO_2max could be categorised into thresholds (e.g. low vs high risk) and modelled using logistic regression, but none of the studies reviewed employed this approach.

Multivariable Regression

Linear regression, similar to logistic regression, is the simplest form of ML and aims to model the relationship between one or more predictor variables and a continuous outcome [183]. It assumes a linear relationship between predictor variables and the outcome, estimating regression coefficients that estimate how much the outcome changes for a unit change in each predictor. When multiple predictors are included, this is referred to as multivariable regression. Linear regression models are particularly useful in clinical contexts because they are relatively interpretable and provide direct effect estimates for individual features.

In the preoperative setting, multiple studies have applied multivariable regression to postoperative outcomes. For example, Bille et al. (2020) used regression to predict the number of absolute complications per patient [97], while Mylius et al. (2021) applied it to estimate the time to functional recovery alongside complication odds ratios [90]. Variations of linear regression with regularisation have also been employed. Zhang et al. (2024) applied both Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge regression to wearable-derived features when predicting VO_2max [89]. These methods add penalty terms to the regression coefficients, which can improve model generalisability in the presence of high-dimensional feature sets.

More directly relevant to this thesis, regression models have also been used to predict VO_2max from wearable-derived features. Novoa et al (2011) used two separate linear regression models with bootstrap robust estimation (1000 iterations) to predict VO_2max in patients scheduled for pulmonary resection. The first model was tested with mean daily distance walked as the independent variable whilst the second model incorporated distance travelled [96]. Jones et al. (2021) explored similar methods in a cohort scheduled for intra-abdominal surgery, combining preoperative wearable-derived step counts with clinical data to estimate CRF [59]. Several other studies have also shown ability to predict VO_2max in samples using wearable-derived features and multivariable regression, although these were in non-preoperative cohorts [186, 187].

Support Vector Machines (SVM)

Support Vector Machines (SVM) are supervised learning methods that classify data by finding an optimal boundary (hyperplane) that maximises the margin between classes [188]. They can also be extended to regression tasks (SVR), making them applicable to both classification and continuous outcome prediction. SVMs are particularly effective in high-dimensional spaces, which is relevant for wearable data where many correlated features may be extracted.

Several studies have explored SVMs in the context of wearable sensor data. Zhang et al. (2024) compared a range of models, including Random Forests (RF), k-nearest neighbours (KNN), XGBoost, and SVMs, and found that SVMs achieved the highest performance for predicting postoperative complications [89]. By contrast, Cos et al. (2021) reported that Gradient Boosted Trees (GBT) outperformed both SVM and RF models when predicting similar outcomes, highlighting that performance is highly dependent on dataset characteristics and validation approaches [88].

SVMs have also been used for pattern recognition within wearable signals. For example, Altini et al. (2016) used SVMs to classify physical activities from accelerometer and HR data collected in lab-based tests, which were then applied to free-living contexts [186]. This illustrates how SVMs can be used not only for direct outcome prediction but also as a precursor to the final model.

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an efficient implementation of gradient-boosted decision trees that is widely used for structured data prediction tasks [189]. Within the preoperative wearable literature, its application has been very limited; Zhang et al. (2024) included XGBoost in a comparison of machine learning models for predicting postoperative complications but found it to perform worse than simpler approaches [89].

Outside of surgical populations, XGBoost has been shown to be promising for VO_2 prediction in exercise settings. Sheridan et al. (2025) compared several ML models, including XGBoost, for estimating oxygen uptake from wearable inertial and physiological sensors during team sports activities. While deep learning models achieved the lowest error, XGBoost and linear models remained competitive [190].

Ensemble Models

Ensemble models combine the predictions of multiple base learners to achieve higher predictive performance than individual models [191]. A common example is the Random Forest (RF), which aggregates the output of multiple decision trees to improve generalisability and reduce overfitting [192]. Within wearable sensor research, RF models have been shown to outperform linear methods such as LASSO regression when predicting laboratory-based measurements including clinical biochemical and cellular measurements. [122, 193]. RF models have also been to predict CRF, but only in heath volunteers following exercise protocols [139]. They were also found to be outperformed by SVM models when predicting postoperative complications [89]. Findings are generally inconsistent across studies and the performance of RF models in this clinical context is not clear.

2.5.4 Deep Learning

No deep learning (DL) techniques employing neural networks were utilised in the research investigating the associations between preoperative wearable sensor data and postoperative outcomes. However, there were multiple studies that utilised DL when predicting clinical fitness measurements. These were employed when analysing large cohort data sets [107, 84, 106].

Deep learning (DL) is a subset of ML that uses artificial neural networks with multiple processing layers to model complex, non-linear relationships. At its core, a neural network consists of layers of interconnected units (or “neurons”) that transform input data through weighted connections, with activation functions. While the simplest networks consist of only a few layers (perceptrons), deep learning methods typically employ many layers to progressively learn abstract representations of the data by changing its internal parameters using the back propogation algorithm [194]. One example of an implementation of this is the ’Step2Heart’ algorithm [84]. This paper proposes the ’Step2Heart’ receiving high-dimensional activity inputs to predict HR response which similarly uses the accelerometer data to predict HR. Stacked CNN and RNN layers are combined where the CNN learns spacial features and the RNN learns temporal features of the data. Aside from predicting a future HR response, this model has also shown to have further clinical value in the preoperative period in that it can be utilised to

predict VO_2 max values. A further example of successful VO_2 max prediction employed two feed-forward layers with 128 units that are densely connected [107]. This model was able to outperform other models and was also able to predict future changes to VO_2 max recordings. However, both of these examples leveraged relatively large datasets to produce these outputs.

In summary, a wide range of ML modelling approaches have been applied to wearable sensor data in the literature, from linear regression and shallow ML models to deep learning models. When predicting postoperative complications or cardiorespiratory fitness, simpler regression-based models and shallow ML models have been most commonly employed, reflecting the small cohort sizes, concerns regarding over-fitting and the need for interpretability. More complex deep learning approaches have primarily been applied in large, non-preoperative cohorts or have been implemented for processing of raw physiological signals, rather than outcome prediction. Taken together, these approaches could be considered complimentary, with large models offering potential value for signal processing and feature extraction, while simpler models support interpretable clinical prediction.

2.5.5 Performance Metrics

Across the research outline, a range of metrics are reported to evaluate predictive performance. The most common and relevant included are:

- **Root Mean Square Error (RMSE):** Measures the average magnitude of prediction errors, with larger errors penalised more heavily. Lower RMSE values indicate better model fit.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.2}$$

- **Mean Absolute Error (MAE):** Similar to RMSE but less sensitive to outliers, representing the average absolute difference between predicted and observed values. Again, a lower MAE indicates a better model fit.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.3}$$

- **Standard error of the Estimate (SEE):** Represents the typical deviation of observed values from the regression line, accounting for the number of predictors in the model.

$$\text{SEE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}} \tag{2.4}$$

- **Coefficient of Determination (R^2):** Quantifies the proportion of variance in the observed outcome explained by the model. Values closer to 1 indicate better fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.5}$$

- **Correlation coefficient (r):** Measures the strength of the linear association between predicted and observed values, independent of scale.

Where y_i are the observed values, $\hat{y}_i$ are the predicted values, $\bar{y}$ is the mean of observed values, n is the number of samples, and k is the number of features in the model.

Studies evaluating VO_2max prediction from wearable data have reported encouraging results, though performance varies by population and study design. In general populations, strong predictive accuracy has been demonstrated. For example, Altini et al. achieved an R^2 of 0.78 and RMSE of 284.7 ml/min, while Beltrame et al. reported a correlation of $r = 0.88$[139, 138]. More recently, Spathis et al. (2022) applied deep learning to large wearable datasets, achieving an R^2 of 0.66, RMSE of 2.96, and $r = 0.82$[107].

In preoperative cohorts, fewer studies are available, but results suggest similar potential. Jones et al. reported an R^2 of 0.74, correlation of $r = 0.86$ [59]. Novoa et al.'s R^2 of 0.93 also indicates strong predictive ability [96].

The range of ML models implemented across research displays how there are several suitable options when predicting outcomes from WS data. However, there is no consensus on which ML techniques are best suited to this data type; particularly in research comparing the application of several ML models there have been confounding conclusions. This is likely due to the range of factors that will influence the performance of these models, including the sections discussed previously.

2.6 Future Challenges and Opportunities

This literature review has identified research utilising WS that is applicable to the preoperative period. Several findings were made that relate to the four subsections of the review. These are outlined below and described in the context of future challenges for research.

2.6.1 Comparison of Sensor Modalities

As shown in Figure 2.4a, accelerometers were the most widely used wearable sensor modality in the reviewed studies, while a smaller subset incorporated cardiac signals

(ECG or PPG). Accelerometer-only devices offer several advantages: they are simple, unobtrusive, require little user input, and generally have low power consumption. For example, the RT3 accelerometer used by Feeney et al. (2011) could record continuously for up to 21 days without user intervention, whereas the ECG necklace device employed by Altini et al. (2016) required daily charging [82, 195, 138]. Reduced charging helps to minimise non-wear periods, lowering the risk of missing data. In addition, unlike ECG and PPG sensors, accelerometers are not dependent on skin contact, further decreasing the likelihood of signal loss (see Figure 2.4c). For applications where cardiac data are not essential, accelerometers therefore provide a practical and reliable means of monitoring preoperative activity levels.

By contrast, combining accelerometer signals with ECG or PPG enables a more comprehensive assessment of patient health. Cardiac data provide insights into cardiovascular function that cannot be derived from accelerometers alone, with HR, HRV, and related features strongly associated with cardiovascular disease and perioperative risk [196]. They also enable the detection of specific abnormalities such as arrhythmias [196]. Multi-modal approaches have demonstrated improved performance in activity recognition, with superior accuracy achieved when accelerometry is combined with ECG, and have been used to enhance signal quality assessment, such as through the joint use of PPG and accelerometer signals [197, 198]. In principle, combining cardiac and movement data allows for a more holistic evaluation of health, richer feature extraction (see Section 4.2.1), and improved classification of activities. However, these advantages come at the cost of increased device complexity, shorter battery life, and more demanding data processing. Selection of sensor modality should therefore be context-dependent, balancing the value of additional physiological information against the practical challenges of data collection and analysis.

In the research included in this review, patients were commonly recruited once surgical treatment had been scheduled or at the preoperative assessment clinic. As a result, any WS data from patients was usually collected in the immediate weeks preceding surgical treatment (see Section 2.1.4). There was no variation in the data collection period between sensor modalities; however, the preoperative period is by definition a broad label for any preoperative assessment (see Figure 2.1). The varying characteristics of sensor modalities outlined above indicate that simple devices with large data storage and battery capabilities may lend themselves to data collection periods extending over a longer period whilst more complex devices that incorporate multiple signals may be useful at shorter intervals to assess vital signs. Therefore, different sensor modalities may be suitable at different stages prior to surgical treatment, but there is no existing literature discussing this.

2.6.2 Missing Data Periods

Missing data was frequently reported across studies, arising from factors such as non-wear, poor electrode contact, or device removal for charging. Approaches to handling these gaps were inconsistent: some studies applied extraction thresholds (e.g., minimum daily wear time), others discarded incomplete data, and a few employed imputation techniques.

Extraction thresholds are sometimes used to identify when there is a suitable volume of data collected either over the course of each day in the monitoring period or over the course of the entire recording. Although some research justifies the selection of an extraction threshold through testing, this was infrequently the case and an arbitrary unit was selected. Some research reports abandoned data whilst others report using imputation techniques. A further limitation is that the underlying causes of missing data were rarely systematically investigated, making it difficult to compare results across studies. Although a small number of papers acknowledged issues such as frequent electrode replacement [138], the field as a whole lacks protocols for classifying and addressing different sources of missingness. Standardised methods for reporting, investigating, and handling missing data would improve reproducibility and ensure that appropriate solutions (e.g., imputation versus exclusion) are applied in future work.

2.6.3 Raw Signal Data

The majority of research that is applicable to the preoperative setting utilises commercially available sensors that employ their own internal proprietary algorithms, a common example being the Fitbit Inspire [103, 89, 83, 91]. From this, researchers are provided with pre-extracted features; for example, heart rate and step count, where no pre-processing is required. This can improve access to research as data from these sensors is computationally simpler to work with; however, it also limits researchers in the agency that they have for pre-processing and feature extraction. In research predicting VO_2max, step-count and floors-climbed were extracted by an internal algorithm from a wrist-worn sensor [59]. Other research has shown that acceleration-derived Metabolic equivalent of Task (METs) and raw acceleration alongside step count data has been shown to be predictive of cardio-respiratory fitness [107]. Further, a variety of pre-processing techniques can be applied to accelerometer signals (see Section 3.2) and when using proprietary algorithms, research may be limited in how it can filter signals. Proprietary algorithms limit researchers in the features they can use from the raw accelerometer signals.

Similar conclusions are particularly pertinent for cardiac signals from ECG or PPG sensors (see Section 2.4.2). Often HR is the only extracted feature from these signals and although HR is a useful measure with strong associations for health (see Sec-

tion 2.2.2), there is a vast amount of information that can be further extracted from cardiac signals. HRV, calculated from the location of the QRS complex (see Section 2.4.2), has been a popular metric of health but requires the lengths of R-R intervals. Research has suggested that not only ECG but PPG signals are adequate for estimating specific HRV features that are relevant for assessing patient deterioration [199]. Further, research has identified the potential for these cardiac signals from wearable sensors to identify patients at high risk of suffering cardiac abnormalities [200]. This highlights a large gap in useful data from both ECG and PPG signals that is not being utilised within research applicable to the preoperative setting. Although proprietary algorithms can simplify data access to wearable sensor data by removing the barriers of preprocessing data, it may in turn limit research capabilities by reducing the dimensions of the data and preventing extraction of features. Future research should investigate the added prognostic value that these underutilised features from raw signal data can bring to the predictive performance.

2.6.4 Predictive Models

There is a wide range of models being employed to investigate differences in outcomes, from simple group comparisons (e.g., active vs. inactive) to predictive models estimating complication risk, hospital readmission, or CPET results. Machine learning models, particularly those incorporating features extracted from HR and accelerometry, have shown success in predicting both postoperative outcomes and clinical fitness measurements. However, many existing studies rely on relatively simple, static features rather than fully exploiting the richness of continuous signals. While temporal features such as HR recovery (see Section 2.4.2) hold promise, their use remains limited in the literature. More broadly, ML methods provide unique advantages for handling non-linear relationships in wearable data, and comparing multiple approaches is often necessary. No single model has consistently outperformed others in the preoperative setting, suggesting that flexibility and methodological variety remain important in future work.

Deep learning (DL) methods have shown strong performance in predicting clinical fitness measures when applied to large datasets. However, their application in preoperative research presents challenges. Preoperative wearable studies typically recruit smaller cohorts, making DL models prone to overfitting [201]. This issue may be compounded by the limited feature sets available from proprietary algorithms (see Section 6.4), which reduce the richness of input data. DL models also demand considerable computational resources, which are not accessible to all research groups. By contrast, shallow ML models have already demonstrated good performance for postoperative outcomes and are practical for smaller datasets. Looking ahead, DL may hold promise when applied to large volumes of raw wearable data, particularly for tasks

such as noise reduction and feature extraction.

2.6.5 Considerations for Implementation

Wearable sensors generate large volumes of highly personal data, often referred to as Patient-Generated Health Data (PGHD) [202]. Ensuring this information is stored securely and used responsibly is critical, particularly as privacy concerns can affect patients' willingness to share data. Despite this, few of the reviewed studies explicitly addressed privacy or data governance, reflecting a focus on methodological rather than implementation. Several regulatory frameworks, such as the General Data Protection Regulation (GDPR), provide guidance for the safe use of individual data. Recent work has highlighted the importance of amplifying user agency over health data to build trust and support participation [203]. While not the focus of this thesis, such considerations are essential for future clinical implementation of wearable-based preoperative assessment.

Beyond technical performance, the successful integration of wearable sensors into preoperative care depends on clinician buy-in. Preliminary work suggests clinicians see value in wearables as tools to provide feedback and support decision-making, but concerns remain around trust, transparency, and control of data [204]. Developing explainable models, where feature contributions can be clearly communicated, may be key to clinician confidence and eventual adoption. Regulatory frameworks also play a critical role. Currently, many commercial wearables fall outside FDA approval requirements as they are classified as lifestyle devices, meaning clinical validation is limited [205]. As evidence of perioperative benefits grows, stricter requirements for regulatory approval and validation may emerge. Finally, issues of generalisability should be acknowledged: study samples are often unrepresentative, with limited reporting of participant demographics, which constrains the translation of findings to broader surgical populations.

Finally, if wearable sensors are to be implemented more widely in the perioperative pathway, their cost-effectiveness must be established. At present, most work in this area is feasibility research, and few studies have systematically evaluated the economic implications, particularly in the preoperative stage. Some evidence suggests that wearables may be especially valuable for patients with limited access to follow-up care (e.g., rural settings), offering a cost-effective alternative to in-person monitoring [206]. More broadly, while preliminary studies highlight potential resource savings, robust cost–benefit analyses are still lacking and should be a prioritised where appropriate.

2.6.6 Conclusions

This review has summarised recent research applying wearable sensors in the preoperative setting, covering sensor modalities, pre-processing methods, feature extraction, and predictive modelling. A consistent limitation is the reliance on proprietary or pre-processed data, restricting the ability to explore richer physiological variables such as HRV. Research making full use of raw ECG data remains scarce, and best practices for signal processing and feature engineering are not established. Similarly, there is no consensus on which machine learning approaches are most effective for this application. Addressing these gaps through systematic use of raw data, further investigation of advanced physiological features, and further model comparison is essential for progressing wearable-based preoperative assessment.

Chapter 3

The REMOTES Dataset

3.1 Introduction to the Study

The remainder of this thesis investigates the potential to estimate VO_2max using wearable sensor data in a preoperative cohort. A dedicated dataset was collected to support this investigation, addressing several gaps identified in Chapter 2 through the use raw wearable ECG signals and corresponding CPET-derived VO_2max values. This chapter therefore describes the dataset in detail, including the study cohort, measured outcomes and the characteristics of the physiological signals collected from wearable sensors.

The Remote Monitoring for Preoperative Risk Assessment for Major Abdominal Surgery (REMOTES) study was a prospective observational clinical trial conducted at Leeds Teaching Hospitals NHS Trust between December 2022 and September 2024. The study aimed to evaluate the feasibility and value of wearable sensors for capturing physiological and movement data from patients undergoing major abdominal surgery, with a focus on informing preoperative risk stratification. Ethical approval for the study was obtained prior to participant enrolment (REC reference: 22/SS/0050), and all patients provided informed consent. The study is registered at ClinicalTrials.gov (Trial ID: NCT06042023).

The dataset comprises multi-day recordings of raw physiological and activity signals, collected from patients during a three-day period from routine daily lives in the lead-up to surgery. Unlike structured exercise protocols, this approach sought to characterise patients' physiological profiles in free-living conditions. The review presented in Chapter 2 highlighted several recurring limitations in existing work: a reliance on features processed using proprietary algorithms and limited freedom to explore features that require raw data (e.g. HRV). The REMOTES dataset presents a valuable opportunity to investigate relationships and physiological features that previously were not possible due to constraints in sensor modality and raw data access. The study was designed to explore several key questions:

- Can passive monitoring of daily routines using wearable sensors provide insight into physiological states relevant to patient fitness and surgical risk?
- Can wearable-derived features be meaningfully linked with routinely collected clinical variables, particularly those obtained in the preoperative stage?

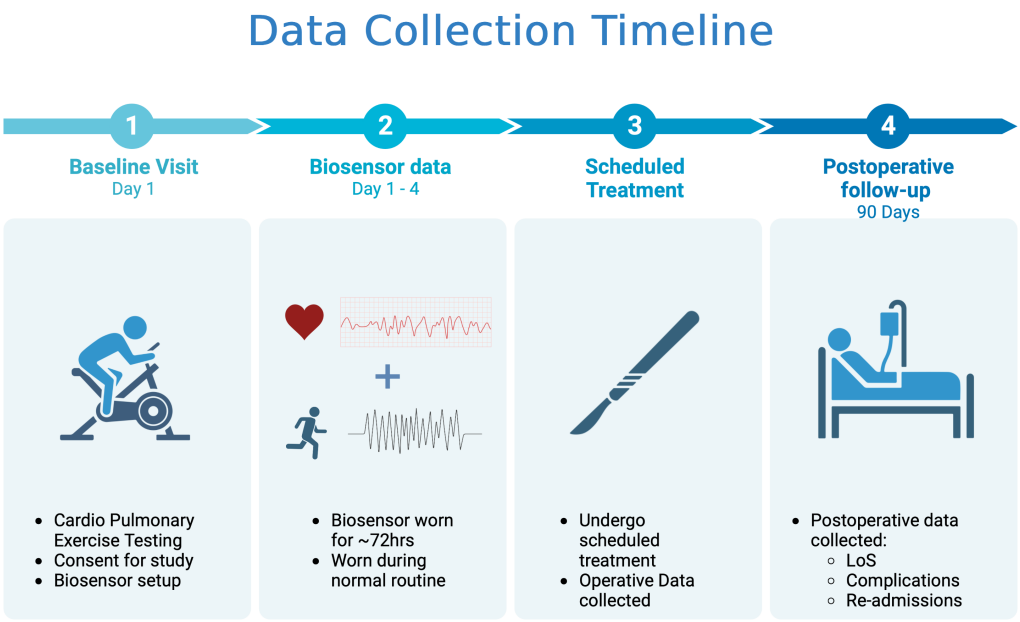


Figure 3.1: Overview of data collected from each participant as they moved through each stage of the study. Only data collected from stages 1 and 2 are presented in this thesis.

Although the REMOTES study included collection of operative and postoperative outcomes, the analysis presented in this thesis focuses exclusively on the preoperative data, as shown in stages 1 and 2 of Figure 3.1. This reflects the specific interest in the use of wearable devices as an alternative to existing preoperative assessments for evaluating patient fitness.

I was not directly involved in the recruitment of patients or the deployment of wearable sensors during the study. I was responsible for the processing and analysis of raw signal data, and was involved in setting up infrastructure for data storage and pipeline development. While I tested the sensor in preparation for analysis, all clinical data collection was undertaken by the Principal Investigator, Dr Alexios Dosis, and a team of research clinicians.

3.1.1 Participant Recruitment and Timing of Data Collection

All individuals scheduled for elective major abdominal surgery (BUPA classification: Major 1 to +5) were screened against inclusion criteria before attending the preopera-

tive clinic (Figure 3.2). This approach aimed to minimise selection bias and ensure a sample representative of the broader preoperative population. Participants underwent their preoperative CPET on a treadmill or cycle ergometer, adhering to the Association for Respiratory Technology and Physiology guidelines, supervised by an anaesthetist or exercise physiologist [207]. If inclusion criteria were met, participants received written study information and gave informed consent. The study aimed to recruit 200 participants, estimated from the number of patients undergoing CPET over 18 months at the study location assuming a 33% acceptance rate. While no formal sample size calculation was performed, this exceeds previous studies' sizes [59, 208, 170, 171]. Following CPET, participants were fitted with a Ubiqvue-Lifesignals LX1550E chest-sensor, worn continuously for 72 hours during daily activities.

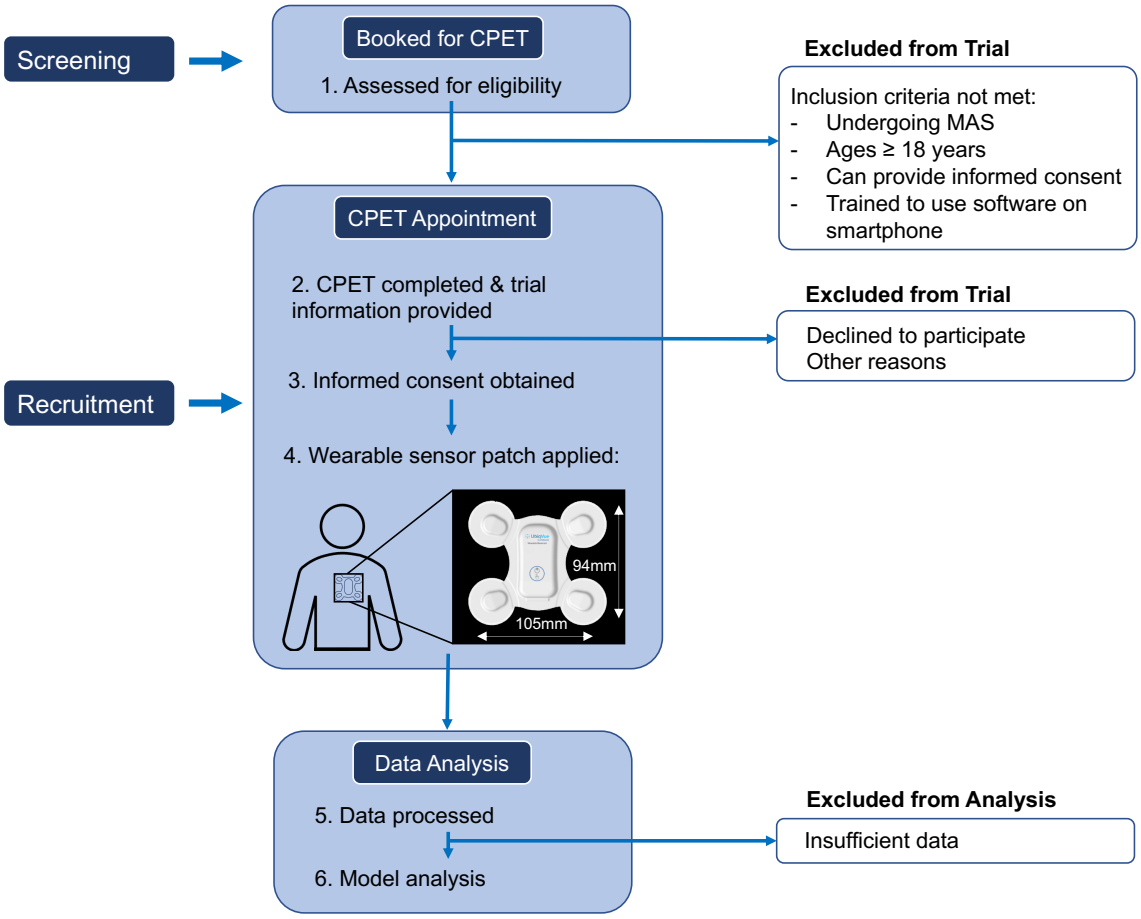


Figure 3.2: Recruitment process for participants. All individuals scheduled for major abdominal surgery (MAS) and booked for CPET during the trial period were screened against inclusion criteria. If met, they were then given study information before providing informed written consent. The wearable sensor (Ubiqvue-Lifesignals LX1550E chest sensor) was attached following CPET via an integrated self-adhesive pad on the left side of the chest, as shown. An outline of the sensor with physical dimensions is shown.

3.1.2 Wearable Sensor Device

Participants were fitted with a Ubiqvue-Lifesignals LX1550E chest-worn sensor immediately following their clinical CPET assessment. The sensor was secured to the left side of the chest using an integrated self-adhesive pad and worn continuously for up to 72 hours during free-living activity. The device recorded two key physiological data streams:

- **ECG**: 2-lead Electrocardiogram, sampled at 244.14 Hz
- **Accelerometer**: Triaxial accelerometer data, sampled at 25 Hz

Importantly, no preprocessing or onboard signal cleaning was performed, resulting in completely raw ECG and movement data. Data were transmitted wirelessly via a paired smartphone to an encrypted cloud server. Although the sensor is capable of recording additional signals such as skin temperature, these were excluded from analysis due to concerns regarding data quality and relevance to the aims of this work.

To transfer data from the device, data were stored on-device for a maximum of 6 hours but required a wireless connection to offload data to the LifeSignals server, as shown in Figure 3.4. To do this, patients were also provided with an android phone setup with a SIM card, provided purely for the aim of data transfer. The connection between the LifeSignals device and the mobile phone was maintained via a personal hotspot, but disruptions such as battery depletion or loss of mobile signal could interrupt the transfer.

(A) (B)

Figure 3.3: (A) The transfer of data from a wearable device to the LifeSignals server via personal hot-spot from the provided mobile device. (B) Image of the LifeSignals wearable biosensor (source: LifeSignals)[1].

[1] https://www.lifesignals.com/wp-content/uploads/2024/04/2A-Ubiqvue-by-lifesignals-Biosensor-1.png

3.2 Participant Characteristics

Demographics

Over the course of the REMOTES study, 198 participants met inclusion criteria and consented to participate. Of these, one device malfunctioned, resulting in 197 participants with at least some quantity of collected wearable sensor data. One further participant had a failed CPET and was removed from analysis. Demographic and clinical characteristics of the remaining 196 participants are summarised in Table 3.1.

Table 3.1: Demographic, physiological, and wearable data characteristics of the participant cohort (N=196), stratified by gender. Values are mean (SD) [range].

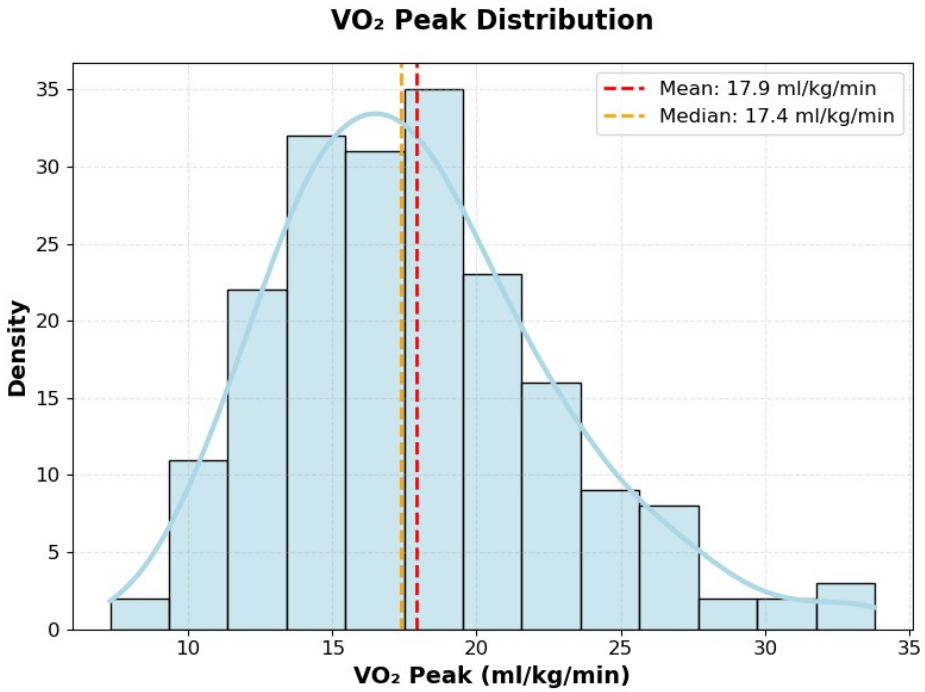
Characteristic	Male (n=140)	Female (n=56)
Age (years)	69.2 (10.2) [38–90]	67.3 (12.6) [29–92]
BMI (kg/m^2)	28.2 (4.9) [18–45]	30.3 (7.4) [17–50]
VO_2max (ml/kg/min)	19.0 (4.8) [11–34]	15.3 (3.9) [7–26]
Wear time (hours)	73.5 (16.9) [11–120]	73.0 (20.6) [22–120]

The age of participants ranged widely, with a mean age of 68.7 years (range: 29–92). The interquartile range (IQR) was 61.0–77.0 years, and the median was 70.0 years, indicating that the cohort was skewed to older participants. This is broadly in keeping with typical age distribution of patients undergoing major abdominal surgery, though this cohort were on average slightly older than those reported in previous research [14].

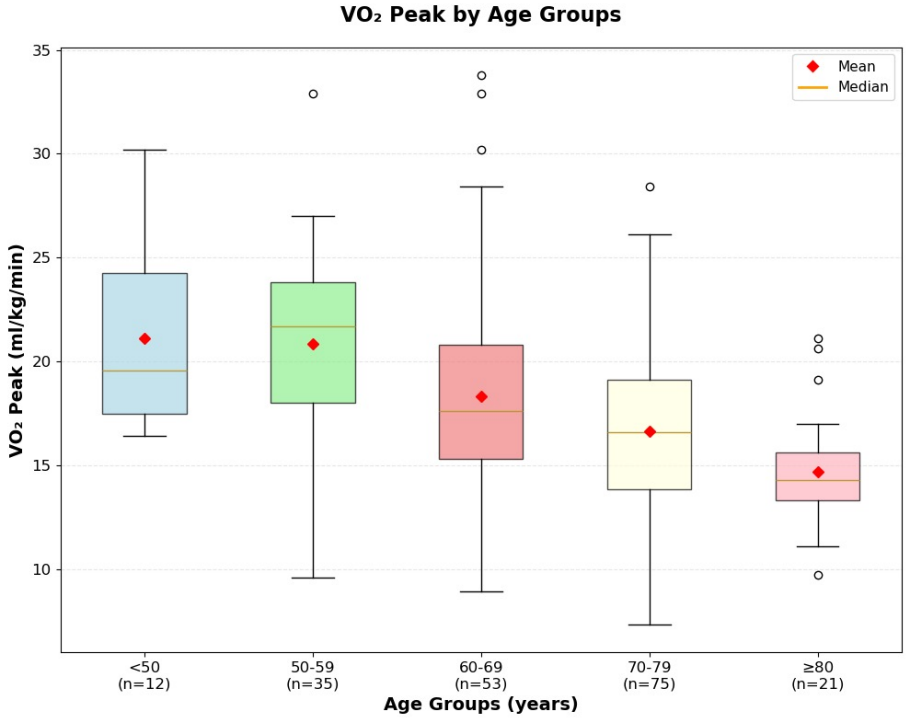
Gender distribution was heavily imbalanced, with 71.6% (n=141) of participants reported as male and 28.4% (n=56) as female. This reflects previously reported male predominance in surgical procedures included in this study, though our cohort is even more male-skewed [14]. Female participants had slightly higher BMI values on average.

As mentioned previously, all participants were scheduled for procedures classified as *major abdominal surgery* according to BUPA classification (1 to +5) [209]. This does not include cardiac or transplant-related interventions. Specific procedures included bowel resections, cystectomies, and total gastrectomies. Many were performed for malignant disease of the gastrointestinal tract, pancreas or liver which are often managed surgically as first-line treatment [11]. Co-morbidities were common, including hypertension, diabetes, and respiratory disease. These clinical variables were not a focus of the present thesis, as the aim was to explore the predictive value of wearable sensor data and basic demographic features alone on cardiorespiratory fitness. This was done to evaluate the utility of readily available features like age, body mass index (BMI) and gender in combination with wearable data, without the need for more complex clinical data.

Cardiorespiratory Fitness



(A)



(B)

Figure 3.4: (A) the distribution of VO_2max values for the entire cohort. It shows a histogram with overlaid kernel density estimate showing a slight right-skewed distribution. (B) boxplot for VO_2max values stratified by age group (< 50, 50–59, 60–69, 70–79, 80+ years), showing a decline in cardiorespiratory fitness with increasing age. Boxplots display median, interquartile range, and range for each group. Red markers show the mean value within each group and the yellow line shows the median values.

Cardiopulmonary fitness was assessed using peak oxygen uptake (VO_2max), collected via preoperative CPET. In this cohort, CPET was performed using either treadmill or cycle ergometer protocols depending on preference and access, as described in Section 3.1.1. There are well-established systematic difference between these modalities, with treadmill derived VO_2max values typically being 5–10% higher due to greater upper body muscle mass involvement. To account for inflated outcomes, VO_2max values recorded from treadmill CPET were reduced by 10% to standardise results across testing modalities [207]. This adjustment aligns with guidance from the Association for Respiratory Technology and Physiology who report that VO_2max is between 5-10% higher during treadmill exercise compared to cycle ergometer.

Across the cohort, VO_2max values have a wide range from 7.3 to 33.8 ml/kg/min (mean 17.9, SD 4.85), with most participants clustered between 14–21 ml/kg/min, indicating a slightly right-skewed distribution (Figure 3.4A).

As shown in Table 3.1, VO_2max was on average lower in female participants (mean 15.3 ml/kg/min) than male participants (mean 19.0 ml/kg/min). This is consistent with expected sex-based physiological differences in oxygen carrying capacity and cardiac output [210]. Fitness values also declined with age (Figure 3.4B), with the highest average values in the subgroup under 50 years (mean 21.1 ml/kg/min) and the lowest in participants aged 80 years or older (mean 14.7 ml/kg/min), consistent with known age-related decrease in aerobic capacity [211]. These baseline variations in VO_2max across the cohort provide context for the predictive analyses presented in later chapters.

3.3 Data Collected from Wearables

3.3.1 Wear Time Distribution

Each participant was scheduled to wear the device for approximately 72 hours. In some cases, extended wear of up to five days (120 hours) was suggested by researchers if there was concern that the initial hours of data collection might be of poor quality.

Figure 3.1 shows the distribution of total wear time, calculated as the time between the first and last recorded sample for each participant, regardless of any gaps or missing segments in the signal. This method of calculating wear time was considered appropriate in this context because the device was applied in the clinic by a researcher and secured with adhesive electrodes. It is therefore unlikely that the device was repeatedly removed and reattached during the recording period, as is sometimes the case with wrist-worn wearables [212]. Mean and median wear times, stratified by sex, are shown in Table 3.1, with almost no difference observed between males and females. Wear times varied substantially across the cohort, ranging from 11 to 120

hours.

A notable feature of the distribution is a sharp peak at around 70 hours, corresponding to one quarter of all participants. This suggests that many participants wore the device for exactly the scheduled protocol duration. Overall, the distribution reflects high compliance with wear instructions, with most participants achieving wear times clustered around the central tendency. This is important for assessing not only the quantity of available physiological data but also the validity of subsequent activity and heart rate measurements in the cohort.

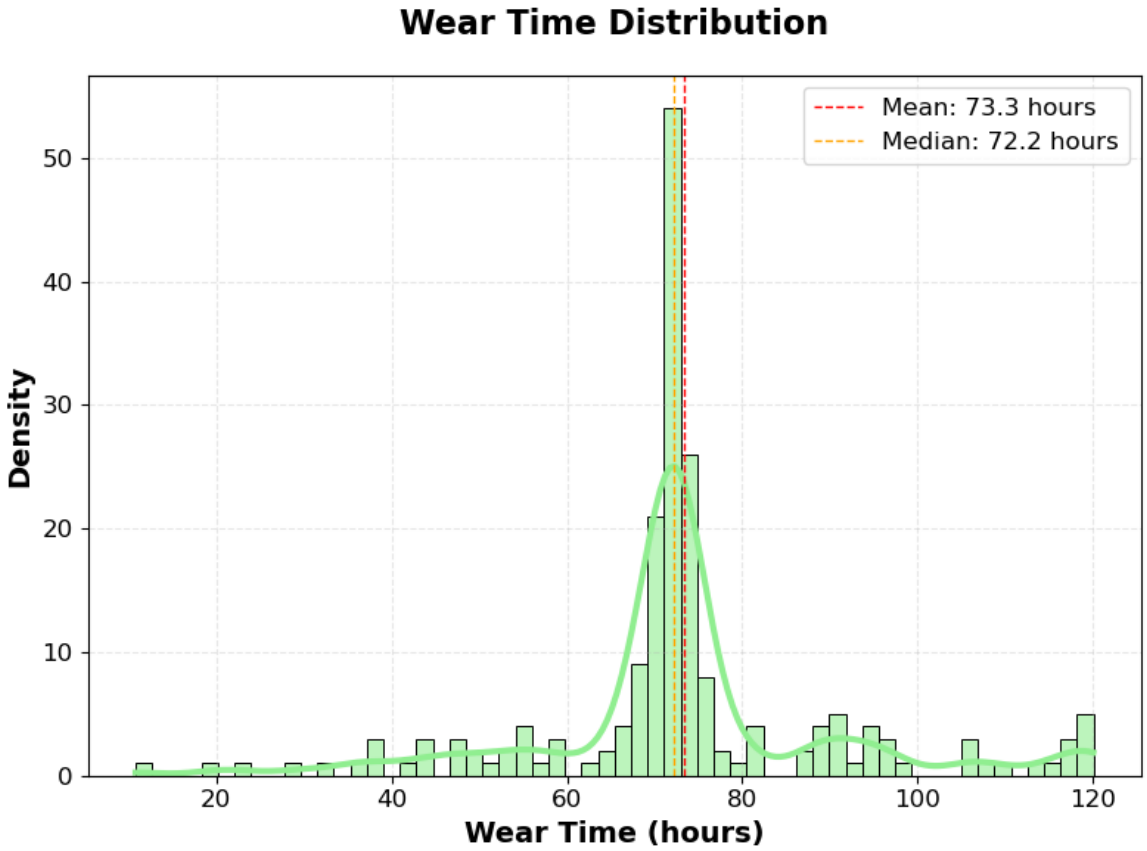


Figure 3.5: Wear Time Distribution. Density plot showing the distribution of device wear time (hours) across all study participants. The histogram displays the frequency distribution with light green bars, while the overlaid green curve represents a kernel density estimate (KDE) used to visualise the underlying distribution given that wear time is non-normally distributed due to the study protocol. Red and orange dashed lines indicate the mean and median wear times, respectively.

Characterising Signals

To better understand the characteristics of the collected data, initial exploratory analysis was carried out using a range of visualisations. This was essential to get familiar with the raw output from the wearable sensor, to interpret how the data are recorded, transmitted, and stored, and to identify what the numerical values in each stream rep-

resent. These visualisations provide the first detailed view of the raw physiological and accelerometer signals in this dataset, highlighting expected patterns and potential artefacts that may influence subsequent analysis.

Figure 3.6 illustrates an example of signal dropout in the raw ECG recordings. In this example of just under 3 days of data, we can see examples of where the signal drops to a fixed amplitude of -2834 mV. This value appears as a static, flat-line segment in the trace. We hypothesise that this represents a default placeholder output from the device when no valid signal is being collected. These dropouts occur with variable duration: sometimes they persist for extended periods as seen in Figure 3.6 where one instance lasts for roughly a quarter of a day. In other cases the dropout is extremely brief, with the signal dropping to the placeholder value and returning to baseline almost immediately. In some participants, this behaviour was observed repeatedly within the same second, reflecting rapid disconnections and reconnections of the ECG signal (see Appendix X for additional examples).

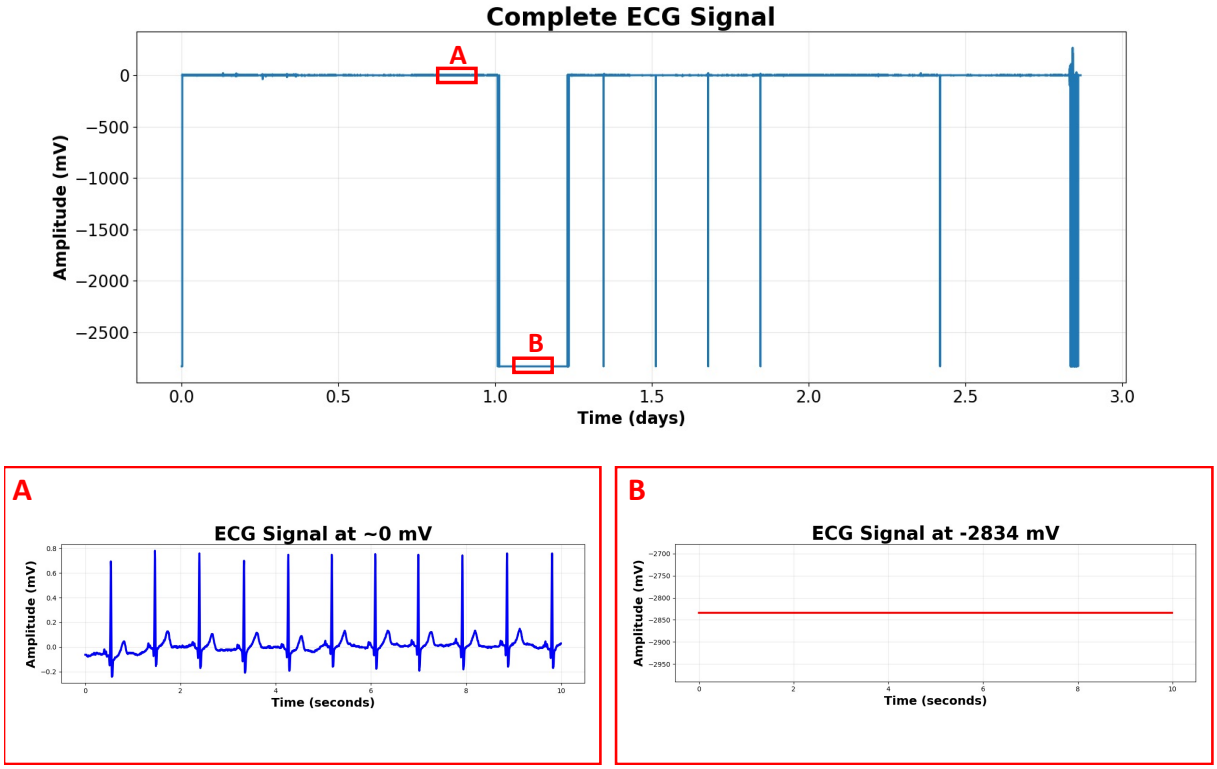


Figure 3.6: Example of ECG signal dropout. The top panel shows a full 72-hour ECG trace from a study participant, with red boxes marking periods of differing amplitude, ranging from a normal baseline around 0 mV to drops below -2500 mV. The lower panel provides a zoomed view of the highlighted segments, illustrating the difference between a standard ECG waveform with clear morphology and a flat-line signal caused by static dropout and a default placeholder value (-2834 mV), indicating signal loss.

To confirm that these placeholder values reflect true signal loss, the same three-day ECG trace was plotted alongside the raw accelerometer recordings from the three

orthogonal axes (Figure 3.8). Here the accelerometer values across all axes drop to exactly -2g during the same time points as the ECG dropouts. In raw accelerometer data such values are not feasible; at least one axis should register a magnitude of +1g due to the effect of gravity. This confirms that no real data were being collected during these periods, and the data presented during these times does not represent valid sensor readings.

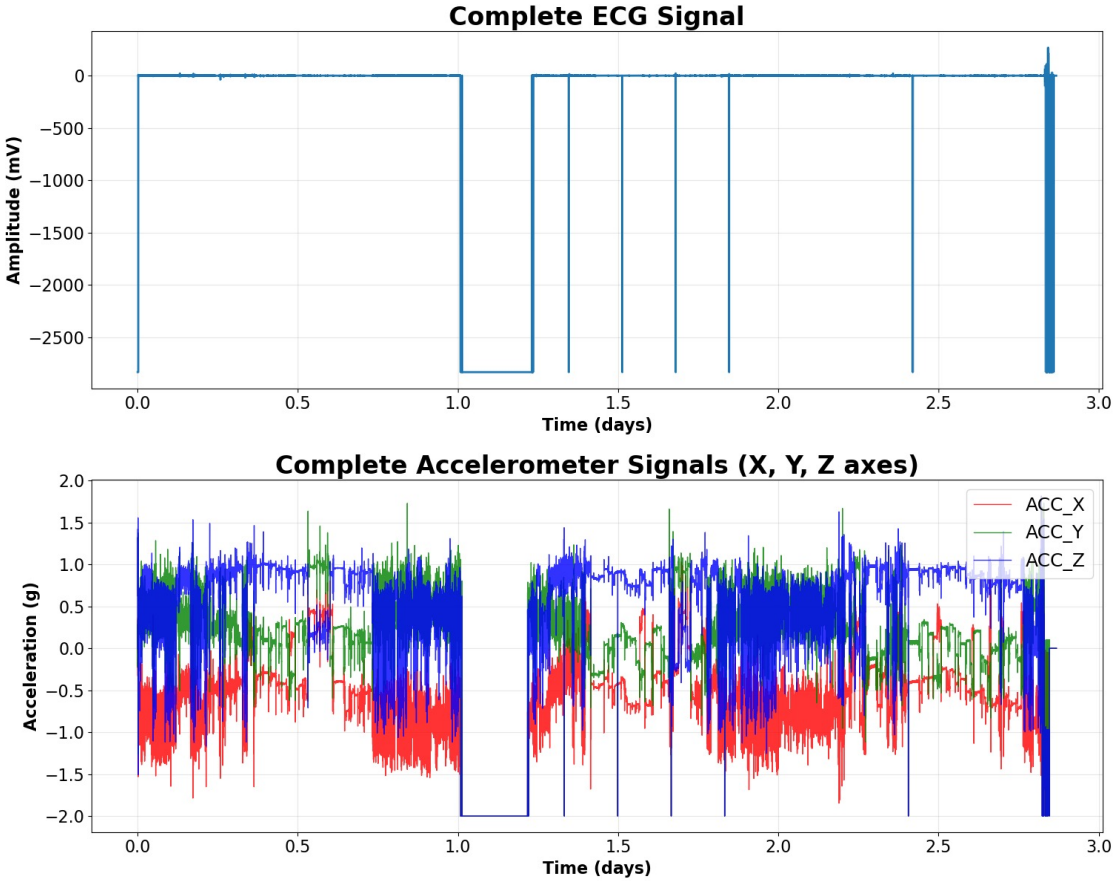


Figure 3.7: Comparison of ECG and accelerometer signals over a three-day recording for the same participant shown in Figure 3.6. Periods where the ECG drops to a fixed value of -2834 mV coincide with accelerometer readings of -2 g across all three axes.

Further investigation of a combination of the ECG and accelerometer data identified that missingness was not always synchronous between ECG and accelerometer signals. From visualisation we could see instances where valid accelerometer readings are collected, but ECG data remained at default placeholder values, as seen in Appendix (Figure B.2). This indicates that the accelerometer and ECG signals require independent characterisation and quantification of missingness.

3.3.2 Quantifying missing data

Using the methods outlined above, we quantified the proportion of valid data collected separately for the ECG and accelerometer channels in order to identify any notable patterns. One clear finding from Figure 3.8 is that, across all participants, there were no instances where a higher percentage of ECG data was collected compared to accelerometer data. In contrast, it was common for the accelerometer to record more data than the ECG — this was the case for 194 participants (99.0%), with only 2 participants (1.0%) showing identical collection rates across both modalities.

On average, the accelerometer achieved a data collection rate of 94.6% (SD: 13.6%), compared with 89.7% (SD: 16.0%) for the ECG. Closer inspection revealed that whenever accelerometer data were missing, ECG data were also missing, but never the reverse. There were many instances where accelerometer data were successfully recorded while the ECG was missing (Figure B.2).

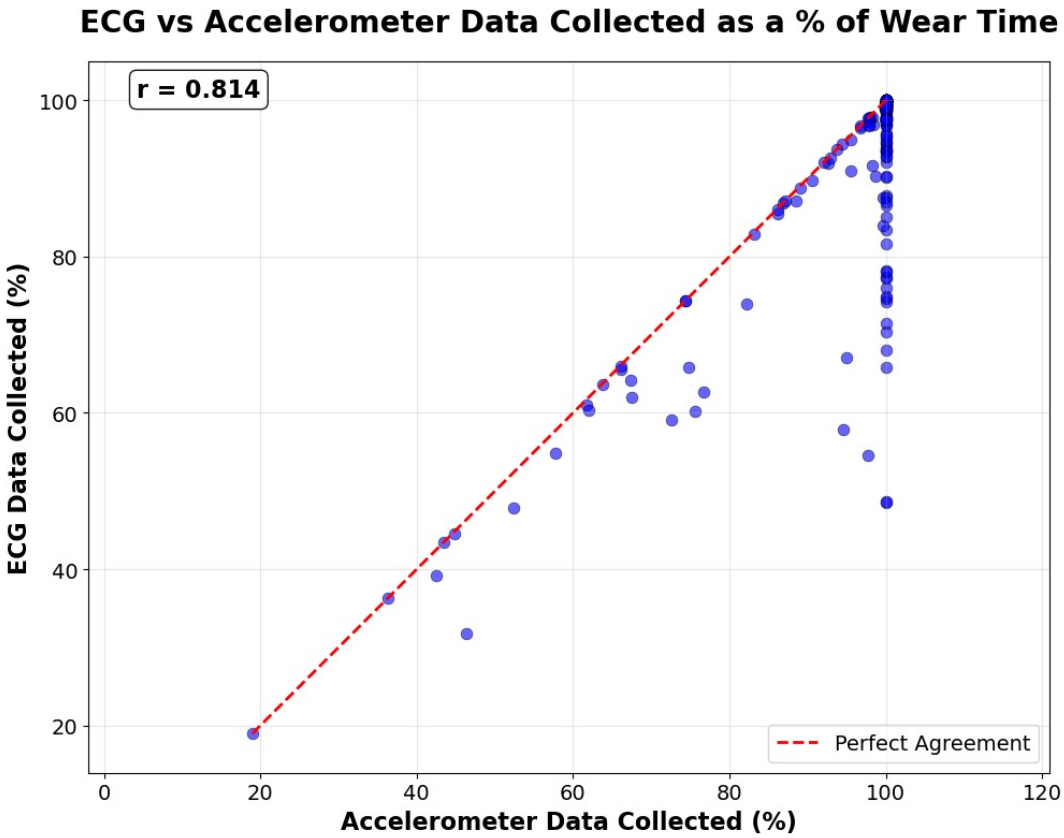


Figure 3.8: ECG vs Accelerometer Data Collection Agreement. Scatter plot showing the relationship between accelerometer and ECG data collection rates (% of wear time) for all study participants (n=196). Each point represents one participant's data collection performance for both modalities. The red dashed diagonal line represents perfect agreement between the two measurement methods. Points close to this line indicate similar collection rates for both ECG and accelerometer data, while deviations suggest differential data quality between modalities.

This suggests the mechanisms for missing data differ between the two signal types. For simultaneous accelerometry and ECG, data loss is likely to indicate a device-level disconnect such as a malfunction or loss of connection with the paired mobile device because if the device was powered and connected as the accelerometer can continue recording, regardless of skin contact. In contrast, ECG recording requires continuous skin–electrode contact to capture the electrical activity of the heart; any disruption to this contact, even if the device remains connected, will result in missing ECG data. As such, ECG data can be lost for reasons that do not affect accelerometer recordings.

Quantifying missing periods correctly, and with some understanding of the underlying mechanisms, is important for several reasons. Participants with a high proportion of missing data in either the accelerometer or ECG channels may not provide a representative picture of their usual physiological state or daily activity. While most participants

in this study contributed between 60–70 hours of usable data, when only considering total wear time three participant had less than 24 hours. However, when counting missing data in individual signals, eight participants (4.1%) had less than 24 hours of ECG data, and six participants had less than 24 hours of accelerometer data. Previous studies have suggested that at least three days of continuous free-living monitoring are needed to capture representative patterns of physical activity, whereas 24 hours is generally insufficient [213]. Identifying these cases early allows for informed decisions about participant inclusion and the reliability of derived features.

Although the exact mechanisms underlying missingness cannot be determined solely from retrospective analysis, the patterns observed suggest different causes for each signal type. Missing accelerometer data is most likely due to device-level failures or disconnection from the paired mobile device, events which are unlikely to be strongly related to participant behaviour and can be considered approximately missing completely at random (see Chapter 3.1 for a description of MCAR). ECG data loss may result from the same device disconnections but can also occur independently due to transient loss of skin–electrode contact, for example through excessive movement or sweating. While such periods may not be completely random, they are still unlikely to reflect deliberate device removal, and they typically occur in short bursts.

Quantifying the proportion of missing data is therefore a necessary first step in dataset characterisation and can be done relatively straightforwardly once device behaviour and dropout mechanisms are understood. However, the presence of data does not guarantee its suitability for analysis. Even complete recordings may contain substantial periods of noise, artefact, or poor electrode contact, all of which can compromise the accuracy of derived metrics such as HR. As discussed in Chapter 2, quality assessment is often overlooked in free-living wearable studies, despite its importance for ensuring reliable downstream analysis. In the next section, we examine the quality of the ECG recordings in this dataset.

3.3.3 ECG Signal Quality

While both ECG and accelerometer data were collected in the REMOTES study, initial analysis of data completeness indicated that ECG recordings were generally shorter and more turbulent than accelerometer recordings. This alone provides a rationale for prioritising ECG in signal quality assessment, as maximising usable ECG data is critical for the downstream extraction of cardiac features.

The nature of wearable ECG acquisition makes it particularly susceptible to quality degradation. Continuous skin–electrode contact is essential, and even minor disruptions—such as partial detachment, changes in skin impedance due to sweating, or patient movement—can introduce substantial noise or signal loss. In contrast, accelerometer data collection is less dependent on skin contact, as the sensor measures

movement and gravity irrespective of skin contact, provided the device remains connected.

Another reason for prioritising ECG quality is that its quality can be assessed against well-defined physiological expectations. The QRS complex and its constituent R-peaks provide clear reference features, and deviations from expected morphology or timing can be objectively detected. In contrast, accelerometer signals lack a universal pattern, with expected waveforms varying greatly depending on activity type, intensity and body location making quality assessment more ambiguous. Given these factors, we primarily focus on assessing ECG quality within the dataset.

We first present an example of a clean ECG segment early in the recording (Fig 3.9), where the extracted heart rate aligns with physiological expectations. This is followed by several segments from the same recording in which visible artefacts (e.g., muscle activity and high-frequency noise) alter waveform morphology. All signals shown have undergone standard noise-reduction prior to beat detection using the NeuroKit pipeline [214]:

- **Band-pass filtering** using a 5th-order Butterworth band-pass filter set to 0.5–50 Hz to suppress baseline drift and high-frequency noise as part of standard signal processing.
- **Baseline detrending** to remove slow trends not fully captured by the band-pass (reduces residual wander).
- **Power-line suppression** (notch at 50Hz) to reduce mains interference.
- **Amplitude standardisation** (z-scoring) to stabilise peak detection across segments.

After this cleaning, Neurokit2 was used to automatically detect R-peaks. Further aggressive de-noising was intentionally avoided to prevent attenuating peak amplitudes, distorting morphology or shifting R-peak timing, any of which could alter HR. The examples below illustrate that, even after standard preprocessing, signal quality can vary substantially within a single recording, motivating the need for explicit quality assessment in subsequent chapters.

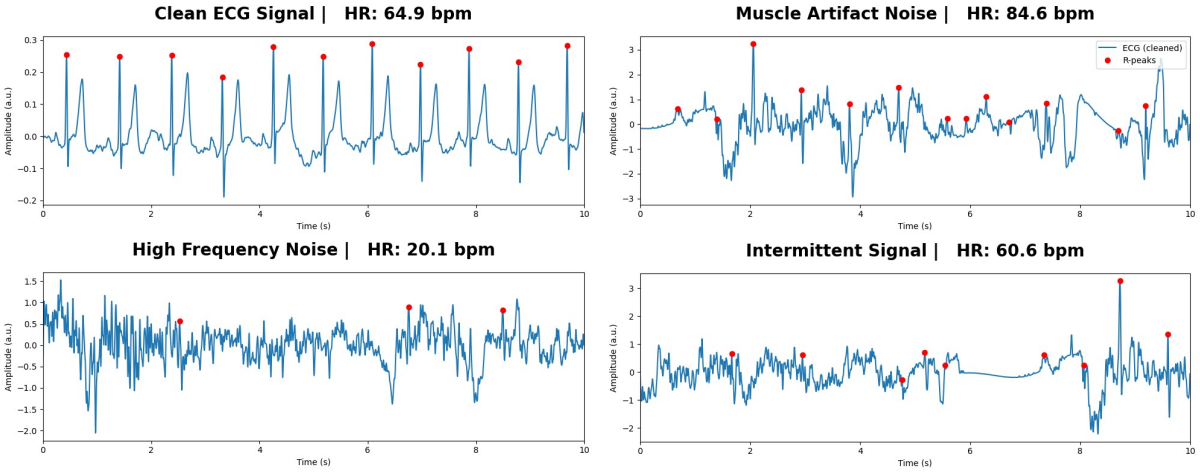


Figure 3.9: Examples of different ECG signal qualities observed within a single recording. Top left: clean ECG trace with stable calculated heart rate. Top right: presence of muscle artefact. Bottom left: intense high-frequency noise. Bottom right: high-frequency noise with intermittent signal loss. For reference, the calculated heart rate is shown alongside each trace.

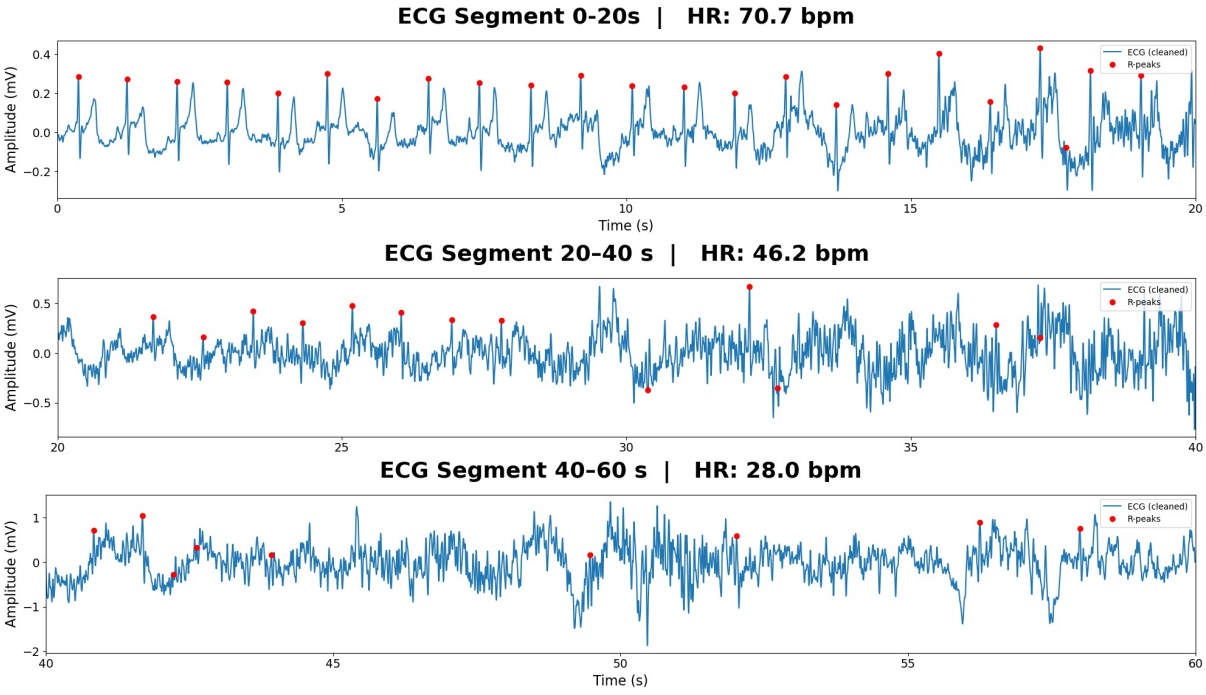


Figure 3.10: Example of rapid ECG quality degradation within a one-minute window. Heart rate estimates from 20-second segments drop from $\sim$70 bpm to 28 bpm as waveform morphology deteriorates, finishing in a segment without any identifiable physiological signal.

In a further example (Figure 3.10, we illustrate how rapidly ECG signal quality can deteriorate and the impact this has on extracted features. Within the space of one minute, we see HR estimates from consecutive 20-second segments fall from an expected resting value of approximately 70 bpm down to 28 bpm, accompanied by visible

changes in waveform quality. In the final segment, no recognisable physiological signal remains, and the extracted heart rate is not representative of any meaningful physiological state. These values were produced despite using the preprocessing methods and robust beat detection algorithm described above. Including segments with this level of distortion would reduce the reliability of derived features and hinder the accurate prediction of cardiorespiratory fitness measures.

Segmenting ECG recordings into short windows (e.g., 10 seconds) is a standard protocol for identifying usable signal periods. This approach prevents distorted signals from contributing unrealistic feature values. However, the scale of the REMOTES dataset makes visual inspection impractical: with nearly five million 10-second ECG segments, manual review is not feasible. Automatic selection of high-quality signal segments is therefore essential before relevant features can be reliably extracted. This requirement motivated the work in the following chapter, in which we benchmark several open-source signal quality indices and evaluate their suitability for free-living ECG data.

Chapter 4

Assessment and Implementation of Signal Quality Indices

4.1 Background

As shown in Chapter 3.3.3, ECG signals collected in the REMOTES study varied substantially in quality from episodes with complete signal loss or frequent motion artefacts to clean ECG morphology. Filtering of signals is a form of preprocessing that aims to remove signal interferences, without removing any useful physiological waveform morphology. However, the problem lies in that physiological morphologies in an ECG signal can overlap with noise, meaning that in some cases it is impossible to remove noise without altering or degrading the morphology of the physiological signal. Therefore, while preprocessing steps can help with noise removal and R-peak detection, there are signals that remain corrupted beyond use. Automatic methods of signal quality assessment offer an alternative option by identifying poor quality signals and those with missing data, while retaining important physiological features in good quality segments. To prevent distorted signals propagating into unrealistic cardiac features, this automatic segmentation is the first key step in analysing the wearable ECG signals from the REMOTES dataset.

Fortunately, this is not a novel challenge but rather one that has had increasing attention. Wearable sensors have seen a jump in popularity over the last decade as they become accessible to the general public, and ECG recordings are increasingly being incorporated into these devices [215]. As ECG recording becomes more common in consumer and research devices, the volume of data that requires interpretation has grown rapidly. Further, because ECG signals are inherently sensitive to electrical activity, they are particularly vulnerable to noise when recorded outside of a clinical setting. The number of trained experts available to review these signals has not increased in parallel, driving the need for automated assessment. Another application for this is that signal quality can rarely be retrospectively substantially improved and

the best option is to assess quality in real-time and adjust the device at the source (i.e. electrode positioning). Instant feedback on quality from SQIs can therefore help non-specialists to set-up devices correctly. The importance of this problem was reflected in the PhysioNet Challenge in 2011, in which teams competed internationally to develop algorithms capable of classifying short ECG recordings as ”acceptable” or ’unacceptable” signal quality for clinical interpretation [216]. Signal quality indices (SQIs) have emerged as a standard approach to these issues.

4.1.1 Signal Quality Indices

SQIs are quantitative measures designed to assess whether a given segment of ECG signal is of sufficient quality for reliable analysis, usually applied as a threshold or as features within a classifier [217]. A wide range of SQIs have been proposed that differ in their underlying approach: some use frequency characteristics of the signal, others apply heuristics to extracted features while more recent methods apply ML. Several studies have evaluated these indices, reporting mixed performance depending on the datasets used and the noise that is present. One review focused primarily on statistical SQIs across six different SQIs considering measures such as signal-to-noise ratios, kurtosis of the signal and relative power of the QRS complex [218]. A different approach compared a much wider range of 26 SQIs but only assessed them on one dataset (PhysioNet 2011 Challenge set) [219]. Finally, a comprehensive review comparing 39 SQIs against multiple databases demonstrated that SQI performance depends strongly on the type of noise present in the ECG signal [217]. The four noise types investigated were motion artefacts (MA), Electromyogram noise (EMG), Power Line Interference (PLI) and additive white Gaussian noise (AGN).

This final review highlights an important conclusion in that while some research has compared different noise types, they are generally constrained by the noise characteristics present in the available datasets. This is combined with a further limitation in that evaluations in this field are determined by labels generated by clinical experts. Whilst a valid measure, this means the criteria for what constitutes a ’usable’ or ’acceptable’ ECG may vary between individuals and may not reflect the robustness to specific noise types. This could make it difficult to directly determine the level of noise at which SQIs begin to fail. To overcome this and assess which SQI would be most applicable to the REMOTES dataset, we propose a different method of assessment.

4.2 Assessment of Open-source SQIs

4.2.1 Synthetic ECG Generator

Synthetic ECG refers to artificially generated signals that aim to replicate the characteristics and morphology of ECG collected in real settings [220]. Generally, these synthetic signals have shown to be strongly representative of real ECG data and have been applied in multiple settings, a common example being to increase the volume of training data for ML models. These tools can employ a range of methods to generate data, but generally implement either physiological methods, parametric approaches or machine learning approaches. The benefits of synthetic ECG signals are twofold: a large volume of signals can be generated with known 'ground truths', and there is full control over the signal characteristics including noise levels and morphology.

Rather than using signals pre-determined quality labels, we can therefore implement a synthetic ECG generator that can add controlled increases to noise in signals. This allows the SQIs to be systematically evaluated under different conditions and directly identify their noise thresholds. This tool enables reproducible benchmarking of SQIs by simulating clean ECG signals and progressively adding motion artefacts, power line interference, Gaussian noise, and variations in HR. By doing so, we can systematically examine the extent to which each SQI tolerates increasing levels of noise and determine the thresholds beyond which segments are labelled as 'unacceptable' by the SQI.

To select the SQIs for testing, we wanted to prioritise those that have publicly available code or implementations. A previous review in 2020 identified that many proposed SQIs are published without corresponding code or reproducible implementations, limiting their applicability in practice [221]. Specifically, none of the 19 SQIs that were reviewed published adjacent code. Recently, several indices have been released either as open-source code with graphical user interfaces. For this project, we therefore selected four SQIs with immediate implementation available, prioritising reproducibility and transparency.

Finally, after having identified the noise thresholds of these open-source SQI tools, there needs to be a method to contextualise these thresholds and investigate their clinical relevance. To do this, once signals with maximum tolerable noise levels were established for each index, these were reviewed by a clinician to evaluate how well the SQI classifications aligned with expert judgment. This reviewing of signals was done with several criteria to reflect different potential applications of signal quality assessment, such as heart rate monitoring. In this way, we not only benchmark SQIs in a controlled synthetic setting, but also explore how their decisions compare to clinical expectations.

4.2.2 Aims

Using the synthetic ECG signals and four open-source SQI tools, we have three aims for this chapter:

1. **To benchmark SQIs using synthetic data** by adding controlled increases in signal noise (motion artefacts, heart rate variation, power line interference, and Gaussian noise). This enables systematic assessment of how each SQI tolerates increasing levels of noise and the thresholds at which they fail.

2. **To compare SQI-derived noise thresholds against clinical judgement** by reviewing maximum tolerable noise levels against expert review. This provides insight into how SQI classifications align with clinician judgment in different applications.

3. **To apply the selected SQI to the REMOTES dataset** in order to evaluate overall signal quality across the cohort and establish the proportion of data suitable for downstream physiological analysis.

4.2.3 Methods

The workflow for this study is illustrated in Figure 4.1. First, synthetic ECG signals were generated using the synthetic tool to provide clean baseline signal with controlled noise addition. Each signal was then processed using a set of publicly available signal quality indices (SQIs), and the point at which the outputted SQI label changes to 'Unacceptable' is identified; this is the noise threshold tolerated by the SQI. Finally, these noise thresholds were compared against clinical expert judgment to evaluate their practical relevance across different applications. The following sections describe each stage of this process in detail.

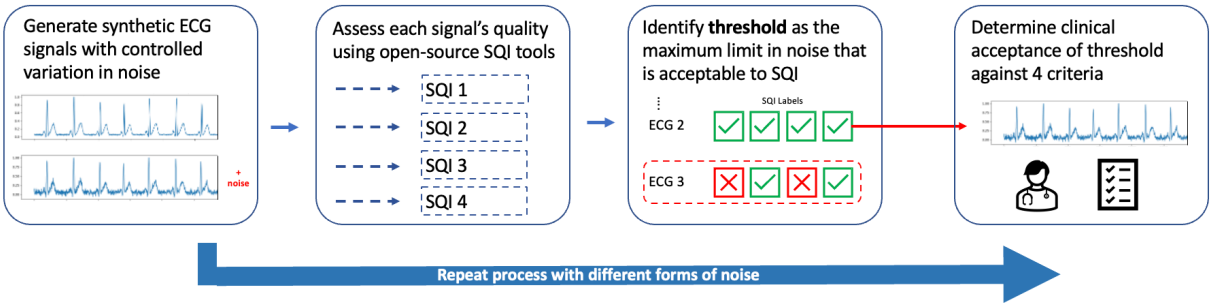


Figure 4.1: Pipeline to assess SQI outcomes.

Synthetic ECG generator

As mentioned in section 4.2.1, a variety of approaches exist to artificially generate ECG signals. Physiological models can capture detailed physiology but are difficult to ran-

domise and computationally intense. Deep learning models like generative adversarial networks can generate signals that mimic real patient data from large datasets, but require a full spectrum of noise variation to be present in those data. Parametric models use mathematical functions to generate signals from controlled input parameters. In this case, that makes them well suited to benchmarking tasks by allowing precise control over noise type and level. For this reason, we employed the Karhinoja et. al (2024) framework for generating synthetic ECG signals [222], which is available at https://github.com/UTU-Health-Research/framework_for_synthetic_biosignals.

It consists of three components: (i) a beat interval generation model that simulates heart rate and respiratory modulation, (ii) a waveform model that constructs each ECG cycle (P, QRS, T) from Gaussian functions with tunable width and amplitude, and (iii) a noise model that adds artefacts based on power spectral densities or recordings from established noise databases (e.g., MIT-BIH Noise Stress Test).

Noise was introduced in several controlled forms to reflect common artefacts encountered in free-living ECG. Point frequency noise was added to the signal to replicate power-line interference (PLI). We used a fixed frequency of 50Hz to model European power supplies and then varied the magnitude in our experiments [223]. White noise, representing electronic thermal noise, was also added with varying magnitude in our experiments. Signals were further augmented through the addition of pre-defined motion artifact types representing muscle artifact noise, hand movement, walking, and baseline wander. Heart rate, although not a form of signal noise, was included as a category of noise to add further common variation to the ECG signals.

Generating signals

To assess the SQIs, the first stage (see Figure 4.1) involves generating signals with varying amounts within four sources: Heart Rate, White Noise, Power-line Interference and Motion Artifacts.

Each source of noise was investigated independently. The amplitude of noise was initially set using the default parameters of the framework to generate a realistic clean signal of heart rate 80 bpm [222]. For heart rate, white noise and power-line interference the amplitude was increased in set increments; increments were selected empirically based on visual inspection to ensure that each step produced a noticeable change in the signal morphology. The remaining types of noise were fixed at the default values. For white noise and power-line interference the signal-to-noise ratios (SNR) are reported as units in the results. SNR was calculated as follows. A clean baseline ECG was first generated but with no added noise. A second signal was then generated using the same beat intervals but with noise added at the selected amplitude. Signal power was computed as the mean squared amplitude of each signal. Noise power was estimated as the difference between the power of the noisy signal and the clean signal.

The signal-to-noise ratio (SNR) was then calculated as:

$$\text{SNR (dB)} = 10 \cdot \log_{10}\left(\frac{P_{\text{signal}}}{P_{\text{noise}}}\right) \tag{4.1}$$

where P_{signal} is the power of the clean ECG and $P_{\text{noise}} = P_{\text{noisy}} - P_{\text{signal}}$, with P_{noisy} the total power of the noisy signal. Larger SNR values indicate cleaner signals. Motion artefacts (e.g., baseline wander, hand movement, muscle artefact) were introduced using default values provided by the framework.

For any set of noise parameters, because the generator introduces stochastic variation depending on the random seed, two signals generated with identical parameters can still differ in their apparent noise content. This could lead to confusion, for example if a signal produced after increasing a noise parameter appeared cleaner than one at the previous level by chance. To address this, we generated 100 signals for each parameter setting and report the proportion of signals classified as acceptable. This averaging approach reduces the influence of random variation and ensures that trends across noise levels reflect the true tolerance of each SQI rather than chance fluctuations.

A parameter set was deemed to produce 'unacceptable' ECGs if fewer than 50% of signals were labelled by an SQI as acceptable, reflecting a majority-rule criterion. Within each noise category, noise levels were increased until all SQIs produced "Unacceptable" classifications.

Open-source SQIs

Each signal generated was passed through four SQIs. Although graphical user interfaces (GUIs) exist for some of these tools (*SQI3* and *SQI4*), these were bypassed for efficiency, instead using their code-based implementations.

SQI1: Orphanidou et al. 2015 The first three are feasibility rules applied to R–R intervals: (i) heart rate between 40–180 bpm, (ii) maximum interval between successive R-peaks $<$3 s, and (iii) ratio of maximum to minimum beat-to-beat interval $<$2.2. If the signal passes these checks, it is then evaluated using an adaptive template-matching step (threshold = 0.66), which measures the regularity of ECG morphology. Signals failing any step are labelled "unacceptable" [224]. Code was taken from [225].

SQI2: Zhao & Zhang 2018 Zhao & Zhang's method to evaluate signal quality combines simple heuristic fusion to extract features and fuzzy logic to evaluate quality [226]. Specifically, three measures are used: (i) the QRS wave power spectrum distribution (pSQI), (ii) kurtosis of the signal (kSQI), and (iii) baseline relative power (basSQI). The original paper also included an additional R-peak detection match (qSQI), but this was omitted in the available implementation. The SQI classifies signals into either 'Unacceptable', 'Barely Acceptable' or 'Excellent'. For consistency, we combine the latter

two into an 'Acceptable' category. We used the SQI as implemented in the `neurokit2` package [214].

SQI3: Kramer et al. 2022 Kramer et al. propose a three stage signal quality classification algorithm [227]. The first stage is a stationary signal check, which detects flatline segments where the ECG remains constant over short (0.2 s) windows. The second stage applies a feasibility rule requiring heart rate to lie between 24–300 bpm. The third stage is a signal-to-noise ratio (SNR) check, defined as the ratio of spectral power in the physiological band (2–40 Hz, covering P, QRS, and T waves) to power outside this range ($<$2 Hz and $>$40 Hz, representing baseline drift, EMG noise, and motion artefacts).The SNR is defined as:

$$\text{SNR} = \frac{\sum_{2}^{40} PSD(f)}{\sum_{0}^{2} PSD(f) + \sum_{40}^{250} PSD(f)} \tag{4.2}$$

where PSD(f) is the power spectral density estimated using a periodogram. Signals with SNR below 0.5 dB, a threshold determined from annotated training data, are classified as unacceptable, along with signals failing any other steps.

SQI4: Elgendi et al. 2023 Elgendi et al. extended Kramer's method by incorporating deep learning [228]. Signals are first transformed into spectrograms using a Short-Time Fourier Transform (STFT), which are then classified using a CNN–LSTM model. This approach allows both time–frequency characteristics and temporal dependencies to be captured.

Clinical Assessment

To contextualise the benchmarking results, we compared SQI outputs against clinical judgement. For each of the four noise sources and each SQI, we identifed the largest noise parameter value that still yielded an "Acceptable" label from the SQI. This represented the maximum noise level that is tolerated, and generate one ECG signal at this level (see Figure 4.1). This produced 16 signals in total (4 noise types × 4 SQIs).

All four SQIs were then applied to each of these 16 signals, meaning that for every signal we obtained four SQI classifications ("Acceptable" or "Unacceptable". Then, four criteria were selected to review these signals against, and were formed with clinical input. The cardiologist, blinded to the SQI outputs, independently assessed each signal using four criteria:

1. Can you estimate a plausible HR?
2. Can you locate all QRS complexes?
3. Can you locate all P and T-waves?
4. Is the signal clinically useful?

For criteria 4, 'clinically useful' was defined as 'allow full assessment of heart rate,

rhythm and beat-to-beat morphology'. We report the agreement between the SQI output and each of the four criteria responses by the cardiologist.

Therefore, for each of the 16 signals, we had an SQI classification of 'Acceptable' or 'Unacceptable', and an answer from the clinician as to whether the criteria was met. Agreement was quantified by counting, across the 16 signals, the number of times each SQI's binary decision (acceptable/unacceptable) matched the cardiologist's assessment under each criterion.

4.2.4 Results

SQI comparison

The threshold at which further increases in variation would lead to a SQI label of 'unacceptable' is recorded in Table 1. No amount of white noise was sufficient for *SQI3* to report the signal as unacceptable, reported as 'N/A'. For the *Motion Artifact* source, four different categories of artefact were assessed. An 'N/A' result here indicates that no *Motion Artifact* led to an unacceptable label; only *SQI1* labelled the ECG with 'walking' artifact as unacceptable whilst all other forms of motion artifact were acceptable by all other SQIs.

Table 4.1: Threshold at which ECG becomes 'unacceptable', for each noise source.

Noise Source	**Threshold for unacceptable label**			
	SQI1	**SQI2**	**SQI3**	**SQI4**
Heart Rate (bpm)	155	925	495	255
White Noise (dB)	4.75	1.41	N/A	1.32
Power Line (dB)	4.13	10.71	0.70	0.70
Motion Artefacts	Walking	N/A	N/A	N/A

SQI1 was the most sensitive to changes in both *Heart Rate* and *White Noise* (see Figure4.2). Figure 4.3 presents example ECG signals generated with *white noise* at a threshold that is 'Acceptable' (threshold < 0.5) for *SQI1* versus for *SQI2*. These ECGs were also labelled as clinically uninterpretable by the cardiologist.

Comparison with clinical expert

The agreement between the responses from the cardiologist with the SQIs is shown in Table 2. Each cell represents the number ECGs in which the cardiologist assessment, for each criterion, agreed with each SQI label.

Overall, *SQI3* had the lowest agreement with only 32 of the 64 total criteria from the cardiologist matching the SQI label across all 16 signals. *SQI2* had slightly higher agreement with 34 criteria matching the SQI label. Both *SQI1* and *SQI4* scored an equal total score for agreement between all four criteria and their SQI labels (42/64).

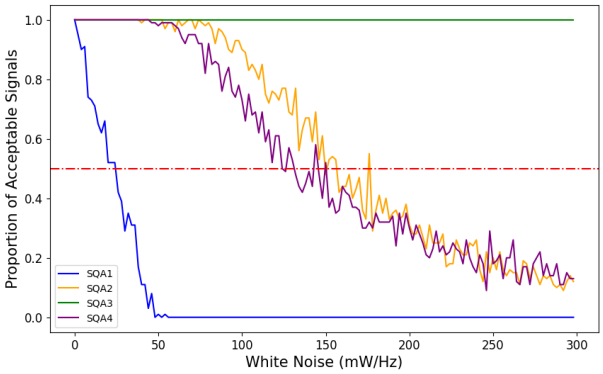


Figure 4.2: Increase in *White Noise* (original units) plotted against the proportion of 'acceptable' labels for each SQI tool. The highlighted threshold of 0.5 indicates the point at which ECG signals are considered to be unacceptable.

Table 4.2: The number of cardiologist assessments that agreed with the SQI labels. This is assessed for each of the 16 signals meaning each value can range 0-16 (a score of 16 shows that the SQI label matches the cardiologist criterion label for all 16 generated signals).

	SQI1	**SQI2**	**SQI3**	**SQI4**
Criterion 1	12	9	12	15
Criterion 2	10	11	12	13
Criterion 3	10	7	4	7
Criterion 4	10	7	4	7

SQI1 was the most consistent and showed similar agreement with the cardiologist across all four criteria (range 10-12). *SQIs 2, 3 and 4* however, displayed highest agreement with the cardiologist for criteria 1 and 2 but less agreement with criteria 3 and 4. *SQI3* in particular agreed with the cardiologist on criteria 1 and 2 for 12/16 ECGs but only agreed with criteria 3 and 4 for 4/16 ECGs.

Agreement between labelling from the SQIs and the cardiologist also differed by source of noise. For signals with *White Noise*, *SQI1* had the most agreement with the cardiologist. When adding *Power Line interference* and *Motion Artefacts*, all four SQIs showed similar levels of agreement with the cardiologist. For increases in *Heart rate*, *SQI4* had the most agreement with labels from the cardiologist.

We further noted that *SQI4* produced inconsistent results as heart rate increased. Although the initial threshold at which ECGs were 'Unacceptable' was at 255 bpm, further increases in heart rate, up to 400 bpm, were deemed to be 'Acceptable' again.

4.2.5 Discussion of Synthetic ECG Labelling

This study investigated the performance of four publicly available SQIs on synthetically generated ECGs with different modes of noise. Outputs were compared against the labels of an experienced cardiologist and several key findings emerged.

First, there was considerable inconsistency both between the SQIs themselves and

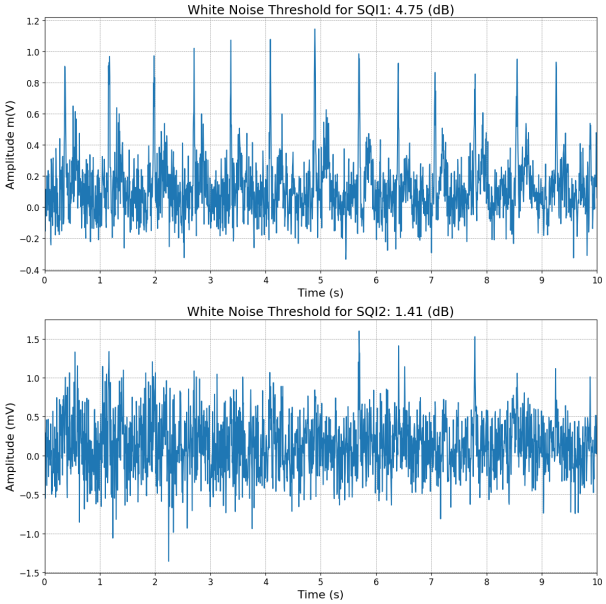


Figure 4.3: Comparison of ECG signals generated with White Noise set at the threshold between unacceptable and acceptable signals for SQI1 and SQI2. Clinical reviewer reported that the upper signal met criteria 1 and 2 but did not meet criteria 3 and 4. The lower signal did not meet any of the four criteria.

within the cardiologist's assessment. Among the four indices, *SQI1* and *SQI4* demonstrated the highest agreement with the expert opinion [224][228]. This was particularly evident when heart rate was increased; while *SQI2* and *SQI3* frequently classified ECGs as acceptable despite implausibly high heart rates, *SQI1* and *SQI4* were more in keeping with expected HR limits. In addition, *SQI1* was the only method to show some sensitivity to the addition of white noise. The overall level of agreement across all SQIs and all noise modes remained relatively low. This finding highlights the importance of critically evaluating SQI outputs, rather than assuming them to be universally reliable.

Second, performance varied depending on the noise type. On average, SQIs showed the highest agreement with the cardiologist when heart rate variation was introduced, and the lowest when white noise was added. The limited sensitivity to white noise additions is surprising, given that this is a common source of degradation in wearable ECG recordings.

Third, the analysis of agreement across the four clinical criteria revealed a clear pattern. The first two criteria related to whether HR can be extracted from the signal. For these, the SQIs showed moderate agreement with the cardiologist. The final two criteria related to other clinical features of the ECG (e.g. identification of P and T-waves) and exhibited lower levels of agreement. This suggests that current SQIs may be better suited for filtering signals for basic heart rate analysis than for more detailed morphological assessment.

It is important to acknowledge that the synthetic generator was used to deliberately

stress-test the SQIs, by introducing noise beyond typical clinical scenarios. This approach allowed systematic exploration of the limits of each SQI and highlighted potential weaknesses that might otherwise remain hidden. However, the heart rate thresholds observed for SQI2 and SQI3 are unreasonably high. Heart rates approaching 500–900 bpm are physiologically impossible in any context, regardless of noise conditions. The fact that these signals were labelled as acceptable highlights the lack of explicit physiological sanity checks in these tools. This represents a key limitation for free-living data used in preoperative assessment settings, where these errors could lead to extreme and misleading heart rate estimates. This raises concerns about the robustness of these SQI's when applied to real-world data in clinical settings.

In conclusion, the tested SQIs were inconsistent both with each other and with an expert cardiologist, limiting their suitability for clinical applications beyond the most basic classification of segments as "Acceptable" or "Unacceptable." These findings underscore the need for SQIs that are designed with specific use cases in mind, and particularly for tools that account for clinically relevant morphology rather than focusing solely on feasibility rules. Based on these findings, the next section outlines the rationale for selecting a single SQI to apply to the ECG signals in the REMOTES dataset.

4.3 Implementation for SQI for the REMOTES dataset

4.3.1 Selection of SQI

Based on the findings above, *SQI1* (Orphanidou et al.) was selected as the most suitable index for application to the REMOTES dataset. While both *SQI1* and *SQI4* demonstrated higher agreement with the cardiologist, *SQI4* displayed inconsistent behaviour when heart rate was varied. In particular, after initially rejecting signals above 255 bpm, it reverted to classifying extremely high and implausible rates (around 500 bpm) as "Acceptable". This poses a significant risk in free-living data, where physiologically implausible signals must be reliably excluded.

In contrast, *SQI1* consistently rejected signals once HR exceeded plausible thresholds, while also demonstrating greater sensitivity to white noise. This is particularly relevant given the prevalence of high-frequency artefacts observed in REMOTES ECGs (see Figure 3.9). Unlike other SQIs that tolerated noise until very low SNR thresholds were reached, *SQI1* was able to identify noisy signals earlier, reducing the likelihood of contaminated segments being passed through to downstream analyses.

For these reasons, *SQI1* was chosen as the primary tool for segmenting ECG signals into acceptable and unacceptable quality within REMOTES, balancing consistency, plausibility, and sensitivity to noise. However, given the the development in ECG processing since the publication of SQI1 (Orphanidou 2015), we propose that there is

scope to update the SQI in areas where specific progress has been made.

4.3.2 Updating SQI: Beat Detector Evaluation

A key component of SQI1 is the reliance on accurate R-peak detection. The feasibility checks within SQI1, such as verifying plausible heart rate and beat-to-beat intervals, are highly reliant on the quality of the beat detector used. Since the publication of SQI1 in 2015, new open-source detectors have been developed, with improved robustness to noisy or low-quality signals. In particular, a benchmarking study in 2024 demonstrated that the NeuroKit2 and NSW detectors consistently outperformed traditional methods such as Hamilton and Pan–Tompkins, particularly on low signal quality telehealth recordings [163]. This suggests that updating SQI1 with a modern detector may increase the proportion of usable signals in free-living ECG data.

To test this, we compared the original beat detectors used in SQI1 (Hamilton and Pan–Tompkins) against the NeuroKit2 detector. The comparison was conducted on the first 20 participants of the REMOTES dataset. For each participant, signals were segmented into 10-second windows and processed with SQI1 using each detector. We then calculated the percentage of windows labelled as 'Acceptable' under each condition.

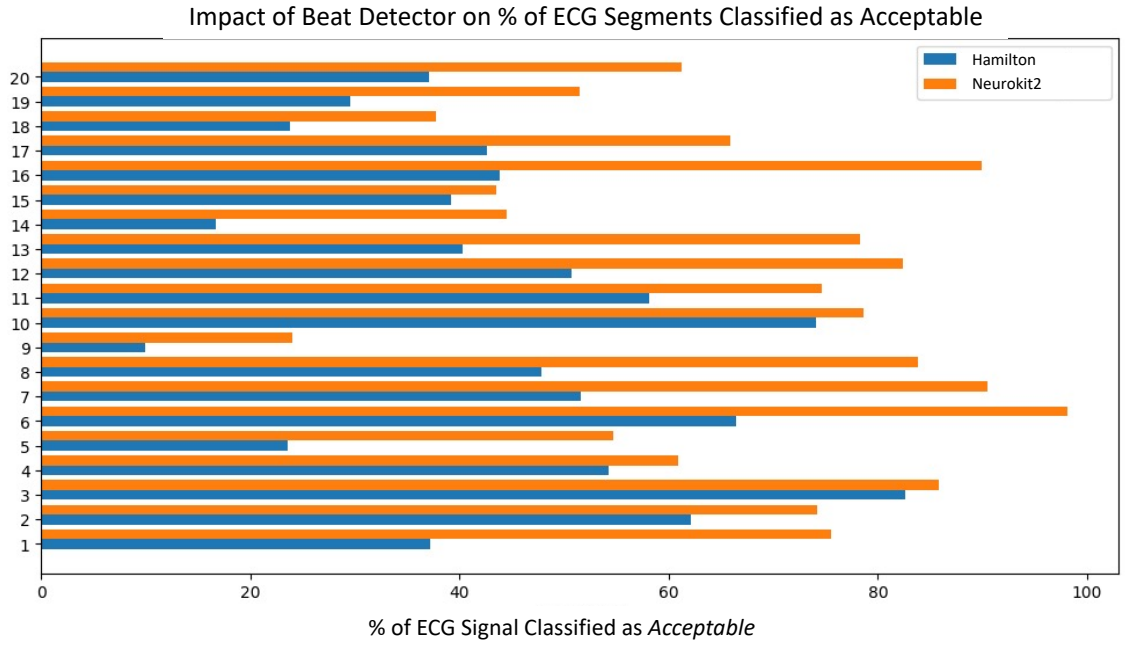


Figure 4.4: Comparison of beat detector performance within SQI1 across the first 20 participants of the REMOTES dataset. The plot shows the percentage of 10-second ECG segments classified as acceptable by SQI1 using different beat detectors. NeuroKit2 (orange bars) consistently allowed a higher proportion of segments to be retained compared with the Hamilton detector (blue bars).

Figure 4.4 shows the proportion of acceptable signals across participants. In all cases, the NeuroKit2 detector resulted in a higher percentage of windows being classified as acceptable compared with the original Hamilton detector. This indicates that NeuroKit2 was better able to tolerate noise and identify plausible R-peaks in challenging segments, thereby preserving more usable data for analysis.

To explore this difference further, we examined the participant with the largest discrepancy between detectors, where 44% of segments were accepted using Hamilton compared to 90% using NeuroKit2. Representative examples from this participant are shown in Figure 4.5. In both examples, the Hamilton detector located an additional beat (highlighted in yellow) that was not detected by NeuroKit2. This extra detection did not align with the expected beat pattern in the 10-second segment and was likely an artefact. The misclassification of this artefact caused *SQI1* to fail its final feasibility rule, in which the ratio of maximum to minimum R–R intervals must be <2.2. By contrast, NeuroKit2 avoided these detections, allowing the segment to be classified as acceptable.

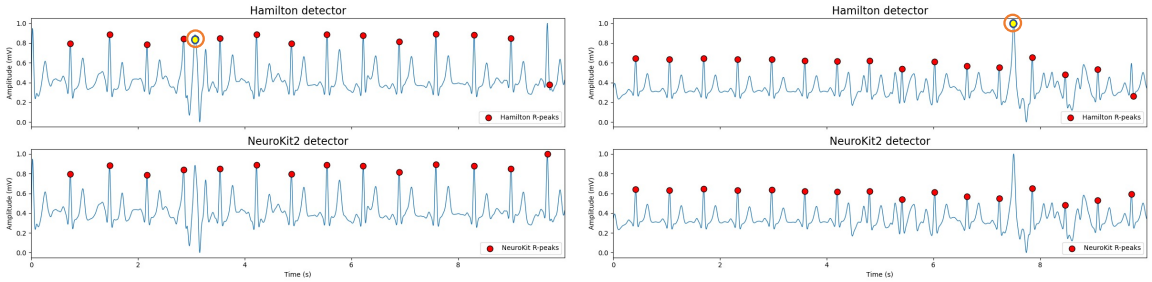


Figure 4.5: Comparison of beat detections from Hamilton (top) and NeuroKit2 (bottom) on two example 10-second ECG segments from the same participant. In each case, the Hamilton detector identified an additional beat (highlighted in yellow) that was not detected by NeuroKit2.

Although this investigation was limited to a smaller subset of participants, these results demonstrate that NeuroKit2 not only increased the proportion of accepted signals overall but did so by handling artefacts more robustly. Having established this improvement, the next question was how these findings translate across the two ECG channels collected in REMOTES.

4.3.3 Comparison of ECG Leads

The Ubiqvue-Lifesignals LX1550E chest sensor used in the REMOTES study records two channels of ECG via four chest-mounted electrodes (Figure 3.4). These leads are positioned adjacently within the device and are designed for long-term wearable monitoring rather than replicating the standard 12-lead clinical configuration. Both channels therefore capture modified chest leads, each subject to varying levels of electrode

noise.

Lead selection using the updated SQI

In processing long-term wearable ECGs, it is common to evaluate multiple channels and prioritise a single channel, rather than combining signals across leads [229, 230]. This approach ensures that downstream features are extracted from the cleanest available channel while avoiding propagation of artefacts. To assess which lead was most suitable in our dataset, we applied the updated *SQI1* (with NeuroKit2 beat detector) to a random subset of 20 participants. For each participant, we calculated the percentage of 10-second windows classified as acceptable in each lead. Results are presented in Figure 4.6.

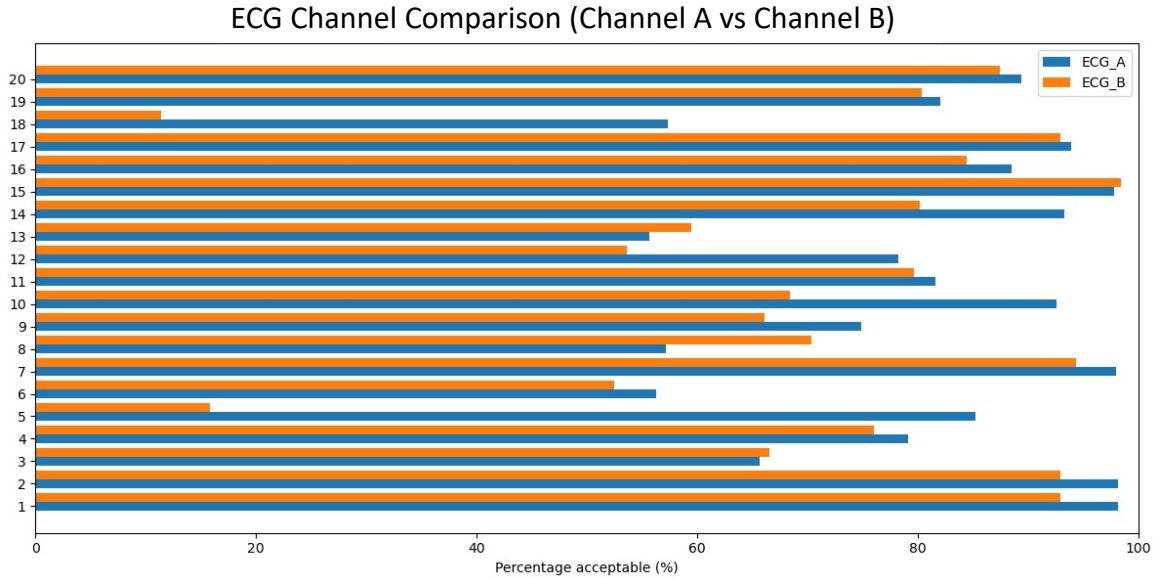


Figure 4.6: Comparison of signal quality between ECG leads. For each of the first 20 entries (y-axis indices 1–20), the horizontal bars show the proportion (%) of windows labelled Acceptable for ECG-A and ECG-B. Higher values indicate a greater fraction of usable data for that entry.

Across most participants (>75%), Lead A consistently yielded a higher percentage of acceptable segments than Lead B. Although this comparison indicates which channel would maximise usable data, further assessment of signal quality could confirm this.

Lead selection using a signal-to-noise ratio (SNR)

To complement the SQI analysis, we estimated the SNR using the calculation employed in *SQI3* as shown in equation 4.2, with identical preprocessing applied to both leads. This proxy is used for relative lead ranking rather than as an absolute SNR estimate. The metric defines “signal” as spectral power between 2–40 Hz and “noise” as power

outside this band. Any ECG content lying outside 2–40 Hz will be treated as noise, biasing the SNR downward. We therefore interpret SNR strictly as a comparative metric between leads under identical filtering.

For each of the 20 participants, we average the SNR calculated from each 10s window across both leads for comparison. Across the same subset, Lead A exhibited higher median SNR than Lead B for the majority of participants (18/20). Given that (i) Lead A yielded a higher proportion of acceptable windows by SQI (Figure 4.6) and (ii) demonstrated higher proxy SNR in most participants, we conducted all subsequent analyses using Lead A only. This choice prioritises data yield and ensures consistency by extracting features from a single channel rather than mixing across leads, which could introduce differences in feature distributions.

4.3.4 Acceptable ECG Data across Participants

All ECG signals were resampled to from the original 244Hz to 250 Hz and segmented into 10-second windows. Each window was processed with the updated SQI1 (NeuroKit2 detector, Lead A), and labelled with either a valid heart rate, or a value of 0 indicating poor quality (including missing data).

As previously described, an average of 89% of recording time was available across the cohort. When incorporating quality assessment, a mean of 74.25% (SD = 20.00) of ECG data was considered of ‘Acceptable’ quality across participants. Expressed in absolute terms, this corresponds to a substantially reduced number of usable hours of ECG data. Figure 4.7 illustrates the distribution of total device wear time (blue) and usable quality ECG data (green) across participants.

The figure shows that, while most participants adhered to the target recording duration (with a clear mode at 70 hours), the amount of usable ECG data was more widely distributed. In particular, many participants had between 40 and 70 hours of acceptable ECG, and the KDE curves highlight a flatter distribution for usable time, with fewer participants retaining very high proportions of clean data. This reflects the impact of motion artefacts, electrode detachment, and signal disconnect in free-living perioperative monitoring.

Despite the high level of missingness, SQI1 enabled extraction of periods of usable ECG for almost all participants. Although the usable yield is reduced compared with total recording time, the proportion retained is not considerably lower. This provides a realistic indication that the device was not optimal for continuous high-quality ECG acquisition, but still yielded data that should be sufficient to conduct downstream analyses.

Windows labelled as ”Unacceptable”/0 were not deleted but retained as explicit indicators of poor-quality or missing data, providing a temporal map of when the ECG signal was unreliable. Retaining these markers ensures transparency in data quality

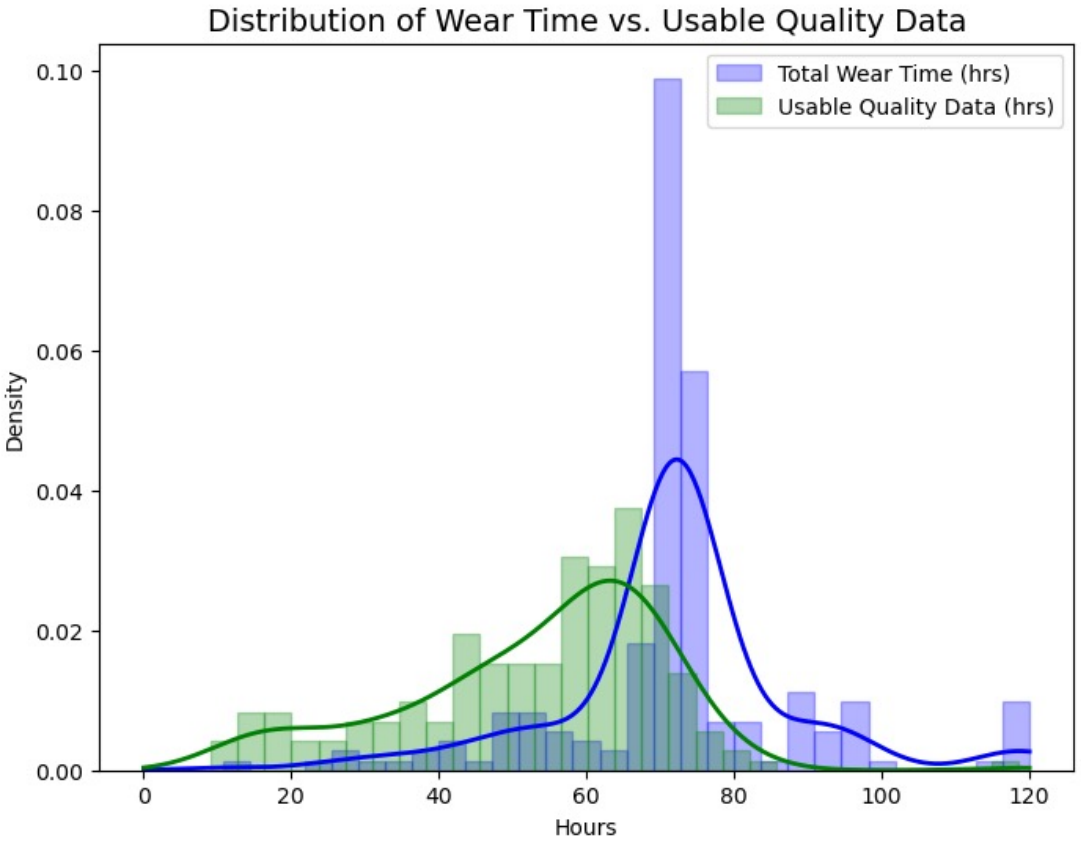


Figure 4.7: Distribution of total device wear time (blue) and acceptable quality ECG recording time (green) across participants. To highlight the distribution shape, each histogram is overlaid with a kernel density estimate, employed due to the non-normal distribution of the wear time. While the total wear time distribution shows a sharp peak around 70 hours, reflecting the target protocol, the usable quality distribution is broader and flatter, with most participants falling between 40 and 70 hours.

and is particularly useful when returning to the ECG data for more detailed analyses, such as HRV analysis using ECG signals. Although ECG segments labelled as 0 cannot be used for physiological analysis, knowing precisely when and how often the signal failed remains informative for assessing overall recording quality.

From the remaining acceptable ECG windows, heart rate (beats per minute) was calculated from each 10 second window and this produced a time-aligned HR signal in which valid heart rate estimates alternated with zero-labelled segments. From this HR (beats-per-minute) was calculated as the mean value across 'acceptable' segments.

The SQI filtering step was therefore a critical preprocessing stage for preparing the REMOTES dataset used in Chapters 5 and 6. It ensured that only physiologically valid data from the ECG signals are carried forward into downstream feature extraction and modelling, while maintaining full traceability of periods affected by noise or missing periods. The variation observed across participants also highlights the challenges of collecting ECG in free-living perioperative populations, where motion artefacts and electrode attachment issues are common. These signals form the basis for subsequent feature extraction and predictive modelling in the next chapter.

Chapter 5

Machine Learning Approaches to Predict VO_2max

5.1 Introduction

The preprocessing described in Chapter 4 provides the foundation for the next stages of the prediction pipeline: feature extraction and predictive modelling. In this chapter, we investigate which features can be extracted from the signals, before going on to assess how well wearable-derived features can predict VO_2max in a preoperative cohort.

As outlined in Chapter 1, VO_2max is a key measure of surgical risk, but CPET, its current clinical gold standard, has limited accessibility. Wearable sensors offer a scalable alternative by enabling continuous monitoring of heart rate and activity in free-living conditions. This approach also provides a more comprehensive view of the daily movements and physiological responses of patients compared to the snapshot provided by short-term tests. Prior research has demonstrated associations between simple features such as resting heart rate, step counts, and physical activity levels with VO_2max [231, 232]. Some research has validated these relationships in preoperative surgical populations, but this is less evidenced [59].

However, as discussed in Chapter 2.6.3, most existing studies rely on proprietary devices and pre-processed outputs, meaning that the precise methods used for feature extraction are often unclear. This limits reproducibility and makes it difficult to evaluate which features are most informative. Here, we use open-source tools where possible to extract a set of activity and heart rate features from raw wearable signals, providing a transparent account of how these features are derived and distributed across the cohort.

Various analytical approaches have previously been employed to predict VO_2max [233]. Multiple linear regression (MLR) remains a commonly used and effective method, while more recent studies have explored machine learning methods such as support

vector machines (SVM), random forests (RF), and multi-layer perceptrons (MLP). Although these approaches have reported promising results, they are typically applied to healthy populations rather than clinical cohorts, and there is little consensus on which models are best suited to wearable-derived features.

This chapter therefore has two of focus: first, to provide a transparent description of feature extraction from wearable ECG and accelerometer data; and second, to establish a benchmark for VO_2max prediction in a preoperative surgical population by comparing regression- and machine learning-based models. Specifically, this chapter will:

1. **Extract** a set of physical activity, and heart rate features from preprocessed wearable data.

2. **Compare the performance** of multiple regression- and machine learning-based models in predicting VO_2max.

3. **Provide an initial assessment of predictive performance** in a preoperative surgical cohort, forming the basis for the subsequent chapter that explore advanced features such as heart rate variability (HRV).

5.2 Materials and Methods

5.2.1 Dataset

This research used data from the REMOTES study, described in Chapter 3. All further details regarding data collection, the device used and cohort group can be found in section 3.1.1.

Given the variability in wearable data quality, the first challenge in the processing pipeline was to exclude participants whose recordings were insufficient to provide a representative estimate of cardiorespiratory fitness (CRF). This required balancing the need for strict thresholds to ensure data representativeness against the risk of excessive participant exclusion. Although there is limited evidence supporting a specific threshold, popular wearable sensors have commonly applied a minimum of 24 hours of data as the cut-off for reliable CRF estimation [234]. As the REMOTES study collected 72 hours of continuous data, less than 24 hours would equate to under one-third of the recording period being usable and would be unlikely to capture participants' typical routines. Therefore, participants were required to have at least 24 hours of acceptable quality data to be included in the analysis.

5.2.2 Feature Extraction

As outlined in Section 3.3.2, missingness patterns differed between ECG and accelerometer streams, motivating feature extraction from each modality to be completed separately where applicable. We begin with features derivable from accelerometry alone, focusing first on step counts, the most frequently reported movement metric.

Step Counts

Estimating steps from raw accelerometry depends on device placement and orientation. Because our accelerometer is mounted on a chest patch (a less common location in the step-counting literature), we selected an open-source, placement-agnostic method by Straczkiewicz and colleagues, released by the Onnela Lab, which has been validated across multiple body locations, conditions, and clinical cohorts (including cancer patients) [235].

Their method is based upon the principle that irrespective of sensor positioning, the accelerometer signal within the device will oscillate around a local average with a frequency matching that of performed steps. Using a continuous wavelet transform, the algorithm identifies the dominant frequency of these oscillations within short, non-overlapping one-second windows. This frequency is then translated into steps per second, and total step counts are obtained by summing across the entire observation period. Reported performance included small mean biases and narrow limits of agreement across validation settings (e.g., cross-body mean bias $\approx -0.5\%$; visually annotated datasets $\approx 0.1\%$; commercial wearable comparison $\approx 3.4\%$ difference). To implement the tool we used the publicly available open-source implementation from onnela-labs [1]. Step counts were extracted at a per-minute level to align with other physiological signals. From this, using the period of time of valid data collection across the period, we average the counts to extract a daily average step count.

MVPA steps

Cadence-based thresholds have been proposed as practical indicators of walking intensity, and time spent walking at a moderate-to-vigorous intensity. A review of 38 studies reported that a cadence of at least 100 steps per minute is a consistent marker of moderate-intensity activity in adults, across both laboratory and free-living settings [158]. Using the per-minute step count values generated from our data, we calculated for each participant, the average daily number of minutes spent walking at or above this threshold. Although this provides a useful indicator of ambulatory intensity, it should be noted that the $\geq$100 steps/min cut-off has not been specifically validated in pre-operative surgical populations. We did not identify an evidence base to support an

[1] https://github.com/onnela-lab/forest

alternative threshold in this setting, and therefore report results using the general population value, accepting that this may underestimate true moderate intensity walking in this cohort. Other measures of PA intensity are implemented alongside this.

Physical Activity Measures

Beyond step counts, activity intensity can be quantified directly from raw accelerometer signals. Several methods exist for categorising intensity into threshold-based groups such as sedentary behaviour (SB), light physical activity (LPA), moderate-to-vigorous physical activity (MVPA), and vigorous physical activity (VPA). As outlined in Section 2.4.1, simple cut-off thresholds are widely used, most commonly from hip- or wrist-worn ActiGraph devices. However, the applicability of such thresholds to chest-mounted accelerometers is not widely evidenced and has generally been implemented in only small sample sizes or in children [236, 237]. When attempting to estimate these thresholds, they appeared to overestimate PA classifications in this cohort.

To overcome this, we implemented the Oxford Biobank Accelerometer Analysis Tool [238], an open-source machine-learning pipeline originally trained on wrist-worn data (https://github.com/OxWearables/biobankAccelerometerAnalysis). Although not chest-specific, this tool produced estimates of activity that were more consistent with expectations in this cohort than simple cut-off approaches. Its use is further justified by the model structure: behaviour is first classified in 30-second windows using a balanced Random Forest trained on labelled data from the CAPTURE-24 study, with 100 trees and 50 rotation-invariant features in the time and frequency domains. To account for temporal dependencies, a Hidden Markov Model (HMM) was then applied, treating the Random Forest predictions as emissions and using the Viterbi algorithm to infer the most likely true behavioural sequence.

The model outputs five behavioural categories (sedentary, light, moderate, vigorous, and sleep). Validation of this approach showed good agreement with ground truth annotations, with performance reported across multiple age groups including adults, over 38 years [238].Although chest-specific validation is lacking, the robustness of the Random Forest features to orientation, combined with the HMM smoothing, supports the use of this model as a practical method for estimating physical activity intensity from chest-worn devices in this study. From these classifications, a signal of physical activity intensities from each participants was extracted and from this, their daily average time spent in each classification reported.

Heart Rate Features

Heart rate (HR) values were available as preprocessed 10-second averages following the procedures described in Chapter 4, see Figure 5.1. These values were considered

sufficiently accurate for feature derivation and were not subjected to further filtering. From these data, several simple summary measures were extracted, including maximum and minimum HR across the full recording period.

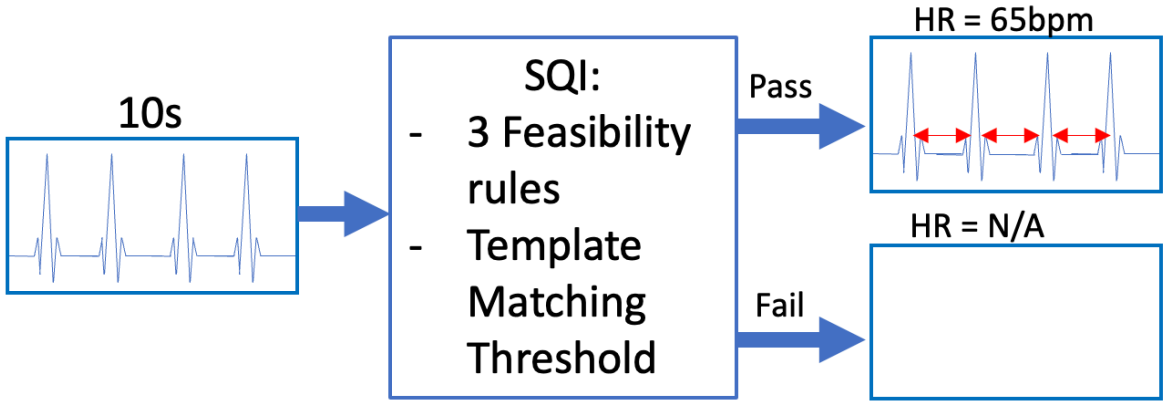


Figure 5.1: Example implementation of the SQI. The ECG is portioned into 10s-segments before passing each segment through the SQI tool. If passed and labelled as acceptable, HR and RR-intervals were extracted from the segment.

Resting heart rate (RHR) was derived separately, as its definition and extraction from wearable data are less standardised (see Section 4.2.3). Section 4.2.3 highlights the absence of a general consensus on how best to calculate RHR from free-living data. One approach proposed by Speed et al. [164], estimates RHR by averaging HR values recorded during the early morning hours (03:00–07:00) when individuals are most likely to be in a resting or sleep state. Following this rationale, we combined HR recordings with accelerometer-derived sleep or sedentary classifications and calculated the mean HR between 03:00 and 07:00 on each day of monitoring. This approach balances ensuring participants are resting whilst also not estimating RHR from a short 5–minute extract, providing a reproducible estimate of RHR from free-living wearable recordings.

Combining Signals

Some features can be derived by combining information from both HR and accelerometer signals. Since missingness patterns differed between modalities, features were initially extracted separately before integrating them. HR was downsampled to match the 30-second resolution of activity classifications, enabling calculation of average HR values during each physical activity class. This provides a simple measure of how HR adapts to differing activity.

Beyond this, previous work has highlighted the utility of combining cadence and HR as a proxy for CRF [231]. In the absence of speed or distance data, a cadence-to-HR ratio can serve as an alternative feature. Following this principle, we calculated the ratio of steps per minute to HR for all minutes with any detected steps (rather than

the 60 steps/min threshold used in prior work, to account for lower activity levels in this preoperative population). The distribution of cadence-to-HR ratios across all active minutes was then summarised at the 25th, 50th, 75th, and 95th percentiles. Lower percentiles reflect efficiency during habitual walking, whereas the upper tail (e.g., 95th percentile) should capture peak effort during higher-intensity activity. The 95th percentile was not included in the original research implementing cadence-to-HR measures, but we hypothesise here that including a high end ratio value could be a useful measure. High-end cadence or cadence-to-HR ratios have been associated with better functional capacity and health outcomes in other populations [239], supporting inclusion of this measure as a potential predictor of VO_2max.

Heart rate recovery (HRR) is often proposed as a useful indicator of cardiovascular fitness, typically defined as the reduction in heart rate within one minute of exercise cessation [83, 177]. In structured testing environments, such as treadmill or cycle ergometer protocols, this can be readily measured as $HRR_1 = HR_{peak} - HR_{1min\ post\text{-}exercise}$.

We attempted to implement this approach by combining accelerometer-derived activity classifications with HR time series. However, the free-living nature of the REMOTES data, combined with frequent missingness and variable signal quality, meant that suitable transitions were rarely captured with both modalities available and aligned. In many participants, missing or noisy ECG segments interrupted continuity at the precise moments needed to calculate HRR. As a result, HRR could not be reliably estimated across the cohort, and was excluded from further analysis.

5.2.3 Predictive Modelling

Using the extracted features, combined with age, gender and BMI, we evaluated five predictive models that have been commonly applied in wearable-based CRF research, see section 5.4.

- Multiple Linear Regression (MLR)
- Support Vector Regression (SVR)
- Random Forest (RF)
- XGBoost
- Multi-layer Perceptron (MLP)

This choice of models was guided by both implementations in previous research and the characteristics of our dataset. Linear regression was included as a transparent baseline model: it has been widely used in prior work and provides interpretable coefficients that allow direct comparison between features. Tree-based ensemble methods

(RF and XGBoost) were selected for their ability to model complex, non-linear relationships and interactions between features, and for good robustness to correlated predictors. SVR was included as a kernel-based method that has previously shown good performance in small-sample wearable studies. Finally, deep learning approaches are often applied in large datasets, and given the relatively small number of outcomes ($<$200) we did not pursue complex DL architectures. Nevertheless, we included an MLP to provide an initial benchmark for neural network performance relative to more shallow machine learning approaches.

5.2.4 Model Development

To account for inflated outcomes, VO2max values recorded from treadmill CPET were reduced by 10% to standardise results across testing modalities [207]. This adjustment aligns with guidance from the Association for Respiratory Technology and Physiology who note that VO_2max is typically between 5-10% higher during treadmill exercise compared to cycle ergometer, due to the involvement of upper body mass.

A correlation matrix (Pearson correlation coefficients) was computed to assess multi-collinearity. Features exhibiting a pairwise correlation coefficients $>$0.9 were identified; of these, only the feature with strongest univariate correlation with VO_2max was retained. We employed participant-level 5-fold cross-validation for model development and evaluation. The dataset was partitioned into five folds, with approximately 20% of participants held out for testing in each fold, and each participant appearing in the test set once. For each iteration, models were trained on four folds and tested on the remaining fold to allow independent evaluation on unseen participants. This allowed us to maximize the use of the small dataset while maintaining data separation between training and testing. Within each training set, feature selection (for Linear Regression and SVR) and hyper parameter tuning (for Random Forest, XGBoost, SVR, and MLP) were performed using internal cross-validation to prevent data leakage. For feature selection for MLR and SVR models, we used Least Absolute Shrinkage and Selection Operator (LASSO) regression with cross-validated alpha within each fold. For all models excluding MLR, hyperparameter tuning was performed using RandomizedSearchCV with 20 iterations and 3-fold cross-validation within each training fold.

The five models were implemented as follows:

- **Linear Regression (MLR):** Trained on LASSO-selected features using ordinary least squares.
- **Random Forest (RF):** Trained on all features. Hyperparameters included number of estimators (100–500) and maximum tree depth (10–50 or `None`). In addition, minimum samples required to split an internal node, minimum samples per leaf, and the number of features considered at each split were also tuned.

- **XGBoost:** Trained on all features. Hyperparameters included number of estimators (100–500), maximum depth (3–8), and learning rate (0.01–0.2). Subsampling of observations and column subsampling were also tuned to improve generalisability.

- **Support Vector Regression (SVR):** Trained on LASSO-selected features. The hyperparameter grid included C values between 0.1 and 100. Both linear and radial basis function (RBF) kernels were evaluated, with kernel-specific parameters tuned where applicable (γ set to either scale or auto for the RBF kernel).

- **Multi-Layer Perceptron (MLP):** Trained on all features. Hyperparameters included hidden layer sizes (3–11 neurons), learning rate (0–1), momentum (0–1), and logistic activation function. Models were trained using stochastic gradient descent with early stopping.

Model Evaluation

Performance was assessed using four metrics and averaged across folds, similar to those employed to evaluate previous prediction models [233]:

1. Root Mean Square Error (RMSE)

2. Standard Error of the Estimate (SEE)

3. Coefficient of Determination (R^2)

4. Pearson Correlation Coefficient (r)

Formal definitions and equations for each metric are provided in Section 2.5.5.

5.3 Results

198 participants were recruited into the REMOTES trial, of which one participant had a device malfunction and one further did not have a valid CPET. After preprocessing using the SQI tool selected and updated as described in 4.3.4, a further 9 participants were removed from analysis due to having under 24 hours of valid heart rate data.

The remaining participants had similar characteristics to those described in Section 3.2. On average, male participants were slightly older (68.9 years vs. 67.1 years) and had a lower BMI (28.2 vs. 30.5 kg/m^2) compared with females, while mean VO_2max values were higher in men (18.9 vs. 15.3 ml·kg^{-1}·min^{-1}).

5.3.1 Features

The characteristics and distributions of features extracted are described below. A full overview of features can be found in table 5.1.

Table 5.1: Descriptive statistics for wearable-sensor and demographic features (n = 187). Values are presented as mean (SD), median [IQR], minimum–maximum, and 95% confidence intervals for the mean.

Feature	Mean (SD)	Median [IQR]	Min–Max	95% CI (mean)
Daily steps (count)	3653.23 (2904.99)	2878.38 [1680.09, 5126.22]	24.58–15900.09	3236.86–4069.60
MVPA steps (count)	9.00 (13.05)	3.65 [0.85, 12.37]	0.00–98.77	7.13–10.87
Resting HR (bpm)	73.03 (9.05)	72.04 [66.62, 78.13]	52.26–106.07	71.73–74.33
Max HR (bpm)	135.11 (15.80)	135.00 [125.00, 145.00]	91.00–200.00	132.85–137.38
Min HR (bpm)	58.82 (7.83)	58.00 [53.00, 64.00]	44.00–84.00	57.70–59.95
Time in MVPA (min)	23.46 (31.20)	10.84 [3.47, 29.55]	0.00–164.78	18.99–27.93
Time in LPA (min)	88.88 (116.53)	29.81 [7.32, 138.14]	0.00–689.74	72.18–105.58
Time in sedentary (min)	778.58 (165.48)	768.16 [671.20, 889.33]	294.49–1336.13	754.86–802.29
Sedentary HR (bpm)	84.49 (9.68)	83.74 [78.57, 90.50]	60.30–116.23	83.10–85.88
Cadence–HR Ratio P25	0.07 (0.03)	0.07 [0.06, 0.08]	0.04–0.36	0.07–0.08
Cadence–HR Ratio P50	0.16 (0.10)	0.14 [0.10, 0.18]	0.05–0.79	0.15–0.18
Cadence–HR Ratio P75	0.37 (0.23)	0.30 [0.21, 0.45]	0.07–1.21	0.34–0.40
Cadence–HR Ratio P95	0.79 (0.26)	0.82 [0.63, 0.98]	0.14–1.37	0.75–0.83
BMI (kg/m^2)	28.86 (5.82)	27.90 [24.80, 32.70]	16.60–50.10	28.02–29.69
Age (years)	68.40 (10.99)	70.00 [61.00, 77.00]	29.00–92.00	66.82–69.97
Sex (n, % male)	135 (72.2%)	—	—	—

Daily Step Counts

Figure 5.2 shows the distribution of average daily step counts across the cohort. The median daily step count was 2,866, with an interquartile range of 1,680–5,035 steps. The distribution was strongly negatively skewed: although the mean was 3,604 steps per day, a substantial number of participants accumulated fewer than 1,000 steps. This indicates that, while a typical preoperative patient achieved low-to-moderate daily activity, a notable subgroup exhibited low mobility. At the upper end of the distribution, a small number of participants recorded very high step counts (up to 15,900 per day). These appear plausible given the broad range of participants included, but they represent only a handful of individuals and do not alter the overall pattern of limited activity in the cohort.

MVPA in Steps

The distribution of minutes spent in moderate-to-vigorous physical activity (MVPA), defined as walking at cadences $>$100 steps/min, is also shown in Figure 5.2B. The median was 3.5 minutes per day, with an interquartile range of 0.75–11.7 minutes. The distribution was sharply negatively skewed: the most common observation was close to zero minutes of MVPA, and the mean was only 8.8 minutes per day. This may reflect the cadence threshold being set too high for this group, or severe levels of inactivity. A small subset of participants did accumulate measurable amounts of MVPA

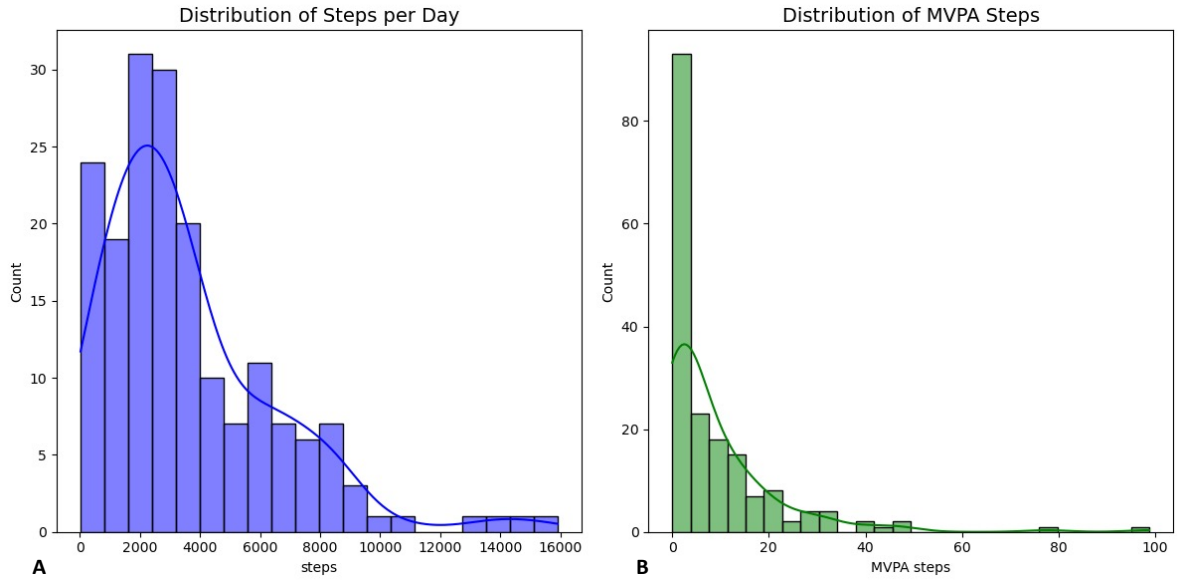


Figure 5.2: Fig A shows average daily step counts distribution across the cohort. Figure B shows the distribution minutes spent in MVPA walking (>100 steps / min), with minutes on the x axis.

(up to 99 minutes), but these were exceptions, underscoring the overall sedentary or low-intensity activity profile of the group.

Time in Physical Activity Intensities

Using the Oxford Biobank accelerometer classification tool, we derived the average daily time spent in different activity intensities (Figure 5.3). Sedentary behaviour (SB) dominated the daily profile, with a median of 772 minutes per day (IQR: 670–889), equivalent to roughly 13 hours. While SB should not normally include sleep, some misclassification between these behaviours may have occurred. Light physical activity (LPA) was limited, with a median of 30 minutes per day (IQR: 7–137), while moderate-to-vigorous physical activity (MVPA) was minimal, with a median of just 11 minutes (IQR: 3–29). Mean values followed a similar but slightly elevated pattern (SB: 780, LPA: 89, MVPA: 23 minutes). A small number of participants recorded unusually high MVPA (up to 165 minutes), but these were rare and did not alter the overall trend.

These results align with the step-count distributions described previously, highlighting the predominantly sedentary lifestyle of this preoperative cohort. However, HR values during LPA and MVPA could not be derived for participants who never engaged in these intensities (17 and 33 participants, respectively). Since these data were structurally absent rather than missing at random, MVPA-related HR features were excluded from further modelling.

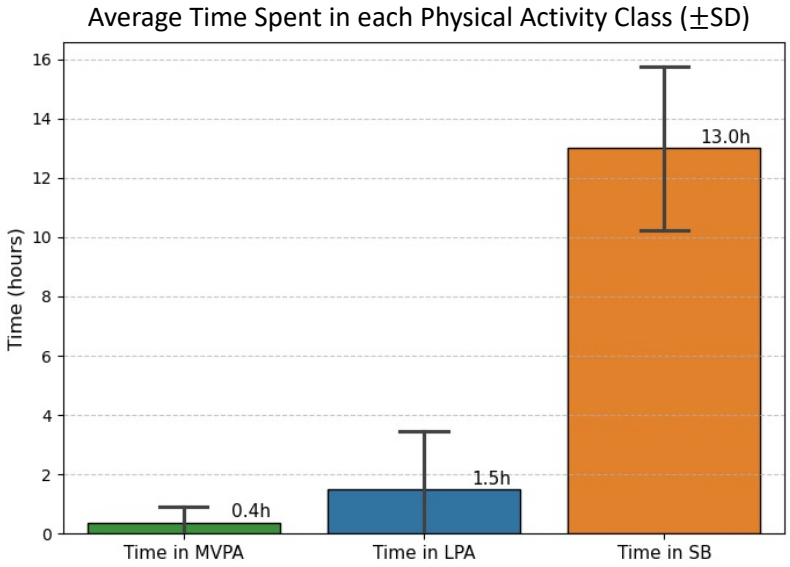


Figure 5.3: Average daily time spent in different physical activity intensities as classified by the Oxford Biobank tool. Bars show mean values with error bars representing standard deviation. Sedentary behaviour (SB) is shown in orange, light physical activity (LPA) in blue, and moderate-to-vigorous physical activity (MVPA) in green.

Resting HR

The distribution of resting HR values is shown in Figure 5.4. The median resting HR across participants was 72 bpm (IQR: 66–78), with a mean of 73 bpm. Almost all participants fell within the British Heart Foundation (BHF) guidance for normal adult resting HR (60–100 bpm) [240]. A small number of participants recorded values below and above this range (minimum 52; maximum 106 bpm), which could reflect underlying health conditions in this preoperative cohort or potential artefacts from noisy segments that were not fully filtered out. Overall, the distribution suggests that most participants presented with physiologically plausible resting HR values, consistent with an at-risk but largely stable preoperative population.

Minimum and Maximum HR

Across the full recordings, maximum HR values had a mean of 135 bpm (IQR: 125–145), with an upper range extending to 200 bpm, likely reflecting episodes of higher exertion in a small number of participants. This value remains within the expected range for maximum heart rates in this cohort, as estimated by the Tanaka formula $(208 - 0.7 \times \text{age})$, which projects an average maximum of around 160 bpm [241]. Minimum HR values had a median of 58 bpm (IQR: 53–64) and a minimum of 44 bpm. However, unlike resting HR, minimum HR was defined simply as the lowest observed value during the recording, without checks for adjacent accelerometer patterns or be-

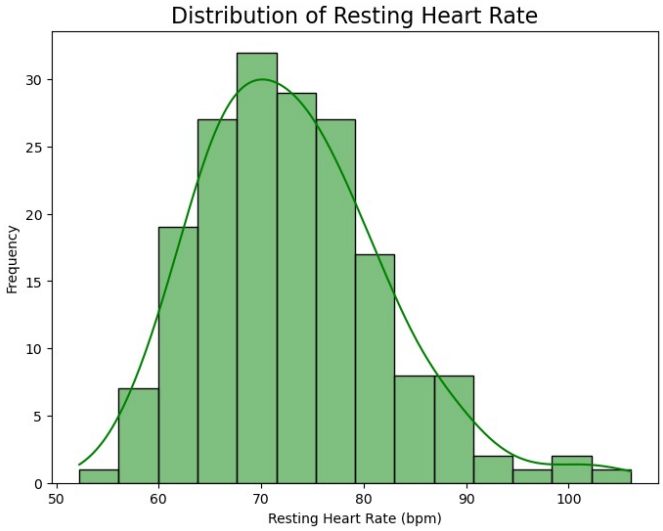


Figure 5.4: Distribution of resting heart rate (HR) across the cohort. Resting HR values are shown in green. Almost all participants fall within the British Heart Foundation guidance for normal adult HR (60–100 bpm) [240].

haviour classifications. As such, it may be more prone to reflecting artefacts or transient drops rather than physiologically representative values. Resting HR is therefore a more useful indicator of participants' lower-range HR, while minimum HR values should be interpreted with caution.

Patterns of Heart Rate and Activity

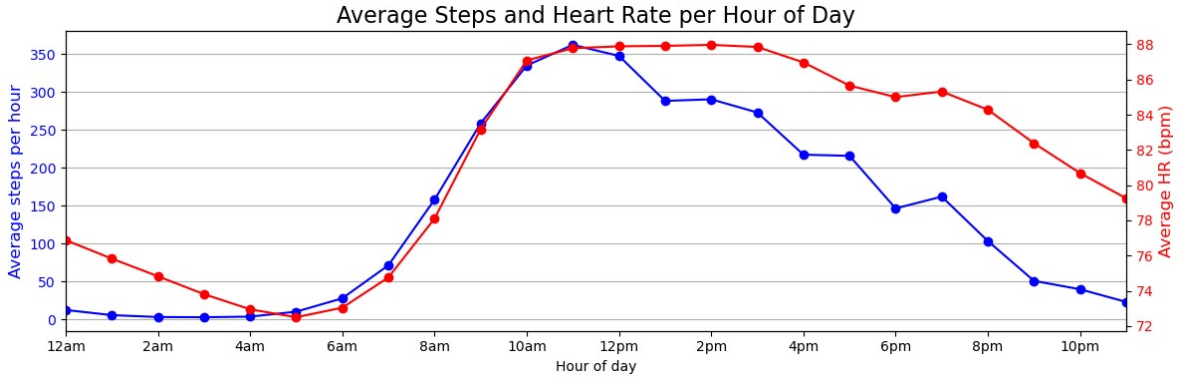


Figure 5.5: Average steps (blue) and heart rate (red) per hour of the day across the cohort. Axis labels for the step counts are shown in blue on the left y-axis and the axis labels for the average HR are shown on the right y-axis in red.

Figure 5.5 shows the average heart rate and step count averaged across each hour across the cohort. Both measures begin to rise around 6 am, increasing in parallel until midday, reflecting the onset of daily activity. Step counts then decline gradually from early afternoon, reaching low levels of under 100 steps per hour by 8 pm, while

heart rate remains elevated until around 4 pm before decreasing more gradually into the evening. Overnight, both step counts and heart rate remain low, with the lowest values occurring between 3–7 am. This coincides with the period used to estimate resting heart rate outlined in the methods, supporting its validity as method to estimate a true resting HR. These patterns are consistent with expected circadian rhythms and typical daily behaviour, indicating that the derived heart rate and step features reflect physiologically plausible activity and recovery cycles.

Cadence-to-HR Ratio

The median cadence-to-heart-rate ratio (50th Percentile) was low (0.14), indicating that in free-living conditions participants typically performed fewer than one step for every seven heartbeats (see appendix for distribution of cadence-to-HR ratios at the 25th, 75th and 95th Percentile, **??**). By contrast, the 95th percentile (P95) was much higher (0.81), showing that most participants were capable of achieving more efficient walking patterns during their most active bouts. This contrast highlights the difference between habitual activity (low-to-moderate efficiency) and occasional higher-intensity periods. While the 50th percentile summarises the central tendency, the 95th percentile may better reflect peak functional ability. The 25th and 75th quartiles showed similar patterns to the median.

5.3.2 Model Predictions

No features had a correlation above 0.9, meaning no features were removed due to high multi-collinearity. Among the models, Multiple Linear Regression (MLR) and Support Vector Regression (SVR) achieved very similar performance, both outperforming the more complex tree-based (RF, XGBoost) and neural network (MLP) models. Specifically, SVR and Linear Regression achieved almost identical performance in RMSE (3.56; 3.57), SEE (4.01; 4.03), R^2 (0.37; 0.36) and APE (15.95; 15.97), with SVR slightly outperforming. In contrast, Random Forest, XGBoost, and MLP all showed weaker predictive performance, particularly in terms of error metrics. These findings suggest that simpler regression-based approaches may be better suited to this dataset than more complex models.

The distribution plots in Figure 5.6B show that both the Linear Regression and SVR models produce predicted VO_2max values with a sharper, more peaked distribution centred around 17 ml·kg^{-1}·min^{-1} compared with the broader spread of the actual values. Despite this, the predicted values are generally well aligned with the overall range of observed data, indicating good model fit at the group level. The Linear Regression and the SVR models produce very similar prediction distributions, with only marginal differences: the Linear Regression model shows a slightly higher density between 17–

Table 5.2: Model Performance Metrics (mean and standard deviation) across the five folds.

Model	RMSE	R^2	Correlation	APE	SEE
MLR	3.57 (0.45)	0.36 (0.17)	0.63 (0.10)	15.97 (3.12)	4.03 (0.44)
RF	3.87 (0.56)	0.27 (0.13)	0.57 (0.11)	17.43 (2.84)	5.25 (0.77)
XGBoost	4.05 (0.65)	0.20 (0.13)	0.56 (0.12)	18.54 (2.81)	5.49 (0.90)
SVR	3.56 (0.45)	0.37 (0.14)	0.64 (0.10)	15.95 (2.86)	4.01 (0.46)
MLP	3.86 (0.71)	0.26 (0.19)	0.53 (0.14)	17.17 (2.51)	5.23 (0.98)

23 ml·kg^{-1}·min^{-1}, whereas SVR predictions are slightly lower and more tightly centred. Overall, the results reinforce that both models perform comparably, with negligible practical differences in prediction patterns.

Figure 5.6A shows the scatterplot of predicted versus actual VO_2max values for the Multiple Linear Regression model. The plot demonstrates a moderate degree of agreement between predicted and observed values, with the regression line indicating a positive correlation (r = 0.66).

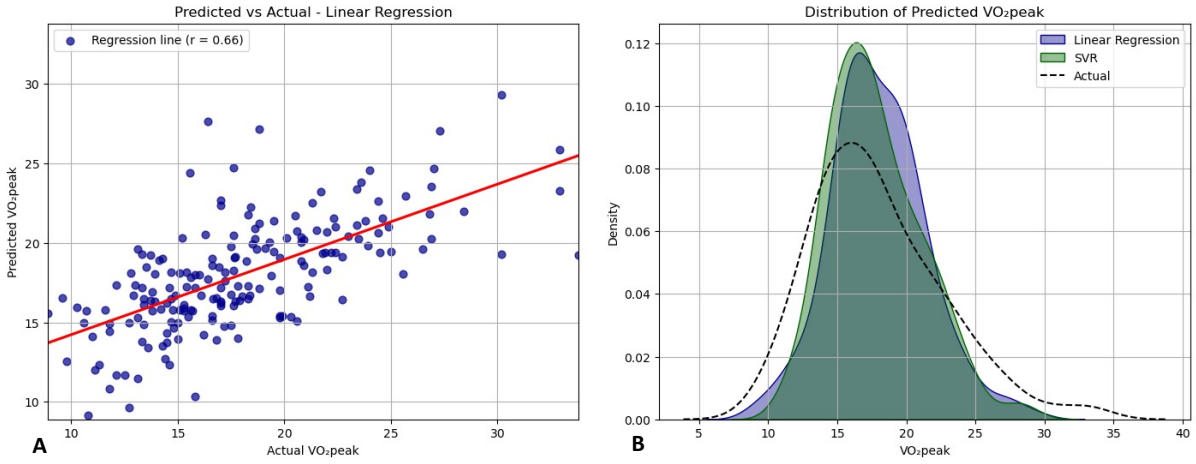


Figure 5.6: Figure A. Scatterplot to show the correlation between the predicted and actual VO_2max values from the Linear Regression model across participants. Figure B. A distribution plot showing the actual VO_2max values (black dotted line) against the predicted values from the SVR (green) and the Linear Regression models (blue).

5.4 Discussion

This study compared the performance of several machine learning models in predicting VO_2max from wearable sensor data in a preoperative cohort. Results showed minimal difference between Multiple Linear Regression (MLR) and Support Vector Regression (SVR), with both outperforming the more complex Random Forest, XGBoost, and Multi-layer Perceptron models. This suggests that, within this dataset, regression-based approaches capture the relevant relationships more effectively than non-linear

methods. This work represents the first direct evaluation of commonly applied machine learning models and wearable sensor features for estimating cardiovascular fitness in a clinical population.

The improved performance of MLR over other ML models may be attributed to sample characteristics. Previous studies in healthy populations have reported that more computationally advanced models outperform linear models [233]. However, this may not generalize well to a heterogeneous clinical cohort. Age and gender are known to be among the strongest predictors of VO_2max and to have relatively linear relationships with CRF. The wide distribution of age and physical activity measures like daily step-counts in this cohort may have reinforced the linearity of these relationships. Additionally, in a clinical sample, a physical activity measure like step counts may not only quantify fitness but also indirectly reflect general health status and disease burden. For example, individuals with fewer co-morbidities may have a higher functional status and therefore achieve higher daily step counts, rather than step count purely representing behavioural activity. This may explain why tree-based or neural network models, which excel in capturing non-linear patterns, provided no clear advantage here.

A further reason why the prediction performance in this study was lower than the highest-performing ML models reported in previous research may be explained by differences in predictor variables ($R>0.8$) [233]. Many high-performing VO_2max models include exercise test-derived variables. By including variables obtained during some form of exercise, even if sub-maximal, these are more likely to provide useful indication or capture aspects of maximal exercise capacity more directly than wearable-derived passive or free-living data. In clinical populations scheduled for major abdominal surgery, exercise test-derived variables are not always feasible and therefore wearable sensors provide an accessible alternative despite this limitation. When comparing against other research using only free-living data in healthy populations, results are somewhat more comparable [232].

A predominance of male participants and the presence of varying co-morbidities may also have influenced model performance, potentially favouring simple models that better capture dominant linear trends across the cohort. We considered developing separate models by sex; however, this approach has not been commonly implemented in previous research and would limit comparability. Additionally, clinical thresholds for VO_2max used in preoperative risk assessment do not differ by sex, and sex was included as an input feature across models to account for sex-specific differences. Maintaining a single model supports clinical application while ensuring sex-related variability is incorporated into model development.

This study also highlights methodological considerations for wearable data in clinical settings. The modest sample size relative to the number of potential predictors favours models that are robust to over-fitting and interpretable to clinicians. Linear regression fulfils both criteria and, given its comparable performance to SVR, represents

an appropriate baseline for subsequent exploration of more advanced features.

MLR and SVR achieved similar performance in predicting VO_2max from wearable data, with no clear benefit from more complex models. The linear model was selected as the basis for further work due to its interpretability and competitive accuracy. To build on CRF estimations, work will explore whether incorporating additional physiological features, such as variability measures, can enhance predictive performance.

Chapter 6

Assessing Heart Rate Variability

6.1 Introduction

In the previous chapter, it is demonstrated that common demographic and wearable-derived features, such as step counts and resting heart rate, could predict VO_2max to a reasonable degree. Among several predictive approaches, regression models performed as well as or better than more complex machine learning methods, with the added advantages of interpretability and clinical acceptability. Having established this baseline, the next step is to investigate whether more detailed physiological features can provide additional predictive value beyond those commonly implemented in previous research.

Heart rate variability (HRV) represents such a feature. Despite strong links between HRV, autonomic function, and health, its use in free-living wearable studies has been limited. This is largely due to technological constraints: most commercial devices provide only averaged heart rate values or photoplethysmography (PPG)-derived signals, which are often unreliable for HRV analysis. By contrast, the present study incorporates multi-day wearable ECG recordings, enabling a more robust investigation of HRV and its potential contribution to VO_2max prediction in a preoperative cohort.

6.1.1 Heart Rate Variability

HRV describes the variation in time intervals between consecutive heartbeats, typically measured as the difference between successive R–R intervals on an ECG recording (Figure 6.1). It reflects the function of the Autonomic Nervous System (ANS) and the Sinoatrial Node (SAN), providing insight into cardiovascular health and adaptability [242, 243].

Reduced HRV has been consistently associated with increased morbidity and mortality, while higher HRV is generally indicative of greater cardiovascular fitness and autonomic regulation [244, 245]. HRV has also been applied across diverse contexts

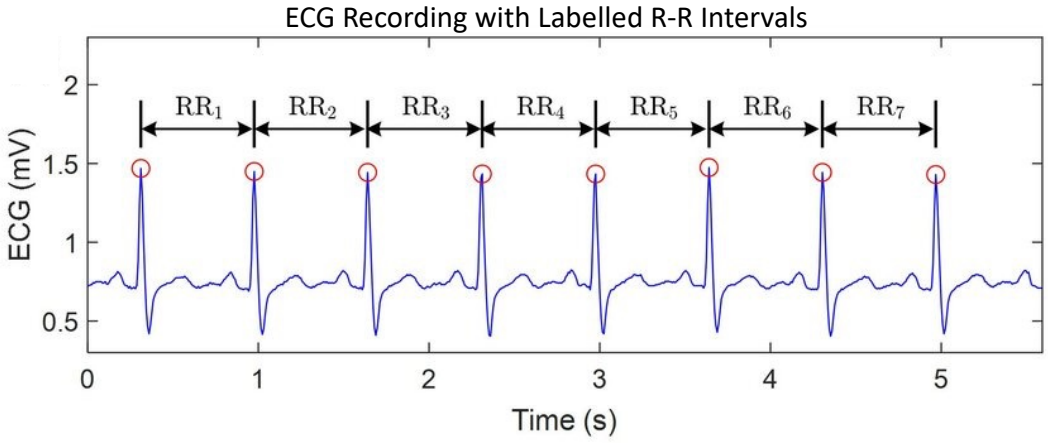


Figure 6.1: Adapted from Lu *et al.*, *Uncertainties in the Analysis of Heart Rate Variability: A Systematic Review*, *IEEE Reviews in Biomedical Engineering*, vol. 17, 2024. Licensed under CC BY 4.0.

including mental health, stress responses, and broader cardiovascular risk assessment [246, 247, 248]. These findings provide a strong rationale for investigating HRV in surgical populations, where cardiovascular health is a key determinant of perioperative risk.

HRV can be quantified in multiple ways. Standard time-domain measures include the standard deviation of normal-to-normal intervals (SDNN) and the root mean square of successive differences (RMSSD). Frequency-domain measures, obtained via spectral analysis, provide information about oscillatory components of autonomic control, such as high-frequency (HF) power linked to parasympathetic activity and low-frequency (LF) power reflecting mixed sympathetic–parasympathetic influence [249]. More advanced non-linear indices, such as entropy and detrended fluctuation analysis, capture the complexity of HR dynamics beyond linear measures. Together, these indices provide complementary perspectives on autonomic function.

6.1.2 Short- and Long-Term HRV

HRV can also be categorised by the duration of the cardiac recording [249]. Short-term HRV (typically 5–10 minutes) is the most common approach and captures autonomic fluctuations in the 0.03–0.4 Hz range. Long-term HRV (24 hours or more) reflects additional influences such as circadian rhythms and slower oscillatory processes, and is thought to more closely reflect SAN activity [243, 250].

Comparative studies suggest only moderate overlap between short- and long-term HRV. For example, Lu Fei et al. [251] reported a correlation of $r = 0.51$ between 5-minute SDNN and 24-hour HRV in post-infarction patients, with long-term HRV showing stronger predictive accuracy for one-year cardiac mortality.

Among long-term indices, the 24-hour standard deviation of N–N intervals (normal-to-normal heartbeats, after removal of artifacts and ectopic beats) (SDNN24) is particularly well established as a prognostic marker of cardiac health [242]. Early work showed that patients with SDNN24 values above 100 ms had over fivefold lower mortality risk following acute myocardial infarction compared to those with values below 50 ms [252]. Its prognostic utility has since been replicated in populations with and without cardiovascular disease [253, 254, 255]. However, SDNN24 requires continuous, noise-free RR interval detection over 24 hours, traditionally from Holter monitoring, which is resource-intensive and prone to artefacts in free-living conditions.

6.1.3 HRV in the context of Cardiorespiratory Fitness

The link between HRV and cardiorespiratory fitness (CRF) has been studied, but mainly using short-term HRV measured under controlled conditions prior to exercise testing [169, 256, 170, 257]. Results are mixed, with several studies reporting only weak associations with VO_2max. The lack of research using wearable sensors for HRV is notable; most studies measure HRV using a non-wearable ECG in the clinic prior to the CPET. This reflects the difficulties in deriving HRV from commercial devices, which often provide only averaged HR or PPG signals. This has reinforced a focus on short-term HRV in prior research, which can be collected in brief stationary conditions without wearables.

One research study did investigate the link between HRV features collected over the course of 24 hours hours using a Holter-ECG device [171]. In this study, SDNN24 was the only measure that was found to be independently associated with VO_2max. This highlights the potential of long-term HRV to provide additional predictive value for CRF. The present study is distinct in that it incorporates multi-day raw ECG from a preoperative cohort, enabling exploration of both short- and long-term HRV in a free-living clinical context.

One challenge remains with the data. As discussed in Section 3.3.3, wearable ECG in the REMOTES dataset is often affected by motion artefacts, signal loss and has variable signal quality. SDNN24 requires near-continuous RR detection, which can be difficult to achieve. Morelli et al. [243] proposed an alternative, SDANN$_{HR}$24, which approximates SDNN24 using semi-continous wearable HR rather than RR intervals thereby addressing these problems. While this approach may enable broader use in non-clinical settings, its application to CRF prediction remains unexplored.

6.1.4 Aims

In the previous chapter, we benchmarked multiple machine learning models, including both linear and non-linear approaches, for predicting VO_2max from wearable data, and

found that linear models had higher performance while offering greater interpretability. Building on this, the current study applies a linear model to evaluate the added value of short- and long-term HRV features over the standard wearable-derived features. We address the following research questions:

- Do HRV measures improve VO_2max predictions in a preoperative cohort?
- What is the relative contribution of short- and long-term HRV measures to VO_2max prediction models?

6.2 Methods

6.2.1 Study Description

A full description of the REMOTES trial, participant characteristics, CPET protocol and any adjustments made to the VO_2max output values are provided in section 3.2.

6.2.2 Signal Preprocessing

ECG preprocessing was performed as described in section 4.3, using the updated signal quality index (SQI) pipeline to identify ”Acceptable” segments and extract valid heart rate time series.

6.2.3 Standard Wearable Features

The set of standard features derived from wearable data (daily step counts, activity classifications, average heart rates by intensity, minimum/maximum heart rate, and cadence-to-HR ratio) were described in detail in section 5.2.2. These features are included here as baseline predictors.

In addition, we extracted a series of short-term and long-term HRV measures from the wearable ECG recordings, described in the following subsections.

6.2.4 Short-term HRV

Accurate estimation of short-term HRV requires clean ECG with reliably detected R–R intervals. Given the high prevalence of artefacts in free-living wearable data, the signal pre-processing protocol described in 4.3 was essential to ensure that only high-quality segments contributed to HRV analysis. This allowed us to identify windows in which RR intervals could be assumed physiologically valid, thereby minimising the risk of unusually small or large intervals caused by noise.

Short-term HRV is conventionally derived from ECG recordings of 5 minutes in length, a duration shown to be sufficient to capture stable autonomic fluctuations across standard indices [249, 242]. Standardising feature extraction across the cohort to a time when participants' autonomic function was least likely to be impacted by external factors was important to ensure validity. HRV was calculated during sleep as this has generally shown to be a stable state for most HRV features, particularly high frequency measures [258]. A single night of sleep has also been shown to provide representative estimates of HRV features, making this approach both feasible and reliable in free-living studies [259].

For each participant, we identified a 5-minute ECG segment recorded between 3:00–7:00 a.m. that was entirely labelled as 'acceptable' by the SQI. Segments also had to have activity classifications of 'sleep' or 'sedentary behaviour'. Segments were excluded if no such clean window was available. Participants without such a period were excluded from analysis. Ectopic beats were identified and linearly interpolated, and participants with 50 ectopic beats within the 5-minute period were excluded, consistent with prior recommendations [260].

From the accepted 5-minute windows, ten HRV measures were derived using the `hrv-analysis` Python package. These included time-domain, frequency-domain, and non-linear indices (Table 6.1), providing a comprehensive representation of short-term autonomic function.

Table 6.1: Short-term HRV features extracted from a 5-minute ECG segment.

Feature	Description
Time-Domain Measures	
SDNN	Standard deviation of NN intervals (overall HRV).
RMSSD	Root mean square of successive differences between NN intervals.
pNN50	Percentage of consecutive NN intervals differing by ¿50 ms.
MeanNN	Average NN interval (mean HR across the period).
Frequency-Domain Measures	
VLF	Power in the very-low-frequency band (0.003–0.04 Hz).
LF	Power in the low-frequency band (0.04–0.15 Hz).
HF	Power in the high-frequency band (0.15–0.40 Hz).
LF/HF	Ratio of LF to HF power.
Non-Linear Measures	
SD1	Short-term variability from the Poincaré plot.
SD2	Long-term variability from the Poincaré plot.

6.2.5 Long-term HRV

In addition to short-term indices, we also investigated long-term HRV measures, which capture slower circadian components of autonomic activity. These measures typically

require 24-hour continuous ECG recordings and have been widely associated with cardiac prognosis [252, 255]. However, their calculation in free-living wearable data is challenging due to noise and missingness.

For consistency across participants, Day 1 of recording was selected, as this showed the lowest average proportion of missing data (17.5%). To ensure sufficient representation of circadian activity, participants with fewer than 16 hours of valid HR data were excluded. While there is no universally agreed minimum duration, the European Society of Cardiology recommends near-complete 24-hour recordings to capture both daytime and nocturnal autonomic patterns [249]. Additionally, recent work has shown that time-domain HRV metrics remain relatively stable up to 35% degradation, whereas frequency-domain metrics degrade much earlier (10%) [261]. Therefore, choosing a threshold that corresponds to 30% missingness should preserve the reliability of time-domain long-term HRV features, whilst capturing autonomic function across the daily period.

SDNN24 The standard deviation of all NN intervals over 24 hours (SDNN24) was calculated from beat-to-beat intervals:

$$SDNN24 = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(RR_i - \overline{RR}\right)^2}$$

where RR_i represents individual RR intervals, $\overline{RR}$ the mean RR interval, and N the total number of RR intervals. SDNN24 remains the most widely documented long-term HRV measure.

SDANN$_{HR}$24 To address the issue of noise and missingness inherent to wearable ECG data, we also implemented SDANN$_{HR24}$, a metric recently proposed by Morelli et al. [243]. This approach replaces RR intervals with heart rate (HR) data, segmented into fixed-length windows of either 1, 5, 10, 30, and 60 minutes, thereby enabling robust estimation even when continuous beat-to-beat data are unavailable. For each segment i, HR was converted to cycle length $(60/HR_i)$, and variability was estimated across segments as:

$$SDANN_{HR24} = \sqrt{\frac{1}{N_{\text{segments}}}\sum_{i=1}^{N_{\text{segments}}}\left(\frac{60}{HR_i} - \frac{1}{N_{\text{segments}}}\sum_{j=1}^{N_{\text{segments}}}\frac{60}{HR_j}\right)^2}$$

Only segments with $\leq 50\%$ missing HR data were included. This method has been validated against standard SDNN24 in free-living wearable datasets and offers a pragmatic alternative for long-term HRV assessment where noise and missing periods of signal are unavoidable [243].

6.2.6 Model Development

The same process as that outlined in in section 5.2.4 was implemented for model development. Inflated outcomes associated with treadmill testing were first reduced by 10% and a correlation matrix was computed to assess multi-collinearity amongst features. Features exhibiting a pairwise correlation coefficients 0.9 were identified; of these, only the feature with strongest univariate correlation with VO_2max was retained. HR during LPA and MVPA was missing for participants who never engaged in these intensities and these features were removed from analysis.

6.2.7 LASSO Regression

In the previous chapter, we used LASSO (Least Absolute Shrinkage and Selection Operator) regression as a feature selection tool for the multivariable linear regression and the SVR. Here we directly implement the LASSO regression (Least Absolute Shrinkage and Selection Operator) to predict VO_2max and embed automatic feature selection [262]. LASSO encourages sparsity by shrinking irrelevant or redundant features to zero and is effective when dealing with highly correlated features. We chose to use LASSO over other dimensionality reduction techniques (e.g. Principal Component Analysis) to prioritise interpretability. Unlike PCA which produces transformed features, LASSO preserves the features. Two separate LASSO models were trained: a baseline model (without HRV features) and a model including HRV features. The dataset was split using 5-fold cross-validation with each participant included in the test set once. In each fold, the model was trained on four subsets and evaluated on the remaining subset. An inner 5-fold cross-validation was performed within each training set to optimize the L1 regularization parameter (λ) to minimise mean squared error. All predictors were standardised (z-scored) ensuring comparability. We also stratified performance by gender.

6.2.8 Model Evaluation

As described in section 5.2.4, model performance was evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Absolute Percentage Error (APE), R2 and Pearson's Correlation. A correlation plot of predicted versus observed VO_2max across all participants was used to assess strength of the relationship between the values. Differences in fold-level metrics between models were assessed using paired t-tests.

6.2.9 Assessing Feature Importance

The size of model coefficients may misrepresent feature importance in the presence of collinearity or feature interactions [263]. We instead calculate SHAP values for each

feature in the HRV model to assess contributions to predictions [264]. This approach is grounded in cooperative game theory that attributes a prediction to each feature by considering its marginal contribution across all possible subsets of features, meaning the sum of feature attributions will equal the models total prediction. Assessing each features contribution to the models prediction is particularly valuable in clinical applications like this study, where interpretability of feature importance is crucial.

SHAP analysis was incorporated into the 5-fold cross-validation framework; SHAP values were computed for test sets in each fold and aggregated to rank features by their mean absolute SHAP value. SHAP values were presented to evaluate absolute feature contributions.

6.3 Results

A total of 198 participants were recruited into the REMOTES study. Exclusions were applied as follows:

- 1 participant: device error
- 5 participants: $<$24 hours of total data collected
- 17 participants: $<$16 hours of heart rate (HR) data on day one
- 5 participants: $>$50 ectopic beats within the 5-minute sleep ECG
- 6 participants: no clean 5-minute ECG available
- 1 participant: incomplete CPET

After exclusions, 163 patients remained (Table 6.2).

Table 6.2: Descriptive statistics, mean (standard deviation), of the research cohort split by gender, showing wearable device wear time, measured VO_2max, and average daily step counts.

Variable	**Men (n=120)**	**Women (n=43)**
Age (years)	68.50 (10.25)	66.91 (13.42)
BMI (kg/m^2)	28.15 (4.98)	30.39 (7.53)
Collected data (hours)	74.33 (16.28)	74.04 (18.22)
VO_2max (ml/kg/min)	18.81 (4.74)	15.19 (3.71)
Average daily step count	3854 (3091)	2749 (2233)

6.3.1 Feature Correlations

Daily steps showed the strongest correlation with VO_2max ($r = 0.51$, $p < 0.01$), followed by age ($r = -0.41$, $p < 0.01$). Among the short-term HRV measures, LF/HF ($r = 0.35$, $p < 0.01$) showed the highest correlation with VO_2max. For long-term HRV measures, $SDANN_{HR}24$ calculated with 60-minute windows demonstrated the strongest correlation across its variations ($r = 0.35$, $p < 0.01$), while SDNN24 displayed the weakest ($r = 0.28$, $p < 0.01$), as shown in Figure 6.2A. All variations of long-term HRV were significantly correlated with VO_2max, with a general trend of increasing correlation strength as the $SDANN_{HR}24$ window length increased from 1 to 60 minutes.

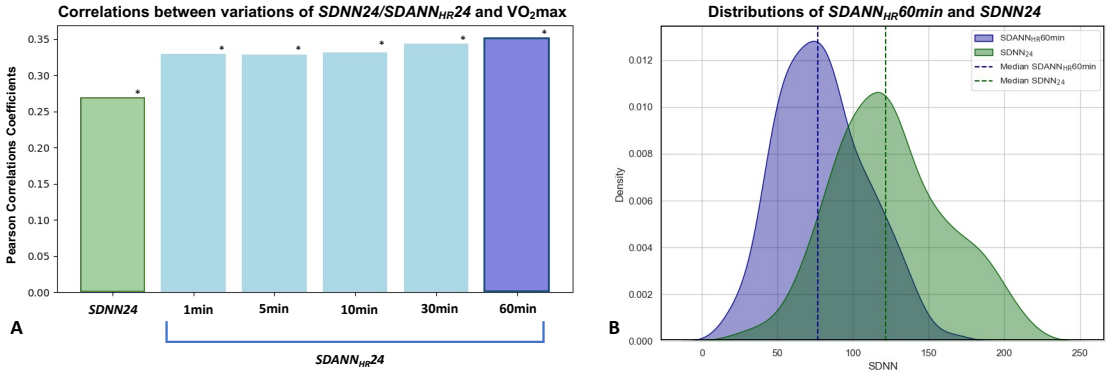


Figure 6.2: A: Correlations between VO_2max and long-term HRV features (SDNN24 and $SDANN_{HR}24$ at varying window lengths). Asterisks (*) denote significant Pearson correlations at $p < 0.05$. B: Distribution plots of SDNN24 (green) and $SDANN_{HR}24$ at 60 minutes (dark blue), with median values shown as dotted lines.

When assessing pairwise correlations between all input variables, SD1, SDNN, RMSSD, SDNN24 (Table 6.1), and all temporal variations of $SDANN_{HR}24$ were excluded due to high multicollinearity ($r > 0.9$), aside from the feature with the highest correlation to VO_2max. This is unsurprising, since the $SDANN_{HR}24$ measures are derived from the same underlying calculation using different segment lengths, making them inherently collinear. The 60-minute $SDANN_{HR}24$,which showed the strongest univariate correlation with VO_2max, was retained for further analysis.

As shown in Figure 6.2B, the distribution of $SDANN_{HR}60$min values is narrower and shifted towards lower magnitudes compared to the corresponding SDNN24 distribution. Specifically, $SDANN_{HR}60$min exhibits a sharper peak with a lower mean (79.5 ms) and standard deviation (28.9 ms), whereas SDNN24 displays a broader distribution extending up to 250 ms, with a higher mean (124.1 ms) and greater variability (37.5 ms). These differences highlight the attenuation introduced when deriving NN values from averaged heart rate segments rather than directly from beat-to-beat NN intervals.

After these exclusions, 16 features were included in the baseline model and 24 features in the HRV model.

6.3.2 Multivariable Regression

Table 6.3: Comparison of model performance between the baseline model and the HRV model (LASSO regression). Values are mean (standard deviation) across five folds. The rightmost column shows the mean (SD) difference: HRV minus Baseline.

Metric	Baseline	HRV	Difference
RMSE (ml·kg^{-1}·min^{-1})	3.54 (0.57)	3.38 (0.53)	-0.17 (0.13)
R^2	0.42 (0.13)	0.47 (0.12)	$+0.05$ (0.04)
Correlation (r)	0.66 (0.10)	0.70 (0.08)	$+0.04$ (0.02)
APE (%)	16.22 (2.45)	15.55 (2.19)	-0.68 (0.97)
MAE (ml·kg^{-1}·min^{-1})	2.77 (0.38)	2.63 (0.34)	-0.14 (0.16)

The baseline model presented here in Chapter 6 is conceptually the same model as the MLR presented in Chapter 5. However, small differences in the results are observed due to a reduced subset of participants being selected for analysis here. This reflects the stricter inclusion criteria required for reliable HRV feature extraction, which further excluded a small number of participants for whom HRV could not be extracted.

In the HRV model, daily step count, age, BMI, sex, LF/HF, and $SDANN_{HR}24$ (60 min) retained non-zero coefficients in every fold, indicating consistent contribution to predictions. Adding HRV improved performance across all metrics (Table 6.3): R^2 increased from 0.42 to 0.47 and correlation from 0.66 to 0.70, with corresponding reductions in RMSE, MAE, and APE. The standard deviations for the RMSE/R^2/correlation differences indicate stable gains across folds, whereas APE/MAE showed slightly greater variability. Performance was higher in women (HRV model R^2 = 0.49) than in men (R^2 = 0.40).

Paired *t*-tests on the fold-level results indicated that none of the observed differences between models reached statistical significance. There were non-significant increases in R^2 ($+0.05 \pm 0.04$, $t = 2.69$, $p = 0.055$) and non-significant reductions in RMSE (-0.17 ± 0.14, $t = -2.60$, $p = 0.060$), MAE (-0.13 ± 0.18, $t = -1.69$, $p = 0.166$), and APE (-0.68 ± 1.09, $t = -1.39$, $p = 0.236$).

6.3.3 SHAP analysis.

SHAP confirmed that anthropometric features (age, BMI, sex) were, as a group, the largest contributors to model output, alongside daily step count. Among HRV features, the long-term index $SDANN_{HR}24$ (60 min) exhibited a higher mean SHAP contribution than short-term indices (e.g., LF/HF, MeanNN), suggesting additional predictive signal in slower, circadian-scale variability (Figure 6.3B).

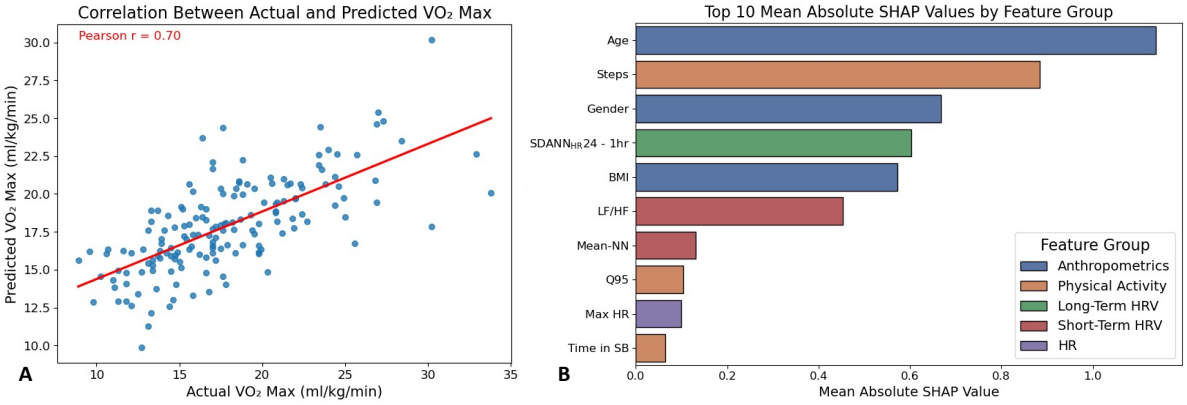


Figure 6.3: (A) Scatterplot of observed vs. predicted VO_2max (all folds combined showing all participants) with regression line. (B) Mean absolute SHAP values (top 10 features) aggregated across folds; features are colour-grouped (anthropometrics, activity, HR/HRV).

6.4 Discussion

This study investigated whether HRV measures derived from wearable sensors improve remote VO_2max prediction in a preoperative clinical cohort. Including HRV features led to modest but consistent improvements over the baseline model without HRV. Notably, measures derived from long-term HRV contributed more strongly to predictive accuracy than short-term HRV measures.

In this study, the HRV model outperformed the baseline, achieving a correlation of 0.70 (compared to 0.66 from the baseline model) and an MAE of 2.63 ml/kg/min, indicating reasonable predictive accuracy. Although the fold-level comparisons consistently favoured the HRV model, the differences did not reach statistical significance, reflecting both the modest effect size and the limited number of folds available for formal statistical testing. Nonetheless, the consistent direction of improvement across all performance metrics suggests that HRV may provide some additional predictive signal. Given the clinical heterogeneity of the cohort and the small evaluation sample, these findings should be interpreted as some preliminary evidence of benefit that can be explored further.

The addition of HRV features was consistent with improved performance across metrics, but a larger sample size would be required to have confidence that the small gain is a true effect. As discussed previously (Section 5.4), prediction performance in this cohort is potentially limited by factors beyond what wearable signals can capture. VO_2max in surgical patients reflects not only autonomic function and daily activity, but also disease status, organ function, and medication use. This broader clinical heterogeneity reduces predictive power compared to studies in healthy populations, where lower inter-individual variability may facilitate stronger associations.

6.4.1 Long-term HRV

When taking a closer look at the distributions of long-term HRV measures, clear discrepancies were observed between SDNN24 and SDANN_{HR}60min (Figure 6.2B). While SDANN_{HR}24 60min displayed a narrower distribution with a mean of 79.5 ms, SDNN24 had a substantially higher mean (124.1 ms) and broader variability, with several participants exceeding 150 ms and even a small number above 200 ms. These values appear implausibly high for a preoperative cohort with a mean age of 68 years. For reference, Kleiger et al. classified 24-hour SDNN values below 50 ms as indicative of poor health, 50–100 ms as compromised, and above 100 ms as healthy [252]. Similarly, Yilmaz et al. [265] reported normative values in healthy adults of 141 $\pm$ 39 ms. In the present study, the high proportion of participants with values well above 150 ms is unlikely to reflect true autonomic function, and instead suggests inflation due to residual noise or artefacts, even after applying a robust SQI pipeline.

This discrepancy does not necessarily indicate that SDANN_{HR}24 is inherently more informative as a marker of cardiorespiratory fitness, but rather that it is more resilient to signal quality limitations. By deriving variability from averaged HR segments, SDANN_{HR}24 attenuates extreme inter-beat fluctuations caused by suboptimal RR detection, yielding distributions that appear more physiologically plausible for this cohort. The relative stability of SDANN_{HR}24 compared to SDNN24 highlights the continued importance of addressing signal quality in wearable ECG, and suggests that HR-derived indices may currently offer the most reliable long-term HRV features for free-living applications.

6.4.2 Feature Contributions

When examining feature contributions, age, step-counts, and gender were the strongest predictors. This aligns with extensive research demonstrating the well-established link between age and CRF, as well as the validity of step counts as a health marker [266, 267]. However, the long-term HRV feature SDANN_{HR}24, specifically calculated from 60-minute windows, also ranked among the top contributors, above BMI and other HRV features. This supports prior findings suggesting that ultra-low frequency (ULF) components of the cardiac signal captured in long-term HRV may better differentiate high-risk patients than the high-frequency components derived from short-term HRV [243].

Importantly, SDANN_{HR}24 offers several advantages for wearable applications that should be considered alongside its predictive gains. Unlike short-term HRV measures, which require precise beat-detection from ECG signals, SDANN_{HR}24 can be derived from semi-continuous HR signals that act as a low-pass filter for inter-beat-intervals. This makes it compatible with existing digital infrastructure in perioperative research

using commercial wearable devices. Even if prediction improvements are modest, its ease of implementation make it a practical addition for remote CRF assessment. Although $\text{SDANN}_{\text{HR}}24$ offers greater resilience to noise, there inevitably remains a risk of inaccurate RR-interval detection and this is a known limitation of free-living wearable sensor data. Long-term HRV values here may not perfectly reflect the underlying autonomic function but represent the best practical approximation in real-world conditions.

6.4.3 Signal Quality

Short-term HRV was derived from clean sleep segments identified by the SQI, but posture, environment, and sleep disturbances remain uncontrolled in free-living settings. Long-term HRV, while more robust, still depends on sustained signal quality, which varied across participants. This suggests that the immediate goals of signal quality assessment, whether beat detection, HR estimation, or HRV extraction, need to be more explicitly defined. In this study, HR-derived indices proved more robust and clinically meaningful, indicating that refining HR extraction pipelines may represent the most effective next step for improving free-living HRV analyses.

In conclusion, this study provides preliminary support for integrating HRV measures into wearable sensor-based VO_2max prediction models. While the effect of HRV measures was modest and not significant, the findings suggest a link to CRF. Notably, long-term HRV measures demonstrated stronger predictive value and offer a practical advantage for remote monitoring as they can be calculated from HR signals. Importantly, these findings emphasise that high-quality HR signals are key to extracting clinically useful measures.

Chapter 7

Development of a Signal Quality Index (SQI) Tool

7.1 Introduction

In the preceding chapters, several stages of the analytical pipeline were combined to process 72 hours ECG signals alongside accelerometer data and predict VO_2max in the REMOTES cohort. The first stage of this process, presented in Chapter 4, evaluate multiple SQIs and selected the most appropriate. While this improved the reliability of extracted signals, Chapter 4 also revealed substantial disagreement between SQIs and the variable clinician agreement with the SQI labels, depending on the task. Further, even after applying the SQI, the inflated magnitude of SDNN24 using raw NN-intervals suggested that noisy or unstable data were still being included in analysis (see section 6.4.1).

Chapters 5 and 6 demonstrated that HR-based features, such as resting HR, maximal HR, and cadence-HR ratios, were useful contributors to VO_2max prediction. The addition of HRV measures provided modest, though statistically non-significant, improvements in model performance, with long-term HRV derived from HR signals emerging as the most informative. Together, this suggests that HR and the features derived from it represent a key predictive component, but that their utility may still be limited by inaccuracies in HR estimation or residual noise.

This provides a rationale to revisit the quality assessment stage of the ECG processing pipeline. Improving how signal quality is evaluated for HR estimation could reduced the propagation of artifacts into downstream feature extraction. By investigating how an SQI could be tailored to this task, it may be possible to enhance the accuracy of HR-derived features and as a results, the predictive power of wearable-derived VO_2max models.

7.1.1 Background

A wide range of SQIs have been proposed, from simple rule-based thresholds to machine learning classifiers [268]. DL-based vision classifiers have also shown some success, for example SQI4 implemented in Chapter 4. One point that these models all have in common is that they typically label segments as “Acceptable” or “Unacceptable,” developed using datasets annotated by human experts. As highlighted by Keutche et al. [269], this approach has limitations: expert-annotated datasets often contain little or no arrhythmic data, and more importantly, the quality thresholds are set by humans. Generally, wearable ECGs collect a large volume of data that requires automated analysis tools. These quality labels therefore may not align with the noise threshold that automated beat detection algorithms can overcome, but rather the noise threshold that an expert annotator deems acceptable. Where a clinician may reject a segment due to noise, a robust QRS detector might still extract reliable inter-beat intervals.

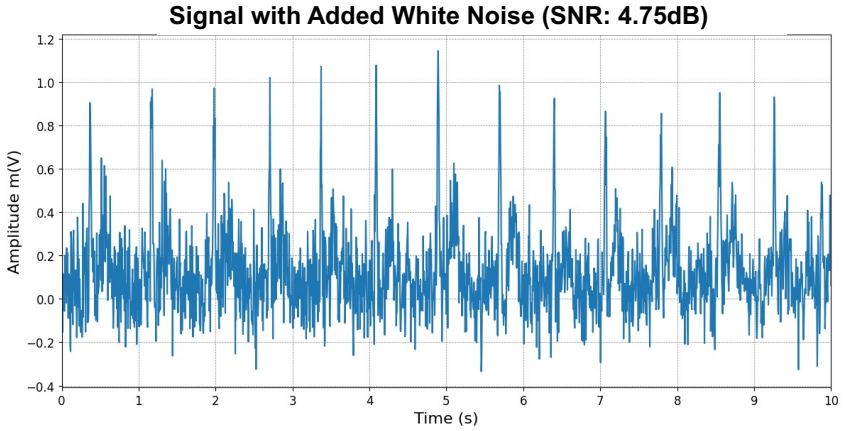


Figure 7.1: 10 second ECG plot taken from [3]. The shows a synthetic ECG signal with added white that was given to a cardiologist for assessment.

Further, the labels that are set by expert annotators in these tasks are generally binary (Acceptable vs Unacceptable) and not specific to any one task [216]. For example, the signal in Figure 7.1 was used in Chapter 4 and when reviewed by a cardiologist, it was labelled as having adequate quality for HR extraction, but not for full clinical assessment. A single label in this case would either allow misleading data into a task, or exclude signals with adequate quality from further processing. By generalising quality assessment across all types of ECG processing and feature extraction, there is a risk that this could undermine clinicians trust in wearable monitoring systems.

This highlights two important considerations. First, the quality labels used to train SQIs themselves should ideally reflect the needs and limitations of the processing algorithms being used, rather than human interpretation alone. Secondly, the level of signal quality required is highly task-dependent [3]. For example, accurate HR estimation primarily depends on QRS detectability, whereas arrhythmia or other cardiac abnormality

detection requires preservation of fine morphological detail. In other words, signal quality is best defined relative to both the processing method and the analytical task.

In this chapter, we propose an SQI that is tailored to specific signal processing and analytical goals, rather than a one-size-fits-all approach. Building on the challenges identified in Section 6.4, we develop and evaluate an SQI designed for HR estimation from free-living wearable ECG using automated algorithms, before applying it to the REMOTES dataset to assess its impact on feature extraction and predictive modelling. While the immediate focus is HR estimation, the framework is generalisable to other analytical tasks where appropriate training labels are available.

The aims of this chapter are therefore twofold:

- **To develop and validate an SQI** capable of identifying ECG segments of sufficient quality for accurate HR extraction using synthetic and open-access datasets. This process is presented in Part 1 below.
- Once the SQI has been developed, it is to be **applied to the REMOTES dataset** and its impact evaluated on downstream feature extraction and VO_2max prediction models. This process is presented in Part 2.

7.2 Part 1: Development of the HR-Specific SQI

7.2.1 Methods

This chapter describes the development and evaluation of a signal quality index (SQI) for wearable ECG. The workflow consisted of three main stages: (i) defining the labelling process of ECG segments, (ii) dataset preparation, (iii) model development and architecture. We begin with the labelling process, which is presented before the datasets as it defines the learning targets and underpins subsequent dataset selection.

Labelling Process

As outlined in the aims, the labels used to train an SQI should be specific to the task under investigation. Since the goal here was to assess whether heart rate (HR) can be reliably extracted from wearable ECG, labels needed to directly indicate whether automated HR estimation was accurate. To achieve this, we applied a standard ECG processing pipeline using the NeuroKit2 beat detector, and compared its outputs to ground-truth (GT) beat annotations. By comparing automated and GT-derived HR values, each segment was labelled as either *Acceptable* (sufficient quality for HR extraction) or *Unacceptable* (insufficient quality). The labelling procedure is summarised in Figure 7.2.

1. **Filtering:** Raw ECG segments were bandpass filtered (0.5–150 Hz, 5th-order Butterworth) to remove baseline wander and high-frequency noise, consistent with established preprocessing standards in modern ECG devices [270].

2. **Beat detection:** R-peaks were detected using the NeuroKit2 `ecg_findpeaks` algorithm.

3. **Beat matching:** Detected beats were matched against GT annotations. If fewer than 50% of detected beats fell within 50 ms of GT beats, the segment was labelled Unacceptable. This step prevents mislabelling segments as Acceptable when HR values coincided by chance, despite poor underlying beat detection.

4. **HR comparison:** HR was computed from both GT and detected beats. If automated HR was within 10% of GT-derived HR, the segment was labelled Acceptable; otherwise it was labelled Unacceptable. The 10% threshold was selected based on accepted accuracy standards for HR monitors [271].

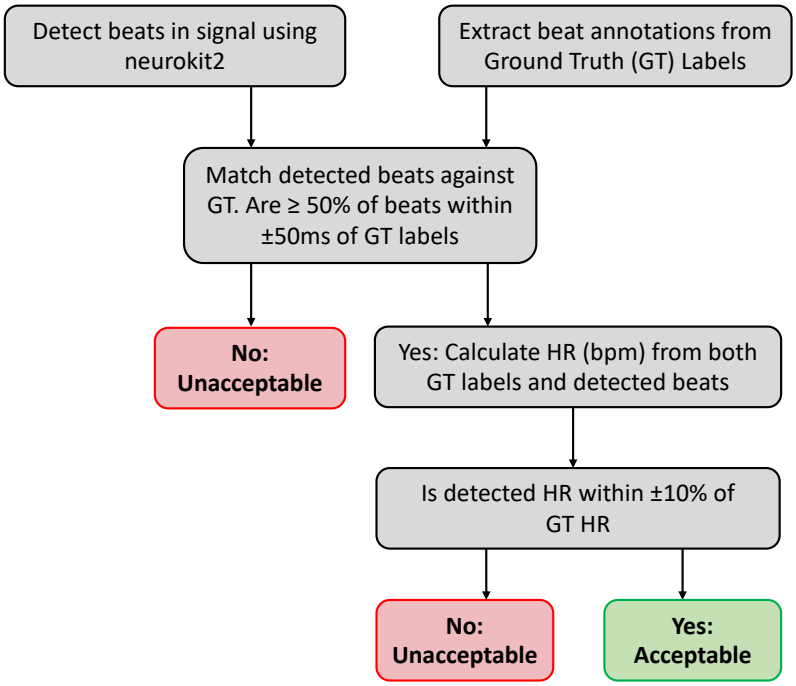


Figure 7.2: Workflow for assigning *Clean* or *Noisy* labels to ECG segments, based on comparison of NeuroKit2 beat detection against ground-truth annotations from open-source ECG datasets.

Datasets

We used a combination of synthetic, semi-synthetic, and real-world ECG datasets, restricted to single-lead recordings. As in Chapter 4.2.1, the primary requirement for inclusion was the availability of accurately annotated beat locations and the presence

of noise variation. Alongside real world-data, we also use the same synthetic beat generator as described in section 4.2.1. A particular advantage of synthetic ECG datasets is that the true beat locations are precisely defined, allowing exact ground-truth comparisons. This enables large-scale generation of labelled examples across a wide spectrum of noise conditions, substantially increasing the training volume available.

Adding in semi-synthetic (real world signals with artificially added noise) and real-world datasets then provides the variability and artifacts encountered in clinical practice, ensuring the model can generalise to real-world conditions. All datasets were partitioned into non-overlapping 10-second windows to ensure consistency with the segmentation approach applied to other SQIs in section 4.2.3. This allowed uniform labelling across datasets and ensured comparability between the synthetic, semi-synthetic, and real-world datasets outlined below:

- **Synthetic Data:** Synthetic ECG signals were generated using the same open-source simulator described in Chapter 4.2.1, capable of adding noise such as HR variability, white noise, power-line interference, and motion artefacts [222]. Signal were generated using random parameter but with the noise functions increased to encourage the presence of noise. Synthetic signals were segmented into 10-second windows for consistency with the other datasets.

- **MIT-BIH Noise Stress Test (NST):** A semi-synthetic dataset consisting of 12 half-hour ECG recordings created by adding calibrated noise (baseline wander, muscle artefact, and electrode motion) to clean ECGs from the MIT-BIH Arrhythmia Database [272]. Noise was added in alternating two-minute intervals, interspersed with clean segments, at varying signal-to-noise ratios (from +24 dB to –6 dB). Because noise-free intervals were not relevant to the labelling task, only the noisy segments were extracted and split into 10-second windows for analysis. This ensured inclusion of a wide spectrum of realistic artefacts while retaining access to ground-truth beat annotations from the underlying clean signals.

- **PhysioNet 2014 Challenge Augmented Dataset:** This dataset includes 100 half-hour ECG recordings originally released for the 2014 PhysioNet Challenge Test set, containing signals from multiple participants in free-living conditions [273]. All signals were divided into 10-second windows with beat annotations used as ground truth. This dataset ensured the SQI was exposed to naturally occurring artefacts in long-term wearable data.

- **TELE ECG Dataset (external validation):** The TELE dataset comprises 250 telehealth ECG recordings collected with dry metal electrodes [274]. While the dataset contained continuous ECG recordings, some files did not include beat annotations in the initial seconds. Therefore, only segments with valid annotated beats were retained, leading to varying numbers of usable 10-second windows

per recording. This smaller dataset was used exclusively as an external validation set to assess generalisability of the SQI to unseen real-world data.

Data Augmentation

To increase the diversity of training examples and improve model robustness to real-world noise, we applied data augmentation to the semi-synthetic (MIT-BIH NST) and real-world (PhysioNet 2014 Challenge) datasets. Each augmented segment preserved its original beat annotations while introducing realistic distortions. Augmentation is particularly valuable in free-living wearable ECG applications, where variability in electrode contact, positioning, posture, and environment can generate signal artifacts not wholly captured in curated datasets. By simulating these perturbations, it is hoped the model is better equipped to generalise to unseen data.

The augmentation pipeline randomly applied up to three transformations per segment, chosen from the set listed in Table 7.1. These augmentation techniques were identified from a review completed by Rahman and colleagues (2023) [275].

Table 7.1: Overview of augmentation transformations applied to ECG signals. Each transformation was applied randomly, with up to three combined per 10-second segment with.

Transformation	Description
Gaussian noise	Adds normally distributed noise scaled to the signal standard deviation.
Amplitude scaling	Multiplies the signal by a random factor (0.6–1.4).
Baseline wander	Superimposes a low-frequency sinusoidal drift (0.05–0.2 Hz).
Amplitude inversion	Flips the signal polarity.
Sine wave artefact	Adds a higher-frequency sinusoidal component (e.g. 5 Hz).
Baseline shift	Adds a constant offset drawn from a scaled random distribution.
Low-pass filter	Filters the signal at a randomly chosen cut-off (e.g. 20 Hz).
High-pass filter	Filters the signal at a randomly chosen cut-off (e.g. 0.5 Hz).

Model Development

Architecture Once segments were labelled as *Acceptable* or *Unacceptable* for accurate HR extraction (Section 7.2.1), we trained a deep learning classifier to discriminate between these classes (Figure 7.3). Deep learning approaches, and in particular convolutional neural networks (CNNs), have shown high performance in ECG classification

problems, as they are able to learn local morphological patterns (e.g., QRS complexes) directly from raw signals without the need for handcrafted features [276].

In principle, the capacity of a deep-neural-network to learn increases as the number of layers in a model increases. However, one problem with having a deep model is that gradients can diminish to near-zero ('vanishing gradients'), hindering convergence towards a useful point. Residual networks (ResNets) extend CNNs by proposing skip connections which aim to overcome the 'vanishing gradient' problem and stabilise optimisation in deep architectures [277]. These have been shown to be effective for ECG classification tasks [278, 279]. Given the sequential nature of ECG signals, we implemented a custom 1D ResNet architecture tailored to signals rather than images. Unlike the standard 2D ResNet-18 commonly used in computer vision, this architecture is lighter consisting of an initial convolutional layer followed by three residual blocks and a global pooling layer. This was selected to reduce the risk of overfitting, which is important given the modest size of real-world ECG datasets available for fine-tuning compared to the synthetic data.

Model Training To maximise generalisability while retaining practical feasibility, we adopted a two-stage strategy, shown in Figure 7.3: (i) *pretraining* on large synthetic and augmented datasets (where beat locations and class labels are known exactly) to learn robust ECG morphology under controlled noise conditions, followed by (ii) *fine-tuning* on semi-synthetic and real-world datasets to adapt to real noise characteristics and device idiosyncrasies. This approach leverages the advantages of deep learning in capturing subtle waveform characteristics, while controlling overfitting by transferring knowledge from synthetic to real data.

The network comprises:

- Initial block: 1D convolution (kernel 5) $\rightarrow$ BatchNorm $\rightarrow$ ReLU $\rightarrow$ MaxPool.
- Residual blocks: three stacked residual units with channel sizes 32, 64, and 128; each unit contains Conv–BN–ReLU layers and an identity/1×1 projection shortcut when downsampling. A squeeze–and–excitation (SE) module refines channel attention within each block.
- Classifier head: global average pooling $\rightarrow$ dense (64) with ReLU and dropout (0.5) $\rightarrow$ sigmoid output (binary).

Pre-training Before input to the model, all signals were resampled to 500Hz and normalised using min-max scaling, and filtered using the same filter as in the labelling process. The model was pre-trained for 80 epochs on synthetic segments to learn noise-robust, beat-centric features. We used the Adam optimiser (learning rate 1×10^{-3},

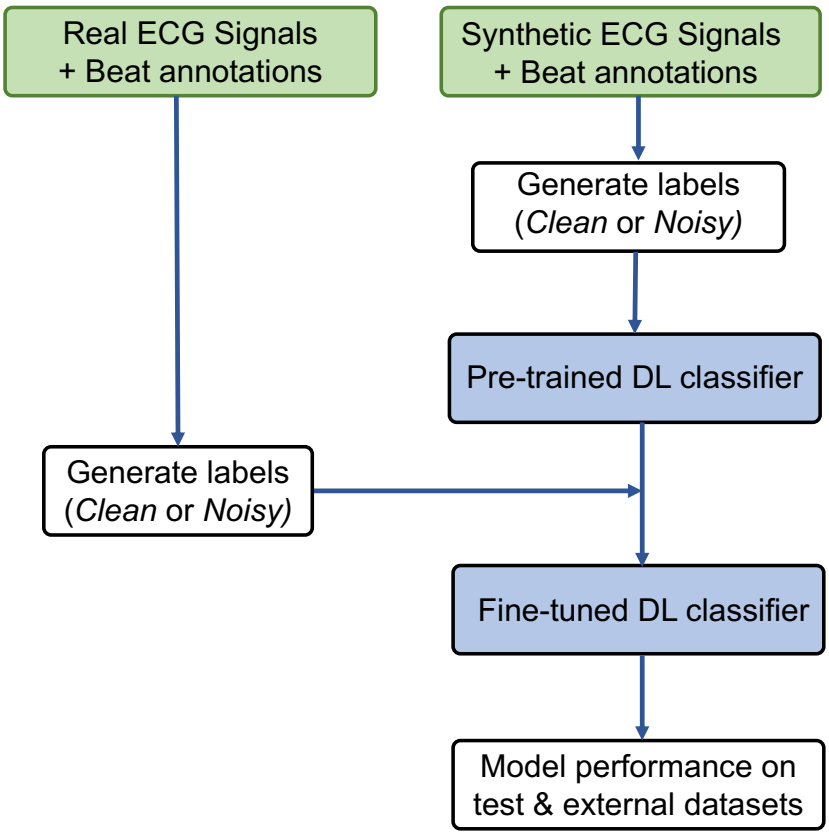


Figure 7.3: End-to-end framework for a task-specific SQI targeting accurate HR extraction. Data inputs (green) include synthetic (for pretraining) and real/semi-synthetic (for fine-tuning). Model development (blue) comprises a 1D ResNet classifier trained with segment-level labels derived from the downstream HR pipeline.

weight decay 1×10^{-4}), binary cross-entropy with logits, and a ReduceLROnPlateau scheduler on validation loss.

Fine-tuning and validation. For fine-tuning on semi-synthetic (MIT-BIH NST, augmented) and real-world (PhysioNet 2014, augmented) data, we split segments 60/20/20 into train/validation/test with stratification by label. Early layers were frozen for the first epochs to retain generic features, then unfrozen for full-network adaptation. We kept the same optimiser and scheduler. The final decision threshold was selected by sweeping $t \in [0.1, 0.9]$ on the (real-only) validation set and choosing the value that maximised F1 score, then fixing that threshold for all test evaluations (including external TELE ECG).

Model Evaluation

Model evaluation focused on classification performance in distinguishing *Acceptable* and *Unacceptable* ECG segments for heart rate estimation. Because this is a binary classification task with an imbalanced class distribution, metrics beyond overall accuracy were used to capture model behaviour. We report accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUROC).

Accuracy reflects the overall proportion of correctly classified segments. Precision ($\frac{\mathrm{TP}}{\mathrm{TP+FP}}$) quantifies the proportion of segments predicted as *Acceptable* that were truly usable, whereas recall ($\frac{\mathrm{TP}}{\mathrm{TP+FN}}$) measures the ability of the model to identify all usable segments. The F1 score, the 'harmonic mean of precision and recall', provides a single balanced measure of discrimination performance and was used to select the decision threshold that maximised performance on the validation set. Finally, AUROC summarises performance across thresholds, indicating the model's overall ability to separate the two classes independent of any fixed cut-off.

This combination of metrics provides a solid assessment of model utility in practice: high recall ensures that usable ECG is rarely discarded, while high precision prevents noisy segments from being misclassified and contaminating downstream HR extraction.

7.2.2 Results

Dataset Overview

In Table 7.2 we report the number of 10-second ECG segments included from each dataset, along with their sampling frequencies and the proportion of segments labelled as Acceptable. Two large randomly generated synthetic datasets were used for pretraining with different sampling frequencies to match those of the real-world datasets.

The PhysioNet 2014 and MIT-BIH NST datasets were used for fine-tuning and evaluation, while the TELE ECG dataset was held out entirely for external validation. As expected, the proportion of Acceptable signals was highest in the PhysioNet 2014 dataset, reflecting its semi-controlled recording environment, and lowest in the MIT-BIH NST dataset due to deliberately added noise.

Table 7.2: Dataset composition, sampling frequency, and proportion of Acceptable signals

Dataset	Hz	Number of Signals	% Clean
Synthetic	500	40,000	37.4%
Synthetic 2	360	32,560	41.8%
MIT-BIH NST	360	804	39.2%
PhysioNet 2014	360	4,452	64.7%
TELE ECG	500	416	43.8%

Performance on Internal Test Sets

After training and fine-tuning, the model was evaluated on held-out partitions from the real-world and semi-synthetic datasets. A decision threshold of 0.32 was selected based on optimisation of the F1 score on the validation set.

Across all internal test data, the model achieved a mean accuracy of 0.89 and an F1 score of 0.89 (Table 7.3). Performance remained well balanced between precision (0.89) and recall (0.89), indicating reliable discrimination between *Clean* and *Noisy* ECG segments. When stratified by dataset, the model achieved an F1 score of 0.90 for the PhysioNet 2014 subset and 0.88 for the MIT-BIH NST subset, reflecting slightly higher performance on the semi-controlled PhysioNet data but still robust generalisation to noisier conditions.

Table 7.3: Model performance on internal test sets.

Dataset	Acc.	Prec.	Rec.	F1
PhysioNet 2014 (PNAUG)	0.898	0.903	0.901	0.899
MIT-BIH NST	0.875	0.877	0.882	0.876

Performance on External Dataset

The model generalised well to the unseen TELE dataset, achieving an accuracy of 0.81, F1 score of 0.80, and AUROC of 0.86. Precision (0.81) was lower than recall (0.90), suggesting that the model successfully identifies most Acceptable segments but occasionally also classifying Unacceptable segments as acceptable. This behaviour prioritises retaining true Acceptable signals for downstream heart rate estimation, at the cost of a small increase in false positives.

Incorrect Classification

To understand model behaviour, we examined examples of incorrectly classified signals. Figure 7.4 shows two examples. Some signals incorrectly labelled as *Acceptable* had large R-peaks, which may have dominated the model's decision despite high-frequency noise that the Neurokit detector struggles with. In contrast, signals incorrectly labelled as *Unacceptable* often had small R-peaks, or a higher number of R-peaks which inherently allows a higher volume of peak detection error, but still to fall within the 10% labelling error. This suggests the model relies heavily on R-peak prominence, potentially overlooking components such as higher-frequency noise.

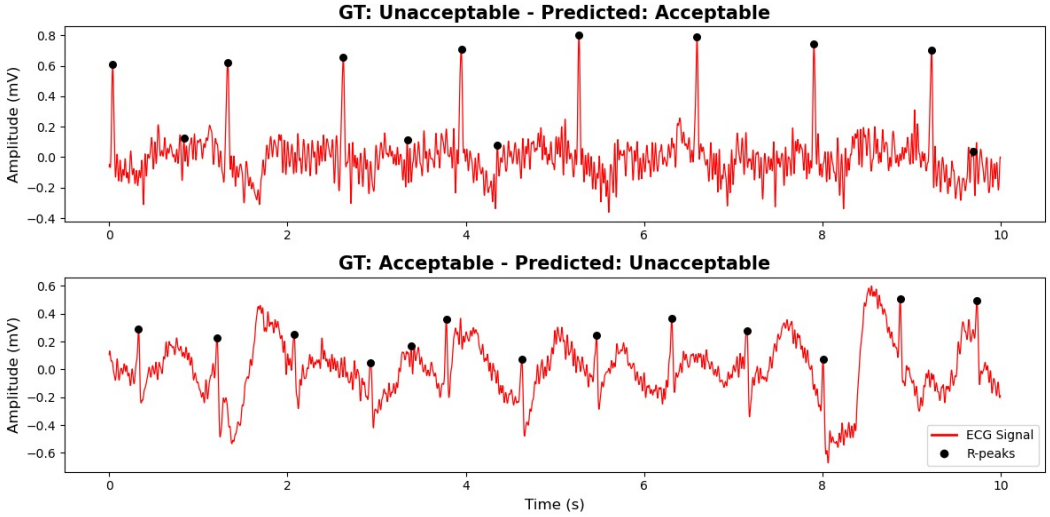


Figure 7.4: Example of incorrectly classified signals from the external dataset using the Ground Truth labels (GT). Top: an *'Unacceptable'* signal, with poor R-peak detection, misclassified as *'Acceptable'*. Bottom: *'Acceptable'* signal, with mostly correctly detected R-peaks, incorrectly classified as *'Unacceptable'*.

7.2.3 Discussion

This project presents a task-specific framework to ECG signal quality assessment, aimed specifically at heart rate estimation using a defined beat detector. By aligning signal quality assessment with analysis goals, we move away from general-purpose SQIs towards a targeted, reproducible approach.

Despite using only a modest amount of real-world data for fine-tuning, our model achieved good performance on both internal and external test sets. Results show that pre-training a deep learning model on synthetic ECG signals simulating realistic noise provides a robust foundation for learning generalisable features. Given the difficulty of obtaining beat locations in noisy real-world data, synthetic signals were especially useful for HR estimation. Performance on edge cases suggests synthetic noise may not fully capture real signal variability, and augmenting real-world data may be key

to bridging this gap. Encouragingly, the model generalised well to an unseen external dataset collected using different sensors (dry electrodes), indicating robustness to different sensor modalities.

A key benefit of this approach is its potential for real-world application in accurate ECG processing. In long-term ECG monitoring, manual inspection of signal quality is impractical and general purpose SQIs may not reliably segment signals for HR estimation. This SQI offers an automated way to flag periods unsuitable for HR estimation that could be deployed within clinical processing pipelines. In doing so, it may increase clinicians' trust in HR derived from wearable ECG's and support the safe integration of this data into clinical workflows.

More broadly, this work presents a generalisable pipeline for task-specific SQI development, linking signal quality assessment directly to the goals of the analysis. Previous research has developed processing pipelines tailored to HR estimation from ECGs, fusing multiple SQIs and HR from multiple ECG leads [280]. While also presenting a generalisable framework, our approach differs in its task-specific design, providing a modular framework specific to the signal processing pipeline itself. For example, it can be used with different beat detectors, stricter thresholds (e.g., $<$5% HR deviation), or extended to other physiological measurements such as respiratory rate or heart rate variability, and applied across sensor modalities.

An important consideration when implementing the SQI is that, because heart rate is estimated over fixed-length windows, the tolerance to the beat detection error is frequency dependent. At higher heart rates, a larger absolute number of beat detection errors may still fall within the acceptance threshold, whereas at lower heart rates, even a small number of missed beats can lead to labelling a segment as unacceptable. This should be considered when interpreting performance across different heart rate ranges. In data with elevated HR's (e.g. exercising), then a stricter acceptance threshold may be more appropriate.

A limitation of our work is that this considers a very simple pipeline where raw ECG is input directly into a beat detector to determine HR, with only limited pre-processing. The effects of this can be seen in the examples of incorrect classification (see figure 7.4, where there are clear R-peaks but the beat-detector may struggle with the presence of high-frequency noise that could be removed with further processing). Alternative pipelines that included additional pre-processing can be used in our labelling process, and should be tested in future to demonstrate utility in more realistic scenarios.

To conclude, this work highlights the importance of task- and pipeline-specific signal quality labelling and offers a reproducible approach for future SQI development. Future work will implement a robust pre-processing strategy before benchmarking this SQI against existing general-purpose SQIs to determine comparative performance in labelling signals for accurate HR extraction.

7.3 Part 2: Application to the REMOTES Dataset

7.3.1 Methods

To evaluate the practical utility of the task-specific SQI, it was applied to the REMOTES dataset introduced in Chapter 3. The SQI was integrated into the same preprocessing pipeline described in Section 4.3. Signals were segmented into non-overlapping 10-second windows; each window was passed through the trained SQI model. If labelled as *Acceptable*, HR was extracted using the NeuroKit2 beat detector; if *Unacceptable*, the window was assigned a value of zero, indicating missing HR due to poor quality. This produced an HR time series consistent with those used in Chapters 4 and 6, but with noise handling guided by the HR-specific SQI rather than rule-based thresholds.

Impact on Data Volume. The proportion of usable ECG data was quantified across the cohort and for individual participants. These values were compared directly with those obtained using the original SQI from Chapter 4. This allowed evaluation of whether the task-specific SQI increased or decreased usable data volume, and whether changes were uniform across participants.

Impact on Feature Extraction and Analysis. To evaluate downstream effects, the feature extraction and modelling pipeline described in Chapter 6 was repeated. Specifically:

- We recalculated all HRV and HR-derived features using the SQI-labelled HR time series.
- Correlations between these features and measured VO_2max were re-assessed and compared with results obtained in Chapter 6.
- The LASSO regression model with HRV features was retrained using the SQI-labelled data, and its performance was compared with the original baseline model (without HRV).

This design allowed us to assess not only changes in data availability but also their downstream effect on feature–outcome associations and model performance. To ensure consistency, statistical comparison of fold-level metrics was repeated using paired *t*-tests, as in Chapter 6.

7.3.2 Results

Change to the Volume of Acceptable Data

We first compared the proportion of usable Acceptable ECG data across participants when applying the ground-truth (GT) SQI from Chapter 4 and the HR-specific SQI. Table 7.4 summarises the average, median, and variability in usable data proportions for both approaches. On average, the difference between GT and ML methods was small (mean difference = –0.3%, SD = 16%), indicating that the HR-specific SQI neither systematically increased nor reduced usable data volume at the cohort level. However, individual-level comparisons revealed large variability, with some participants showing up to 40–60% more segments labelled as Acceptable by the HR-specific SQI, and others showing the opposite pattern. These findings suggest that while the HR-specific SQI does not alter the overall volume of usable data at the cohort level, its variable impact across individuals highlights the potential for signal quality assessment to influence downstream analyses.

Table 7.4: Percentage of Acceptable data across participants under ground-truth (GT) SQI and the HR-specific SQI. Values are reported as mean, standard deviation (SD), and median.

	Mean (%)	SD (%)	Median (%)
SQI	74.50	19.60	79.30
HR-SQI SQI	74.30	20.10	81.00

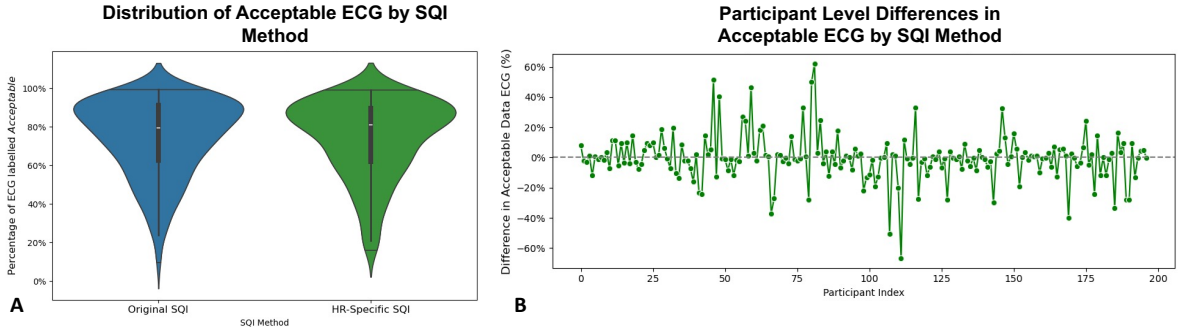


Figure 7.5: Comparison of usable ECG data between the original SQI and the HR-specific SQI. **(A)** Violin plots showing the distribution of the percentage of 10-second segments classified as *Acceptable* across all participants. The overall distributions are highly similar between methods, indicating no systematic shift at the cohort level. **(B)** Participant-level differences in usable data percentage (HR-specific SQI minus original SQI). While the mean difference was close to zero, some individuals exhibited wide discrepancies, with differences of up to 40–60% in either direction.

Using the HR-specific SQI, a greater proportion of ECG segments were retained as Acceptable, resulting in 167 participants meeting the inclusion criteria for HRV feature

extraction compared with 163 under the original SQI. Specifically, less participants were excluded for having ¡16 hours of valid HR data in the first day (17 excluded when using original SQI; 12 when using the HR-specific SQI), although one extra participant was excluded for their prevalence of ectopic beats.

Feature Correlations with the New SQI

Applying the HR-specific SQI produced broadly similar correlation patterns to those reported with the original SQI, with some notable differences (Table 7.5). Daily steps remained the strongest correlate of VO_2max ($r = 0.54$, $p < 0.01$), consistent with established predictors. However, the biggest change was observed for maximum HR, which showed a stronger correlation with VO_2max ($r = 0.40$, $p < 0.01$) compared to the original SQI ($r = 0.31$). This suggests that improved handling of noisy segments may have enhanced the stability of this feature, making it a more reliable indicator of cardiac capacity.

Table 7.5: Correlations of HR-related features with VO_2max under the original SQI and the HR-specific SQI.

Feature	**Original SQI (r)**	**HR-specific SQI (r)**
Max HR	0.31*	0.40*
Resting HR	-0.16*	-0.16*
Min HR	-0.17*	-0.12
25th Percentile (P25)	0.15	0.16*
50th Percentile (P50, Median)	0.21*	0.26*
75th Percentile (P75)	0.28*	0.34*
95th Percentile (P95)	0.33*	0.39*

* $p < 0.05$

Almost all HR-related features also showed improvements. The higher-order cadence-to-HR ratios (e.g., the 95th percentile) increased in strength ($r = 0.39$ vs. $r = 0.33$), and the 75th percentile similarly rose ($r = 0.34$ vs. $r = 0.28$), indicating that the HR-specific SQI may better capture relationships between higher-intensity ambulatory periods and VO_2max. Resting HR and minimum HR retained weak negative correlations, though both were slightly attenuated, suggesting no substantive impact of the new SQI on lower-bound HR measures. These shifts highlight that the most meaningful gains from the HR-specific SQI were within features relating to faster HR recordings (maximum HR and higher order cadence-to-HR ratios), and as a result these appear to have a stronger link to cardiorespiratory fitness.

Impact of the HR-specific SQI on HRV Model Performance

To first directly compare the impact of the SQI model on model performance, we compare the predictions between the HRV model using the original SQI, against that using

the HR-specific SQI.

Table 7.6: Performance of the HRV model using the original SQI compared with the HR-specific SQI. Values are mean (± SD) across cross-validation folds.

Metric	**Original SQI (HRV model)**	**HR-specific SQI (HRV model)**
Correlation	0.70 (0.08)	**0.73** (0.07)
R^2	0.48 (0.12)	**0.51** (0.08)
MAE (ml·kg^{-1}·min^{-1})	2.63 (0.34)	**2.62** (0.17)
RMSE (ml·kg^{-1}·min^{-1})	3.38 (0.53)	**3.35** (0.25)
APE (%)	15.55 (2.19)	**15.37** (1.61)

Table 7.6 shows the performance of the HRV model when applied with the original SQI and with the new HR-specific SQI. Improvements with the new SQI were modest in absolute terms but consistent across all metrics, with slightly higher correlation ($r = 0.73$ vs $r = 0.70$) and explained variance ($R^2 = 0.51$ vs $R^2 = 0.47$), alongside small reductions in error (MAE, RMSE, and APE), with smaller standard deviations.

While these changes are small, they indicate that refining the quality of ECG input makes the HRV features slightly more informative for VO_2max prediction. Notably, as shown in the appendix (see Table E.1), the relative benefit of including HRV (vs non-HRV) was also larger when using the HR-specific SQI across metrics, compared the original SQI. Together, these results suggest that the HR-specific SQI improves the stability of ECG-derived features and allows HRV to contribute more consistently to predictive performance.

Feature Contributions (SHAP Analysis)

Figure 7.6 compares SHAP feature importance values for the HRV model using the original SQI (left) and the HR-specific SQI (right). The most notable change was a substantial increase in the contribution of maximum HR (highlighted with an asterisk in the figure). In contrast, the contribution of SDANN$_{HR}$24 decreased slightly. This pattern may reflect overlap in the information captured by these features: as maximum HR gained influence, the model relied somewhat less on SDANN$_{HR}$24. Such shifts are consistent with multicollinearity between ECG-derived features and highlight how refinements in signal quality can change the balance of predictors within the model.

7.3.3 Final comparison of Baseline vs Optimised Model

To summarise the cumulative optimisation steps presented in this thesis, the baseline model (standard features with the original SQI) was compared against the optimised model incorporating HRV features and the HR-specific SQI. To do this only participants who met the inclusion time of over 24 hours of valid HR, across of both different SQI tools were included, so as to perform a complete comparison across participants with

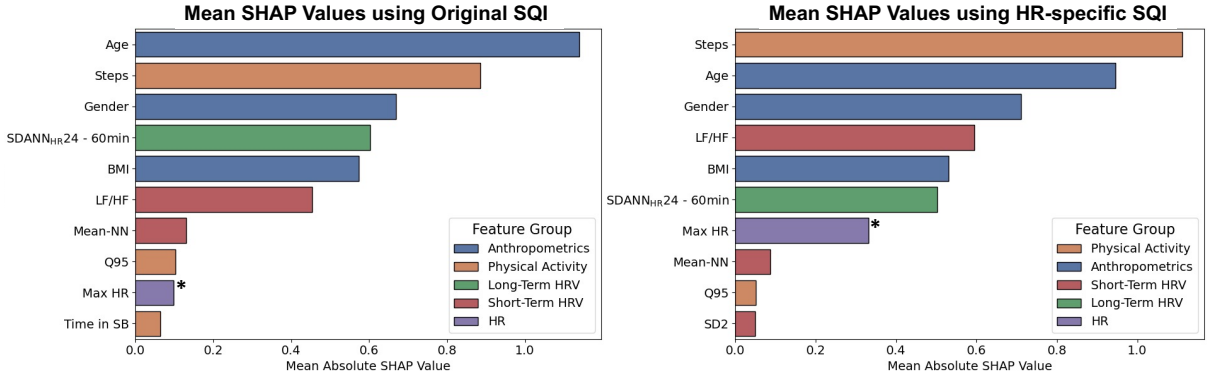


Figure 7.6: Comparison of SHAP feature importance between the HRV model with the original SQI (left) and with the HR-specific SQI (right). The increased contribution of maximum HR is highlighted with an asterisk.

results from both prediction models. Performance was assessed across all five cross-validation folds, and paired statistical tests were conducted on fold-level performance metrics to determine whether differences between models were significant.

Explained variance (R^2) improved from 0.44 in the baseline model to 0.51 in the optimised model, with corresponding increases in correlation (0.68 $\rightarrow$ 0.73) and reductions in error metrics (MAE, RMSE, APE). Although absolute gains were modest, they were consistent across folds. Figure 7.7B illustrates how the optimised model redistributed predicted VO_2max values, producing a closer alignment with the observed distribution compared to the baseline.

Table 7.7: Paired *t*-tests comparing baseline and optimised (HRV + HR-specific SQI) models across cross-validation folds. Asterisks denote statistical significance.

Metric	**Baseline**	**Optimised**	**Mean diff.**	***t*-test (p)**
R^2	0.44 (0.09)	0.51 (0.09)	+0.07 (0.04)	0.025*
Correlation	0.68 (0.08)	0.73 (0.07)	+0.05 (0.03)	0.025*
MAE (ml·kg^{-1}·min^{-1})	2.79 (0.19)	2.62 (0.19)	-0.16 (0.14)	0.055
RMSE (ml·kg^{-1}·min^{-1})	3.58 (0.33)	3.35 (0.28)	-0.23 (0.16)	0.032*
APE (%)	16.34 (1.85)	15.37 (1.80)	-0.97 (0.83)	0.058

*$p < 0.05$

Paired t-tests of the differences in metrics (Table 7.7) confirmed that the optimised model significantly outperformed the baseline across in several metrics (R^2, correlation, RMSE). These results indicate that, while incremental, the combined refinement of physiological features (via HRV) and signal quality (via a task-specific SQI) delivers measurable improvements in VO_2max prediction accuracy.

These results demonstrate that even modest refinements in feature space and signal quality assessment can deliver measurable improvements in VO_2max prediction, underscoring the value of task-specific optimisation in free-living wearable data.

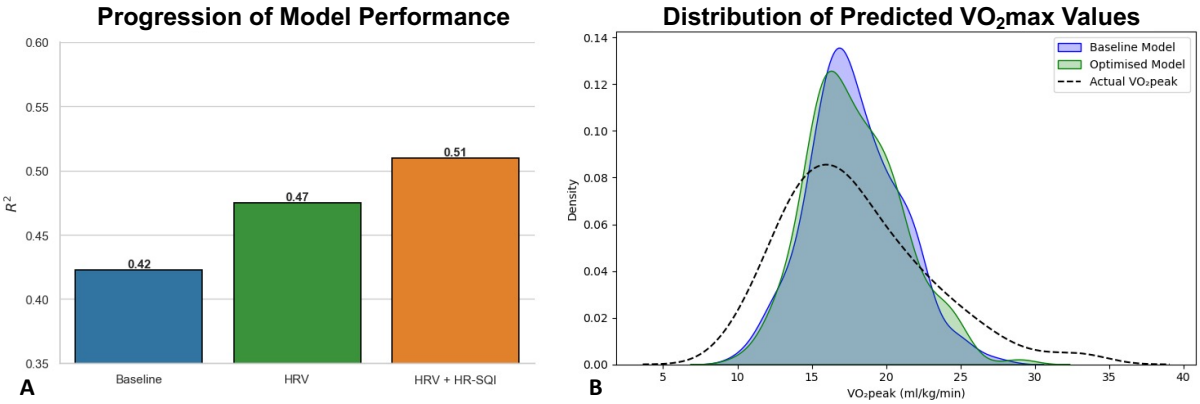


Figure 7.7: Final model comparison. (A) R^2 improvements from the baseline model in Chapter 5 to the optimised HRV + HR-specific SQI model. (B) Redistribution of predictions showing closer alignment with observed VO_2max under the optimised model.

Chapter 8

Conclusion & Future Work

8.1 Summary of Findings

Cardiorespiratory fitness (CRF), expressed as maximal oxygen uptake (VO_2max), is a critical predictor of perioperative risk and long-term health outcomes. Despite its clinical value, the gold-standard test, CPET, is resource-intensive, requires specialist oversight, and is not feasible for all patients. This motivated the search for scalable, passive alternatives based on wearable sensors. While promising results have been reported in healthy populations, previous work has relied on commercial devices with limited transparency around preprocessing and has generally avoided advanced signal-derived features such as HRV. This thesis has addressed these gaps by (i) systematically evaluating existing SQIs, (ii) testing the added value of HRV features for VO_2max prediction in a preoperative surgical cohort, and (iii) developing a task-specific SQI tailored to HR estimation to improve VO_2max predictions.

The thesis followed a stepwise approach. First, existing SQIs were evaluated using synthetic and annotated ECG data, revealing inconsistencies in their outputs and agreement with expert labelling. The best performing SQI was selected and applied to the REMOTES dataset to process the ECG data. Second, using this processed signal HRV features were extracted from the REMOTES dataset and integrated into regression models predicting VO_2max. While this improved performance modestly across all metrics, differences between the HRV and non-HRV models when paired with the original rule-based SQI were not statistically significant. Third, this motivated the development of a new SQI, designed specifically to support HR estimation by aligning labels with the output of an automated beat detector.

The final comparison demonstrated that explained variance (R^2) improved from 0.42 in the baseline model in Chapter 5 to 0.51 in the optimised model in Chapter 7, with corresponding increases in correlation (0.66 $\rightarrow$ 0.73) and reductions in RMSE. Paired *t*-tests between the differences in metrics confirmed that these improvements were significant for R^2, correlation, and RMSE. These results indicate that, while incremen-

tal, the methods for optimising the model presented in this thesis, first through feature enhancement (HRV) and then through a task-specific SQI, yielded meaningful improvements in prediction accuracy. These improvements also confirm the intended benefit of a task-specific SQI: by directly targeting HR estimation, the method enhanced the utility of specific HR-derived features. In particular, the increased stability and predictive contribution of maximum HR demonstrates that aligning SQI design with analytical goals can improve the reliability of key features.

It is important to acknowledge that absolute model performance remains lower than that reported in some studies, particularly those conducted in healthy cohorts or using exercise-derived predictor variables (often reporting $R > 0.8$) [233]. This likely reflects both the heterogeneity of the surgical population and the reliance on passive, free-living data rather than exercise tests which are more directly tied to maximal capacity. However, when compared to research using only free-living data in healthy populations, the results achieved here are more broadly comparable [232]. Within this context, the findings provide insights into the feasibility of wearable-based prediction in preoperative care.

8.2 Clinical Implications

When considering the clinical implications of these findings, it is important to acknowledge that direct VO_2max measurements have inherent variability, with measurement errors of ±5% widely accepted as clinically tolerable [281]. A meta-analysis reported an average standard error of measurement of 2.63 ml/kg/min for test-retest reliability, while other studies have observed VO_2max differences of 5–8% solely due to variations in data processing, supervision, or equipment calibration [282, 283]. Although the absolute prediction error in this model remains too high to directly replace CPET (APE: 15.55%), it should be considered within the context of underlying variability in VO_2max measurements. There may be potential for this model to support triaging and efficient allocation of testing resources.

A VO_2max value below 15 ml/kg/min is a recognised clinical threshold for identifying high-risk patients at preoperative assessment [52]. As wearable devices become more embedded across the surgical pathway, models like this could offer a scalable solution for prioritising CPET using data that is already passively collected by commercial wearables. By accounting for the prediction error in the model and integrating this into risk stratification tools, patients whose predicted VO_2max is well above 15 ml/kg/min could be triaged as low priority for CPET, while those close to or below the threshold could be prioritised for further evaluation. This stratified approach could optimise CPET allocation and streamline preoperative pathways using data from digital health tools to reduce the clinical burden of testing.

8.3 Limitations

Several limitations of the present work must be acknowledged. The data were collected exclusively from patients scheduled for major abdominal surgery, limiting direct generalisability to other clinical groups or to healthy populations. Gender imbalance (74% male) may also constrain broader applicability. The modest sample size and reliance on 5-fold cross-validation also leave scope for over-fitting in this cohort. External validation in an independent, balanced cohort would be required to further test the generalisability of this model.

Another limitation concerns the inclusion thresholds applied during preprocessing (e.g., minimum hours of data required for participant inclusion or for calculating long-term HRV). In several cases, limited prior evidence was available to select these thresholds, and thresholds were implemented based less on empirical research but rather clinical reasoning and feasibility. While these choices were necessary to ensure data quality and to proceed with the analysis, future work should investigate optimal thresholds systematically.

8.4 Future Work

Future research should aim to validate these findings in larger and more diverse cohorts, ideally with external datasets that balance gender and surgical subtypes. Incorporating clinical variables such as co-morbidities, medication use, or disease status alongside wearable features may reduce unexplained variance and improve prediction accuracy.

Further refinement of the SQI framework is also warranted. Future studies should test stricter tolerances (e.g. $<$5% HR deviation) and extend the approach to other signal modalities such as photoplethysmography (PPG), respiratory rate, or HRV-specific applications. Although the current HR-specific SQI is relatively lightweight compared to many deep learning frameworks and had short implementation times per ECG segment, its feasibility for real-time or on-device deployment remains uncertain and should be explored. For HRV development, future work should investigate whether the long-term features identified here (e.g. $SDANN_{HR}24$) provide value across different populations and contexts, and whether further refinements or approximations could increase their robustness in free-living data.

Beyond technical development, the integration of these models into perioperative workflows could be tested to evaluate whether wearable-derived predictions can meaningfully influence clinical decision-making. In particular, the use of such models to triage CPET allocation requires careful evaluation of risk thresholds to ensure that high-risk patients are not incorrectly de-prioritised, alongside a cost–benefit analyses

to quantify efficiency gains. In the longer term, as wearable devices are increasingly embedded into the perioperative care pathway using task-specific SQIs and optimised feature sets directly into wearable platforms could accelerate trust adoption and reliability in real-world monitoring devices.

8.5 Final Remarks

To conclude, this thesis demonstrates that both HRV feature integration and task-specific signal quality assessment can deliver measurable improvements in VO_2max prediction from wearable data in a preoperative population. Although absolute performance remains modest, the optimised model outperformed the baseline, underscoring the value of further investigating features such as HRV and refining preprocessing pipelines. Together, these findings provide a methodological foundation for future clinical research of wearable-based fitness assessment.

Bibliography

[1] Aron Syversen et al. “Wearable Sensors as a Preoperative Assessment Tool: A Review”. In: *Sensors* 24.2 (May 2024), p. 482. ISSN: 14248220. DOI: 10.3390/S24020482/S1.

[2] Aron Syversen et al. “Remote Prediction of Cardiorespiratory Fitness in a Preoperative Cohort: Exploring Short and Long-term Heart Rate Variability.” In: *Research Square, PrePrint* (Aug. 2025).

[3] Aron Syversen et al. “Assessment of ECG Signal Quality Index Algorithms Using Synthetic ECG Data”. In: (May 2024). DOI: 10.22489/CINC.2024.270.

[4] Aron Syversen et al. “Machine Learning for VO2max Predictions: A Comparison of Methods using Wearable Sensor Data.” In: *Proceedings of IEEE EMBC (In Press)* (June 2025).

[5] Aron Syversen et al. “A Framework for Task- Specific Signal Quality Assessment: A Case Study in Heart Rate Estimation.” In: *In Proceedings of 52nd International Computing in Cardiology Conference (In Press)*.

[6] Aron Syversen et al. ““How can we involve Patients?” - Students’ perspectives on embedding PPIE into a doctoral training centre for AI in medical diagnosis and care”. In: *Research Involvement and Engagement* 11.1 (Dec. 2025), p. 77. ISSN: 20567529. DOI: 10.1186/S40900-025-00750-Y.

[7] Alexios Dosis et al. “Estimating postoperative mortality in colorectal surgery- a systematic review of risk prediction models”. In: *International Journal of Colorectal Disease* 38.1 (May 2023), pp. 1–13. ISSN: 14321262. DOI: 10.1007/S00384-023-04455-0/FIGURES/4.

[8] Xiaoyu Wang et al. “Two-Stage Domain Adversarial Learning to Identify Chagas Disease from ECG and Patient Demographic Data”. In: *In Proceedings of 52nd International Computing in Cardiology Conference (In Press)*. Sept. 2025.

[9] Sarah Louise Watson et al. “The lifetime risk of surgery in England: a nationwide observational cohort study”. In: *British Journal of Anaesthesia* 133.4 (Oct. 2024), pp. 768–775. ISSN: 14716771. DOI: 10.1016/j.bja.2024.06.028.

[10] Tom E.F. Abbott et al. “Frequency of surgical treatment and related hospital procedures in the UK: a national ecological study using hospital episode statistics”. In: *BJA: British Journal of Anaesthesia* 119.2 (Aug. 2017), pp. 249–257. ISSN: 0007-0912. DOI: 10.1093/BJA/AEX137.

[11] Seiichi Shinji et al. “Recent Advances in the Treatment of Colorectal Cancer: A Review”. In: *J Nippon Med Sch* 89.3 (2022). DOI: 10.1272/jnms.JNMS.2022{_}89-310.

[12] Maximilian Brunner et al. “Current Clinical Strategies of Pancreatic Cancer Treatment and Open Molecular Questions”. In: *International Journal of Molecular Sciences* 20.18 (Sept. 2019). ISSN: 14220067. DOI: 10.3390/IJMS20184543.

[13] Michele Orditura et al. “Treatment of gastric cancer”. In: *World Journal of Gastroenterology : WJG* 20.7 (Feb. 2014), p. 1635. ISSN: 22192840. DOI: 10.3748/WJG.V20.I7.1635.

[14] Candice L. Downey et al. “Impact of in-hospital postoperative complications on quality of life up to 12 months after major abdominal surgery”. In: *BJS* 110.9 (Aug. 2023), pp. 1206–1212. ISSN: 13652168. DOI: 10.1093/BJS/ZNAD167.

[15] Esmee Van Helden et al. “Early postoperative pain and 30-day complications following major abdominal surgery: A retrospective cohort study”. In: *Regional Anesthesia and Pain Medicine* (Aug. 2024). ISSN: 15328651. DOI: 10.1136/RAPM-2024-105277.

[16] Angelica Armellini et al. “The hospital costs of complications following major abdominal surgery: a retrospective cohort study”. In: *BMC Research Notes* 17.1 (Dec. 2024), pp. 1–7. ISSN: 17560500. DOI: 10.1186/S13104-024-06720-Z/FIGURES/2.

[17] Carl Caspersen, Kenneth Powell, and Christenson Gregory. “Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research”. In: *Public Health Rep* 100.2 (1985), pp. 126–31. ISSN: 0033-3549. DOI: 10.2307/20056429.

[18] Denny Z.H. Levett et al. “Perioperative cardiopulmonary exercise testing (CPET): consensus clinical guidelines on indications, organization, conduct, and physiological interpretation”. In: *British journal of anaesthesia* 120.3 (Mar. 2018), pp. 484–500. ISSN: 1471-6771. DOI: 10.1016/J.BJA.2017.10.020.

[19] Jonathon Moran et al. “Role of cardiopulmonary exercise testing as a risk-assessment method in patients undergoing intra-abdominal surgery: a systematic review”. In: *British journal of anaesthesia* 116.2 (Feb. 2016), pp. 177–191. ISSN: 1471-6771. DOI: 10.1093/BJA/AEV454.

[20] Tom B. Smith et al. “Cardiopulmonary exercise testing as a risk assessment method in non cardio-pulmonary surgery: a systematic review”. In: *Anaesthesia* 64.8 (Aug. 2009), pp. 883–893. ISSN: 1365-2044. DOI: 10.1111/J.1365-2044.2009.05983.X.

[21] Ben Li, Arjun Mahajan, and Dylan Powell. “Advancing perioperative care with digital applications and wearables”. In: *npj Digital Medicine 2025 8:1* 8.1 (Apr. 2025), pp. 1–4. ISSN: 2398-6352. DOI: 10.1038/s41746-025-01620-3.

[22] Alexander Hunter. “Integrating wearable devices into perioperative medicine: The potential, and future challenges”. In: *Future Healthcare Journal* 11.3 (Sept. 2024), p. 100169. ISSN: 25146645. DOI: 10.1016/J.FHJ.2024.100169.

[23] Jerome H. Liu et al. “The Increasing Workload of General Surgery”. In: *Archives of Surgery* 139.4 (Apr. 2004), pp. 423–428. ISSN: 0004-0010. DOI: 10.1001/ARCHSURG.139.4.423.

[24] World Cancer Research Fund. *Bowel Cancer*. June 2023. URL: https://www.wcrf.org/preventing-cancer/cancer-types/bowel-cancer/.

[25] Eric Van Cutsem et al. “Improving outcomes in colorectal cancer: Where do we go from here?” In: *European Journal of Cancer* 49.11 (July 2013), pp. 2476–2485. ISSN: 0959-8049. DOI: 10.1016/J.EJCA.2013.03.026.

[26] Cancer Research UK. *Bowel Cancer*. June 2023. URL: https://www.cancerresearchuk.org/about-cancer/bowel-cancer.

[27] Eva J.A. Morris et al. “Thirty-day postoperative mortality after colorectal cancer surgery in England”. In: *Gut* 60.6 (June 2011), pp. 806–813. ISSN: 0017-5749. DOI: 10.1136/GUT.2010.232181.

[28] Cameron I. Wells et al. “”Failure to Rescue” following Colorectal Cancer Resection: Variation and Improvements in a National Study of Postoperative Mortality”. In: *Annals of surgery* 278.1 (July 2023). ISSN: 1528-1140. DOI: 10.1097/SLA.0000000000005650.

[29] Stijn. H.J. Ketelaers et al. “Significant improvement in postoperative and 1-year mortality after colorectal cancer surgery in recent years”. In: *European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology* 45.11 (Nov. 2019), pp. 2052–2058. ISSN: 1532-2157. DOI: 10.1016/J.EJSO.2019.06.017.

[30] Shanika De Silva et al. “Postoperative Complications and Mortality Following Colectomy for Ulcerative Colitis”. In: *Clinical Gastroenterology and Hepatology* 9.11 (Nov. 2011), pp. 972–980. ISSN: 1542-3565. DOI: 10.1016/J.CGH.2011.07.016.

[31] Arnaud Alves et al. “Postoperative Mortality and Morbidity in French Patients Undergoing Colorectal Surgery: Results of a Prospective Multicenter Study”. In: *Archives of Surgery* 140.3 (Mar. 2005), pp. 278–283. ISSN: 0004-0010. DOI: 10.1001/ARCHSURG.140.3.278.

[32] Sarah E. Tevis and Gregory D. Kennedy. “Postoperative complications and implications on patient-centered outcomes”. In: *The Journal of surgical research* 181.1 (May 2013), p. 106. ISSN: 00224804. DOI: 10.1016/J.JSS.2013.01.032.

[33] Elizabeth C. Wick et al. “Readmission rates and cost following colorectal surgery”. In: *Diseases of the colon and rectum* 54.12 (Dec. 2011), pp. 1475–1479. ISSN: 1530-0358. DOI: 10.1097/DCR.0B013E31822FF8F0.

[34] Maleck Louis et al. “Financial burden of postoperative complications following colonic resection: A systematic review”. In: *Medicine* 100.27 (July 2021). ISSN: 1536-5964. DOI: 10.1097/MD.0000000000026546.

[35] Guy L. Ludbrook. “The Hidden Pandemic: the Cost of Postoperative Complications”. In: *Current anesthesiology reports* 12.1 (Mar. 2022). ISSN: 1523-3855. DOI: 10.1007/S40140-021-00493-Y.

[36] A. Zambouri. “Preoperative evaluation and preparation for anesthesia and surgery”. In: *Hippokratia* 11.1 (2007), p. 13. ISSN: 11084189.

[37] Developed by the National Guideline Centre. *Routine preoperative tests for elective surgery*. Tech. rep. Royal College of Physicians, 2016.

[38] Mike . P.W. Grocott and R. M. Pearse. “Perioperative medicine: the future of anaesthesia?” In: *BJA: British Journal of Anaesthesia* 108.5 (May 2012), pp. 723–726. ISSN: 0007-0912. DOI: 10.1093/BJA/AES124.

[39] Mike P.W. Grocott et al. “Peri-operative care pathways: re-engineering care to achieve the ‘triple aim’”. In: *Anaesthesia* 74 (Jan. 2019), pp. 90–99. ISSN: 1365-2044. DOI: 10.1111/ANAE.14513.

[40] Peter G. Noordzij et al. “Prognostic Value of Routine Preoperative Electrocardiography in Patients Undergoing Noncardiac Surgery”. In: *The American Journal of Cardiology* 97.7 (Apr. 2006), pp. 1103–1106. ISSN: 0002-9149. DOI: 10.1016/J.AMJCARD.2005.10.058.

[41] Jacqueline E.M. Vernooij et al. “Performance and usability of pre-operative prediction models for 30-day peri-operative mortality risk: a systematic review”. In: *Anaesthesia* 78.5 (May 2023), pp. 607–619. ISSN: 1365-2044. DOI: 10.1111/ANAE.15988.

[42] Douglas P. Wagner and E. A. Draper. “Acute physiology and chronic health evaluation (APACHE II) and Medicare reimbursement”. In: *Health Care Financing Review* 1984.Suppl (1984), p. 91. ISSN: 01958631.

[43] H. J. S. Jones and L. De Cossart. “Risk scoring in surgical patients”. In: *British Journal of Surgery* 86.2 (1999), pp. 149–157. ISSN: 0007-1323. DOI: 10.1046/j.1365-2168.1999.01006.x.

[44] Graham P. Copeland. “The POSSUM System of Surgical Audit”. In: *Archives of Surgery* 137.1 (Jan. 2002), pp. 15–19. ISSN: 0004-0010. DOI: 10.1001/ARCHSURG.137.1.15.

[45] Alexios Dosis et al. “Estimating postoperative mortality in colorectal surgery- a systematic review of risk prediction models”. In: *International Journal of Colorectal Disease* 38.1 (Dec. 2023), pp. 1–13. ISSN: 14321262. DOI: 10.1007/S00384-023-04455-0/FIGURES/4.

[46] Nisha Pradhan et al. “Attitudes about use of preoperative risk assessment tools: a survey of surgeons and surgical residents in an academic health system”. In: *Patient Safety in Surgery* 16.1 (Dec. 2022), pp. 1–9. ISSN: 17549493. DOI: 10.1186/S13037-022-00320-1/TABLES/4.

[47] Luigi Goffi et al. “Preoperative APACHE II and ASA scores in patients having major general surgical operations: prognostic value and potential clinical applications”. In: *The European journal of surgery = Acta chirurgica* 165.8 (1999), pp. 730–735. ISSN: 1102-4151. DOI: 10.1080/11024159950189483.

[48] Ross Arena et al. “Assessment of Functional Capacity in Clinical and Research Settings”. In: *Circulation* 116.3 (July 2007), pp. 329–343. ISSN: 00097322. DOI: 10.1161/CIRCULATIONAHA.106.184461.

[49] João J. Ferreira et al. “Wearable technology and consumer interaction: A systematic review and research agenda”. In: *Computers in Human Behavior* 118 (May 2021), p. 106710. ISSN: 0747-5632. DOI: 10.1016/J.CHB.2021.106710.

[50] Nancy E. Mayo et al. “Impact of preoperative change in physical function on postoperative recovery: Argument supporting prehabilitation for colorectal surgery”. In: *Surgery* 150.3 (Sept. 2011), pp. 505–514. ISSN: 0039-6060. DOI: 10.1016/J.SURG.2011.07.045.

[51] Preet G S Makker et al. “Preoperative functional capacity and postoperative outcomes following abdominal and pelvic cancer surgery: a systematic review and meta-analysis”. In: *ANZ Journal of Surgery* 92.7-8 (July 2022), pp. 1658–1667. ISSN: 1445-2197. DOI: 10.1111/ANS.17577.

[52] Earlene Silvapulle and J Darvall. “Objective methods for preoperative assessment of functional capacity”. In: *BJA Education* 22 (2022), pp. 312–320. DOI: 10.1016/j.bjae.2022.03.003.

[53] Denny Z.H. Levett et al. “Perioperative cardiopulmonary exercise testing (CPET): consensus clinical guidelines on indications, organization, conduct, and physiological interpretation”. In: *British Journal of Anaesthesia* 120.3 (Mar. 2018), pp. 484–500. ISSN: 14716771. DOI: 10.1016/J.BJA.2017.10.020/ATTACHMENT/EC8AA2ED-4274-431B-9378-941C1B68FF96/MMC2.PDF.

[54] Khaled Albouaini et al. “Cardiopulmonary exercise testing and its application”. In: *Heart* 93.10 (Nov. 2007), p. 1285. ISSN: 00325473. DOI: 10.1136/HRT.2007.121558.

[55] Philip J. Hennis, Paula M. Meale, and Michael P.W. Grocott. “Cardiopulmonary exercise testing for the evaluation of perioperative risk in non-cardiopulmonary surgery”. In: *Postgraduate medical journal* 87.1030 (Aug. 2011), pp. 550–557. ISSN: 1469-0756. DOI: 10.1136/PGMJ.2010.107185.

[56] Denny Z.H. Levett and Michael P.W. Grocott. “Cardiopulmonary exercise testing for risk prediction in major abdominal surgery”. In: *Anesthesiology clinics* 33.1 (Mar. 2015), pp. 1–16. ISSN: 1932-2275. DOI: 10.1016/J.ANCLIN.2014.11.001.

[57] T. Reeves et al. “Cardiopulmonary exercise testing (CPET) in the United Kingdom-a national survey of the structure, conduct, interpretation and funding”. In: *Perioperative medicine (London, England)* 7.1 (Dec. 2018). ISSN: 2047-0525. DOI: 10.1186/S13741-017-0082-3.

[58] George A. Rose et al. “‘Fit for surgery’: the relationship between cardiorespiratory fitness and postoperative outcomes”. In: *Experimental Physiology* 107.8 (Aug. 2022), pp. 787–799. ISSN: 1469445X. DOI: 10.1113/EP090156.

[59] Laura Jones et al. “Can wearable technology be used to approximate cardiopulmonary exercise testing metrics?” In: *Perioperative Medicine* 10.1 (Dec. 2021). DOI: 10.1186/S13741-021-00180-W.

[60] Vini Vijayan et al. “Review of Wearable Devices and Data Collection Considerations for Connected Health”. In: *Sensors (Basel, Switzerland)* 21.16 (Aug. 2021). ISSN: 14248220. DOI: 10.3390/S21165589.

[61] Matthew Smuck et al. “The emerging clinical role of wearables: factors for successful implementation in healthcare”. In: *npj Digital Medicine 2021 4:1* 4.1 (Mar. 2021), pp. 1–8. ISSN: 2398-6352. DOI: 10.1038/s41746-021-00418-3.

[62] Health Quality Ontario. “Long-Term Continuous Ambulatory ECG Monitors and External Cardiac Loop Recorders for Cardiac Arrhythmia: A Health Technology Assessment”. In: *Ontario Health Technology Assessment Series* 17.1 (2017), p. 1. ISSN: 19157398.

[63] Mintu P. Turakhia et al. “Diagnostic Utility of a Novel Leadless Arrhythmia Monitoring Device”. In: *American Journal of Cardiology* 112.4 (Aug. 2013), pp. 520–524. ISSN: 0002-9149. DOI: 10.1016/J.AMJCARD.2013.04.017.

[64] Jiali Yao et al. “Number of daily measurements needed to estimate habitual step count levels using wrist-worn trackers and smartphones in 212,048 adults”. In: *Scientific Reports 2021 11:1* 11.1 (May 2021), pp. 1–10. ISSN: 2045-2322. DOI: 10.1038/s41598-021-89141-3.

[65] Luiza Isnardi Cardoso Ricardo et al. “Number of days required to estimate physical activity constructs objectively measured in different age groups: Findings from three Brazilian (Pelotas) population-based birth cohorts”. In: *PLoS ONE* 15.1 (Jan. 2020). ISSN: 19326203. DOI: 10.1371/JOURNAL.PONE.0216017.

[66] Andreas B. Böhmer, Frank Wappler, and Bernd Zwißer. “Preoperative Risk Assessment—From Routine Tests to Individualized Investigation”. In: *Deutsches Ärzteblatt International* 111.25 (June 2014), p. 437. ISSN: 18660452. DOI: 10.3238/ARZTEBL.2014.0437.

[67] Benoit Lequeux, Charles Uzan, and Michaela B. Rehman. “Does resting heart rate measured by the physician reflect the patient’s true resting heart rate? White-coat heart rate”. In: *Indian heart journal* 70.1 (Jan. 2018), pp. 93–98. ISSN: 2213-3763. DOI: 10.1016/J.IHJ.2017.07.015.

[68] Shraddha D. Deshmukh and Swati N. Shilaskar. “Wearable sensors and patient monitoring system: A Review”. In: *2015 International Conference on Pervasive Computing: Advance Communication Technology and Application for Society, ICPC 2015* (Apr. 2015). DOI: 10.1109/PERVASIVE.2015.7086982.

[69] Walter Maetzler et al. “Quantitative wearable sensors for objective assessment of Parkinson’s disease”. In: *Movement Disorders* 28.12 (Oct. 2013), pp. 1628–1637. ISSN: 1531-8257. DOI: 10.1002/MDS.25628.

[70] Nadtinan Promphet et al. “Cotton thread-based wearable sensor for non-invasive simultaneous diagnosis of diabetes and kidney failure”. In: *Sensors and Actuators B: Chemical* 321 (Oct. 2020), p. 128549. ISSN: 0925-4005. DOI: 10.1016/J.SNB.2020.128549.

[71] Giorgio Quer et al. “Wearable sensor data and self-reported symptoms for COVID-19 detection”. In: *Nature Medicine 2020 27:1* 27.1 (Oct. 2020), pp. 73–77. ISSN: 1546-170X. DOI: 10.1038/s41591-020-1123-x.

[72] Priyanka Kakria, N. K. Tripathi, and Peerapong Kitipawang. “A real-time health monitoring system for remote cardiac patients using smartphone and wearable sensors”. In: *International Journal of Telemedicine and Applications* 2015 (2015). ISSN: 16876423. DOI: 10.1155/2015/373474.

[73] Jian-Dong ; Huang et al. “Applying Artificial Intelligence to Wearable Sensor Data to Diagnose and Predict Cardiovascular Disease: A Review”. In: *Sensors 2022, Vol. 22, Page 8002* 22.20 (Oct. 2022), p. 8002. ISSN: 1424-8220. DOI: 10.3390/S22208002.

[74] Cameron I Wells et al. “Wearable devices to monitor recovery after abdominal surgery: scoping review”. In: *BJS open* 6.2 (May 2022). ISSN: 2474-9842. DOI: 10.1093/BJSOPEN/ZRAC031.

[75] van der Stam et al. “Machine Learning for Postoperative Continuous Recovery Scores of Oncology Patients in Perioperative Care with Data from Wearables”. In: *Sensors 2023, Vol. 23, Page 4455* 23.9 (May 2023), p. 4455. ISSN: 1424-8220. DOI: 10.3390/S23094455.

[76] Timothy J. Daskivich et al. “Association of Wearable Activity Monitors With Assessment of Daily Ambulation and Length of Stay Among Patients Undergoing Major Surgery”. In: *JAMA Network Open* 2.2 (Feb. 2019), e187673–e187673. ISSN: 25743805. DOI: 10.1001/JAMANETWORKOPEN.2018.7673.

[77] Pedro A. Esteban et al. “Evaluating patients’ walking capacity during hospitalization for lung cancer resection”. In: *Interactive cardiovascular and thoracic surgery* 25.2 (Aug. 2017), pp. 268–271. ISSN: 1569-9285. DOI: 10.1093/ICVTS/IVX100.

[78] Juha K.A. Rinne et al. “Evaluation of a wrist-worn photoplethysmography monitor for heart rate variability estimation in patients recovering from laparoscopic colon resection”. In: *Journal of Clinical Monitoring and Computing* 37.1 (Feb. 2023), pp. 45–53. ISSN: 15732614. DOI: 10.1007/S10877-022-00854-W/FIGURES/3.

[79] Martine J.M. Breteler et al. “Reliability of wireless monitoring using a wearable patch sensor in high-risk surgical patients at a step-down unit in the Netherlands: a clinical validation study”. In: *BMJ Open* 8.2 (Feb. 2018), e020162. ISSN: 2044-6055. DOI: 10.1136/BMJOPEN-2017-020162.

[80] Ellen Waller et al. “Prehabilitation with wearables versus standard of care before major abdominal cancer surgery: a randomised controlled pilot study (trial registration: NCT04047524)”. In: *Surgical Endoscopy* 36.2 (Feb. 2022), p. 1008. ISSN: 14322218. DOI: 10.1007/S00464-021-08365-6.

[81] Christian M. Beilstein et al. “Multimodal prehabilitation for major surgery in elderly patients to lower complications: protocol of a randomised, prospective, multicentre, multidisciplinary trial (PREHABIL Trial)”. In: *BMJ open* 13.1 (Jan. 2023), e070253. ISSN: 20446055. DOI: 10.1136/BMJOPEN-2022-070253.

[82] C. Feeney, J. V. Reynolds, and J. Hussey. “Preoperative physical activity levels and postoperative pulmonary complications post-esophagectomy”. In: *Diseases of the Esophagus* 24.7 (Sept. 2011), pp. 489–494. ISSN: 1120-8694. DOI: 10.1111/J.1442-2050.2010.01171.X.

[83] Massimiliano Greco et al. “Wearable Health Technology for Preoperative Risk Assessment in Elderly Patients: The WELCOME Study”. In: *Diagnostics 2023, Vol. 13, Page 630* 13.4 (Feb. 2023), p. 630. ISSN: 2075-4418. DOI: 10.3390/DIAGNOSTICS13040630.

[84] Dimitris Spathis et al. “Self-supervised transfer learning of physiological representations from free-living wearable data”. In: *ACM CHIL 2021 - Proceedings of the 2021 ACM Conference on Health, Inference, and Learning* 21 (Apr. 2021), pp. 69–78. DOI: 10.1145/3450439.3451863.

[85] Stephanie Soon et al. “Wearable devices for remote vital signs monitoring in the outpatient setting: an overview of the field”. In: *BMJ Innovations* 6.2 (Apr. 2020), p. 55. ISSN: 2055-8074. DOI: 10.1136/BMJINNOV-2019-000354.

[86] John M. Taylor and Michael A. Gropper. “Critical care challenges in orthopedic surgery patients”. In: *Critical Care Medicine* 34.SUPPL. 9 (Sept. 2006). ISSN: 00903493. DOI: 10.1097/01.CCM.0000231880.18476.D8.

[87] Neal Robert Haddaway et al. “The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching”. In: *PLOS ONE* 10.9 (Sept. 2015), e0138237. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0138237.

[88] Heidy Cos et al. “Predicting Outcomes in Patients Undergoing Pancreatectomy Using Wearable Technology and Machine Learning: Prospective Cohort Study”. In: *J Med Internet Res 2021;23(3):e23595 https://www.jmir.org/2021/3/e23595* 23.3 (May 2021), e23595. ISSN: 14388871. DOI: 10.2196/23595.

[89] Jingwen Zhang et al. “Predicting Post-Operative Complications with Wearables”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6.2 (July 2022). ISSN: 24749567. DOI: 10.1145/3534578.

[90] Caspar F. Mylius et al. “Objectively measured preoperative physical activity is associated with time to functional recovery after hepato-pancreato-biliary cancer surgery: a pilot study”. In: *Perioperative Medicine 2021 10:1* 10.1 (Oct. 2021), pp. 1–12. ISSN: 2047-0525. DOI: 10.1186/S13741-021-00202-7.

[91] Alessandra Angelucci et al. “Fitbit Data to Assess Functional Capacity in Patients Before Elective Surgery: Pilot Prospective Observational Study”. In: *Journal of Medical Internet Research* 25 (Apr. 2023), e42815. ISSN: 1438-8871. DOI: 10.2196/42815.

[92] Traci L. Hedrick et al. “Wearable Technology in the Perioperative Period: Predicting Risk of Postoperative Complications in Patients Undergoing Elective Colorectal Surgery”. In: *Diseases of the colon and rectum* 63.4 (Apr. 2020), pp. 538–544. ISSN: 1530-0358. DOI: 10.1097/DCR.0000000000001580.

[93] Afsaneh Doryab et al. “Modeling Biobehavioral Rhythms with Passive Sensing in the Wild”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.1 (Mar. 2019), pp. 1–21. DOI: 10.1145/3314395.

[94] Virginia Sun et al. “Wireless Monitoring Program of Patient-Centered Outcomes and Recovery Before and After Major Abdominal Cancer Surgery”. In: *JAMA surgery* 152.9 (Sept. 2017), pp. 852–859. ISSN: 2168-6262. DOI: 10.1001/JAMASURG.2017.1519.

[95] Hiroki Nakajima et al. “How Many Steps Per Day are Necessary to Prevent Postoperative Complications Following Hepato-Pancreato-Biliary Surgeries for Malignancy?” In: *Annals of surgical oncology* 27.5 (May 2020), pp. 1387–1397. ISSN: 1534-4681. DOI: 10.1245/S10434-020-08218-X.

[96] Nuria Maria Novoa et al. “Value of the average basal daily walked distance measured using a pedometer to predict maximum oxygen consumption per minute in patients undergoing lung resection”. In: *European Journal of Cardio-Thoracic Surgery* 39.5 (May 2011), pp. 756–762. ISSN: 1010-7940. DOI: 10.1016/J.EJCTS.2010.08.025.

[97] Andrea Billé et al. “Preoperative Physical Activity Predicts Surgical Outcomes Following Lung Cancer Resection”. In: *Integrative cancer therapies* 20 (2021). ISSN: 1552-695X. DOI: 10.1177/1534735420975853.

[98] Simon J. G. Richards et al. “The association between low pre-operative step count and adverse post-operative outcomes in older patients undergoing colorectal cancer surgery”. In: *Perioperative medicine (London, England)* 9.1 (Dec. 2020). ISSN: 2047-0525. DOI: 10.1186/S13741-020-00150-8.

[99] Che Chang Yang and Yeh Liang Hsu. “A Review of Accelerometry-Based Wearable Motion Detectors for Physical Activity Monitoring”. In: *Sensors (Basel, Switzerland)* 10.8 (Aug. 2010), p. 7772. ISSN: 14248220. DOI: 10.3390/S100807772.

[100] Katherine S Hall et al. “Systematic review of the prospective association of daily step counts with risk of mortality, cardiovascular disease, and dysglycemia”. In: *International Journal of Behavioral Nutrition and Physical Activity* 17.1 (May 2020), pp. 1–14. ISSN: 14795868. DOI: 10.1186/S12966-020-00978-9/FIGURES/2.

[101] Min Zhao et al. “Recommended physical activity and all cause and cause specific mortality in US adults: prospective cohort study”. In: *BMJ* 370 (July 2020), p. 2031. ISSN: 1756-1833. DOI: 10.1136/BMJ.M2031.

[102] Pedro F. Saint-Maurice et al. “Moderate-to-Vigorous Physical Activity and All-Cause Mortality: Do Bouts Matter?” In: *Journal of the American Heart Association* 7.6 (Mar. 2018). ISSN: 20479980. DOI: 10.1161/JAHA.117.007678.

[103] William J. Kane et al. “Wearable technology and the association of perioperative activity level with 30-day readmission among patients undergoing major colorectal surgery”. In: *Surgical endoscopy* 36.2 (Feb. 2022), pp. 1584–1592. ISSN: 1432-2218. DOI: 10.1007/S00464-021-08449-3.

[104] Lorenzo A. Rossi et al. “Predicting post-discharge cancer surgery complications via telemonitoring of patient-reported outcomes and patient-generated health data”. In: *Journal of surgical oncology* 123.5 (Apr. 2021), pp. 1345–1352. ISSN: 1096-9098. DOI: 10.1002/JSO.26413.

[105] Micah T. Eades et al. “Smartphone-recorded physical activity for estimating cardiorespiratory fitness”. In: *Scientific Reports 2021 11:1* 11.1 (July 2021), pp. 1–6. ISSN: 2045-2322. DOI: 10.1038/s41598-021-94164-x.

[106] Haraldur T Hallgrímsson et al. “Learning Individualized Cardiovascular Responses from Large-scale Wearable Sensors Data”. In: (Dec. 2018). DOI: 10.48550/arxiv.1812.01696.

[107] Dimitris Spathis et al. “Longitudinal cardio-respiratory fitness prediction through wearables in free-living environments”. In: *npj Digital Medicine 2022 5:1* 5.1 (Dec. 2022), pp. 1–11. ISSN: 2398-6352. DOI: 10.1038/s41746-022-00719-1.

[108] Wim H.M. Saris and R. A. Binkhorst. “The use of pedometer and actometer in studying daily physical activity in man. Part I: Reliability of pedometer and actometer”. In: *European Journal of Applied Physiology and Occupational Physiology* 37.3 (Sept. 1977), pp. 219–228. ISSN: 03015548. DOI: 10.1007/BF00421777/METRICS.

[109] H. W. Cui et al. “The association of pre-operative home accelerometry with cardiopulmonary exercise variables”. In: *Anaesthesia* 73.6 (June 2018), pp. 738–745. ISSN: 1365-2044. DOI: 10.1111/ANAE.14181.

[110] Jermana L Moraes et al. “Advances in Photopletysmography Signal Analysis for Biomedical Applications”. In: *Sensors 2018, Vol. 18, Page 1894* 18.6 (May 2018), p. 1894. ISSN: 1424-8220. DOI: 10.3390/S18061894.

[111] Creative Commons. *CC BY 4.0 Deed*. URL: https://creativecommons.org/licenses/by/4.0/.

[112] Dustin T. Weiler et al. “Wearable heart rate monitor technology accuracy in research: A comparative study between PPG and ECG technology”. In: *Proceedings of the Human Factors and Ergonomics Society* 2017-October (2017), pp. 1292–1296. ISSN: 10711813. DOI: 10.1177/1541931213601804.

[113] William B. Kannel et al. “Heart rate and cardiovascular mortality: The Framingham study”. In: *American Heart Journal* 113.6 (June 1987), pp. 1489–1494. ISSN: 0002-8703. DOI: 10.1016/0002-8703(87)90666-1.

[114] Dongfeng Zhang, Xiaoli Shen, and Xin Qi. “Resting heart rate and all-cause and cardiovascular mortality in the general population: A meta-analysis”. In: *CMAJ* 188.3 (Feb. 2016), E53–E63. ISSN: 14882329. DOI: 10.1503/CMAJ.150535/-/DC1.

[115] Tom E F Abbott et al. “Preoperative heart rate and myocardial injury after non-cardiac surgery: results of a predefined secondary analysis of the VISION study”. In: *BJA: British Journal of Anaesthesia* 117.2 (May 2016), p. 172. ISSN: 14716771. DOI: 10.1093/BJA/AEW182.

[116] William K. Freeman and Raymond J. Gibbons. “Perioperative Cardiovascular Assessment of Patients Undergoing Noncardiac Surgery”. In: *Mayo Clinic Proceedings* 84.1 (2009), p. 79. ISSN: 00256196. DOI: 10.4065/84.1.79.

[117] Mohamed Elgendi et al. “The use of photoplethysmography for assessing hypertension”. In: *npj Digital Medicine* 2.1 (Dec. 2019). ISSN: 23986352. DOI: 10.1038/S41746-019-0136-7.

[118] Denisse Castaneda et al. “A review on wearable photoplethysmography sensors and their potential future applications in health care”. In: *International journal of biosensors & bioelectronics* 4.4 (2018), p. 195. ISSN: 2573-2838. DOI: 10.15406/IJBSBE.2018.04.00125.

[119] Robert Avram et al. “Real-world heart rate norms in the Health eHeart study”. In: *NPJ Digital Medicine* 2.1 (Dec. 2019). ISSN: 23986352. DOI: 10.1038/S41746-019-0134-9.

[120] Robert Wang et al. “Accuracy of Wrist-Worn Heart Rate Monitors”. In: *JAMA Cardiology* 2.1 (May 2017), pp. 104–106. ISSN: 2380-6583. DOI: 10.1001/JAMACARDIO.2016.3340.

[121] Hirofumi Tanaka, Kevin D. Monahan, and Douglas R. Seals. “Age-predicted maximal heart rate revisited”. In: *Journal of the American College of Cardiology* 37.1 (Jan. 2001), pp. 153–156. ISSN: 0735-1097. DOI: 10.1016/S0735-1097(00)01054-8.

[122] Jessilyn Dunn et al. “Wearable sensors enable personalized predictions of clinical laboratory measurements”. In: *Nature Medicine 2021 27:6* 27.6 (May 2021), pp. 1105–1112. ISSN: 1546-170X. DOI: 10.1038/s41591-021-01339-0.

[123] HEXOSKIN. *HEXOSKIN PROSHIRT - MEN'S*. 2023. URL: https://hexoskin.com/products/hexoskin-proshirt-mens.

[124] Marco Altini. *(PDF) Personalization of energy expenditure and cardiorespiratory fitness estimation using wearable sensors in supervised and unsupervised free-living conditions*. 2015. URL: https://www.researchgate.net/publication/287204378_Personalization_of_energy_expenditure_and_cardiorespiratory_fitness_estimation_using_wearable_sensors_in_supervised_and_unsupervised_free-living_conditions.

[125] Marjolein E. Haveman et al. “Feasibility and patient's experiences of perioperative telemonitoring in major abdominal surgery: an observational pilot study”. In: *Expert Review of Medical Devices* 19.6 (2022), pp. 515–523. ISSN: 17452422. DOI: 10.1080/17434440.2022.2108703.

[126] Yan Li et al. “Evaluation of a Physical-Psychological Integrative (PPI) intervention for community-dwelling spinal cord injury survivors: Study protocol of a preliminary randomized controlled trial”. In: (2023). DOI: 10.1371/journal.pone.0282846.

[127] Yvonne Kiera Bartlett, Thomas L. Webb, and Mark S. Hawley. “Using Persuasive Technology to Increase Physical Activity in People With Chronic Obstructive Pulmonary Disease by Encouraging Regular Walking: A Mixed-Methods Study Exploring Opinions and Preferences”. In: *Journal of Medical Internet Research* 19.4 (Apr. 2017). ISSN: 14388871. DOI: 10.2196/JMIR.6616.

[128] Sanchit Kumar et al. “Wearables in Cardiovascular Disease”. In: *Journal of Cardiovascular Translational Research* (2022). ISSN: 19375395. DOI: 10.1007/S12265-022-10314-0.

[129] *What is an electrocardiogram (ECG)? - InformedHealth.org - NCBI Bookshelf*. URL: https://www.ncbi.nlm.nih.gov/books/NBK536878/.

[130] World Health Organisation. *Cardiovascular diseases (CVDs)*. June 2021. URL: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

[131] Steve Meek and Francis Morris. “ABC of clinical electrocardiography: Introduction. I—Leads, rate, rhythm, and cardiac axis”. In: *BMJ : British Medical Journal* 324.7334 (Feb. 2002), p. 415. ISSN: 14685833. DOI: 10.1136/BMJ.324.7334.415.

[132] Stefan Sammito and Irina Böckelmann. "[Options and limitations of heart rate measurement and analysis of heart rate variability by mobile devices: A systematic review]". In: *Herzschrittmachertherapie & Elektrophysiologie* 27.1 (Mar. 2016), pp. 38–45. ISSN: 1435-1544. DOI: 10.1007/S00399-016-0419-5.

[133] Mikkel Nicklas Frandsen et al. "Preoperative heart rate variability as a predictor of perioperative outcomes: a systematic review without meta-analysis". In: *Journal of Clinical Monitoring and Computing* 36.4 (Aug. 2022), pp. 947–960. ISSN: 15732614. DOI: 10.1007/S10877-022-00819-Z.

[134] Steven R. Steinhubl et al. "Effect of a Home-Based Wearable Continuous ECG Monitoring Patch on Detection of Undiagnosed Atrial Fibrillation: The mSToPS Randomized Clinical Trial". In: *JAMA* 320.2 (July 2018), pp. 146–155. ISSN: 0098-7484. DOI: 10.1001/JAMA.2018.8102.

[135] Sameer Prasada et al. "Preoperative Atrial Fibrillation and Cardiovascular Outcomes After Noncardiac Surgery". In: *Journal of the American College of Cardiology* 79.25 (June 2022), pp. 2471–2485. ISSN: 1558-3597. DOI: 10.1016/J.JACC.2022.04.021.

[136] Finlay A McAlister et al. "A comparison of four risk models for the prediction of cardiovascular complications in patients with a history of atrial fibrillation undergoing non-cardiac surgery". In: *Anaesthesia* 75.1 (May 2020), pp. 27–36. ISSN: 1365-2044. DOI: 10.1111/ANAE.14777.

[137] Yu Wu et al. *Turning Silver into Gold: Domain Adaptation with Noisy Labels for Wearable Cardio-Respiratory Fitness Prediction*. Nov. 2022. URL: http://arxiv.org/abs/2211.10475.

[138] Marco Altini et al. "Cardiorespiratory fitness estimation in free-living using wearable sensors". In: *Artificial Intelligence in Medicine* 68 (Mar. 2016), pp. 37–46. ISSN: 0933-3657. DOI: 10.1016/J.ARTMED.2016.02.002.

[139] Thomas Beltrame et al. "Prediction of oxygen uptake dynamics by machine learning analysis of wearable sensors during activities of daily living". In: *Scientific Reports 2017 7:1* 7.1 (Apr. 2017), pp. 1–8. ISSN: 2045-2322. DOI: 10.1038/srep45738.

[140] Sylvia Cho et al. "Factors Affecting the Quality of Person-Generated Wearable Device Data and Associated Challenges: Rapid Systematic Review". In: *JMIR mHealth and uHealth* 9.3 (May 2021). ISSN: 22915222. DOI: 10.2196/20738.

[141] Roderick J.A. Little and Donald B. Rubin. "Statistical analysis with missing data". In: *Statistical Analysis with Missing Data* (Jan. 2019), pp. 1–449. DOI: 10.1002/9781119482260.

[142] Janus Christian Jakobsen et al. “When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts”. In: *BMC Medical Research Methodology* 17.1 (May 2017), pp. 1–10. ISSN: 14712288. DOI: 10.1186/S12874-017-0442-1/TABLES/2.

[143] Jay Darji et al. “Handling missing data in the time-series data from wearables”. In: *Time Series Analysis - Recent Advances, New Perspectives and Applications [Working Title]* (Aug. 2023). DOI: 10.5772/INTECHOPEN.1002536.

[144] Mia S. Tackney et al. “A framework for handling missing accelerometer outcome data in trials”. In: *Trials* 22.1 (Dec. 2021), pp. 1–18. ISSN: 17456215. DOI: 10.1186/S13063-021-05284-8/FIGURES/8.

[145] Marjolein E Haveman et al. “Determining the Reliable Measurement Period for Preoperative Baseline Values With Telemonitoring Before Major Abdominal Surgery: Pilot Cohort Study.” In: *JMIR perioperative medicine* 5.1 (2022), e40815. ISSN: 2561-9128. DOI: https://dx.doi.org/10.2196/40815.

[146] Julia Y Lin, Ying Lu, and Xin Tu. “How to avoid missing data and the problems they pose: design considerations”. In: *Shanghai Archives of Psychiatry* 24.3 (May 2012), p. 181. ISSN: 10020829. DOI: 10.3969/J.ISSN.1002-0829.2012.03.010.

[147] Phayung Meesad and Kairung Hengpraprohm. “Combination of KNN-based feature selection and KNN-based missing-value imputation of microarray data”. In: *3rd International Conference on Innovative Computing Information and Control, ICICIC’08* (2008). DOI: 10.1109/ICICIC.2008.635.

[148] Syed Khairul Bashar et al. “Noise Detection in Electrocardiogram Signals for Intensive Care Unit Patients”. In: *IEEE access : practical innovations, open solutions* 7 (2019), p. 88357. ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2926199.

[149] Seokhoon Kang, Anand Paul, and Gwanggil Jeon. “Reduction of mixed noise from wearable sensors in human-motion estimation”. In: *Computers & Electrical Engineering* 61 (July 2017), pp. 287–296. ISSN: 0045-7906. DOI: 10.1016/J.COMPELECENG.2017.05.030.

[150] Oliver Stegle et al. “Gaussian process robust regression for noisy heart rate data”. In: *IEEE transactions on bio-medical engineering* 55.9 (Sept. 2008), pp. 2143–2151. ISSN: 1558-2531. DOI: 10.1109/TBME.2008.923118.

[151] Sridhar Krishnan and Yashodhan Athavale. “Trends in biomedical signal feature extraction”. In: *Biomedical Signal Processing and Control* 43 (May 2018), pp. 41–63. ISSN: 1746-8094. DOI: 10.1016/J.BSPC.2018.02.008.

[152] Samina Khalid, Tehmina Khalil, and Shamila Nasreen. “A survey of feature selection and feature extraction techniques in machine learning”. In: *Proceedings of 2014 Science and Information Conference, SAI 2014* (Oct. 2014), pp. 372–378. DOI: 10.1109/SAI.2014.6918213.

[153] Ngoc-Huynh Ho et al. “Step-Detection and Adaptive Step-Length Estimation for Pedestrian Dead-Reckoning at Various Walking Speeds Using a Smartphone”. In: (2016). DOI: 10.3390/s16091423.

[154] Xiaokun Yang and Baoqi Huang. “An accurate step detection algorithm using unconstrained smartphones”. In: *Proceedings of the 2015 27th Chinese Control and Decision Conference, CCDC 2015* (July 2015), pp. 5682–5687. DOI: 10.1109/CCDC.2015.7161816.

[155] Scott W. Ducharme et al. “A Transparent Method for Step Detection using an Acceleration Threshold”. In: *Journal for the measurement of physical behaviour* 4.4 (Dec. 2021), p. 311. ISSN: 25756613. DOI: 10.1123/JMPB.2021-0011.

[156] Andrea Mannini and Angelo Maria Sabatini. “Machine learning methods for classifying human physical activity from on-body accelerometers”. In: *Sensors (Basel, Switzerland)* 10.2 (Feb. 2010), pp. 1154–1175. ISSN: 1424-8220. DOI: 10.3390/S100201154.

[157] Long Luu et al. “Accurate Step Count with Generalized and Personalized Deep Learning on Accelerometer Data”. In: *Sensors 2022, Vol. 22, Page 3989* 22.11 (May 2022), p. 3989. ISSN: 1424-8220. DOI: 10.3390/S22113989.

[158] Catrine Tudor-Locke et al. “How many steps/day are enough? For older adults and special populations”. In: *International Journal of Behavioral Nutrition and Physical Activity* 8.1 (July 2011), pp. 1–19. ISSN: 14795868. DOI: 10.1186/1479-5868-8-80/METRICS.

[159] Ali Neishabouri et al. “Quantification of acceleration as activity counts in ActiGraph wearable”. In: *Scientific Reports 2022 12:1* 12.1 (July 2022), pp. 1–8. ISSN: 2045-2322. DOI: 10.1038/s41598-022-16003-x.

[160] Carla Elane Silva dos Santos and Cassiano Ricardo Rech. “Association between different cutoff points for objectively measured moderate-to-vigorous physical activity and cardiometabolic markers in older adults”. In: *Archives of Gerontology and Geriatrics* 91 (2020). DOI: 10.1016/j.archger.2020.104238.

[161] Willis J Tompkins. *A Real-Time QRS Detection Algorithm*. Tech. rep. 3. 1985.

[162] Manuel Merino-Monge et al. “Heartbeat detector from ECG and PPG signals based on wavelet transform and upper envelopes”. In: *Physical and Engineering Sciences in Medicine* 46.2 (June 2023), pp. 597–608. ISSN: 26624737. DOI: 10.1007/S13246-023-01235-6/TABLES/4.

[163] Florian Kristof et al. “QRS detection in single-lead, telehealth electrocardiogram signals: Benchmarking open-source algorithms”. In: *PLOS Digital Health* 3.8 (Aug. 2024), e0000538. ISSN: 2767-3170. DOI: 10.1371/JOURNAL.PDIG.0000538.

[164] Cathy Speed et al. “Measure by measure: Resting heart rate across the 24-hour cycle”. In: *PLOS Digital Health* 2.4 (Apr. 2023), e0000236. DOI: 10.1371/JOURNAL.PDIG.0000236.

[165] Zhan Liu et al. “Resting heart rate as a preoperative predictor of postoperative atrial fibrillation after pulmonary thromboendarterectomy”. In: *Journal of cardiac surgery* 37.6 (June 2022), pp. 1644–1650. ISSN: 1540-8191. DOI: 10.1111/JOCS.16407.

[166] Karim S. Ladha et al. “Association between preoperative ambulatory heart rate and postoperative myocardial injury: a retrospective cohort study”. In: *British Journal of Anaesthesia* 121.4 (Oct. 2018), pp. 722–729. ISSN: 14716771. DOI: 10.1016/j.bja.2018.06.016.

[167] Nikhil Singh et al. “Heart Rate Variability: An Old Metric with New Meaning in the Era of using mHealth Technologies for Health and Exercise Training Guidance. Part One: Physiology and Methods”. In: *Arrhythmia & Electrophysiology Review* 7.3 (May 2018), p. 193. ISSN: 20503377. DOI: 10.15420/AER.2018.27.2.

[168] Petr Reimer et al. “Role of heart-rate variability in preoperative assessment of physiological reserves in patients undergoing major abdominal surgery”. In: *Therapeutics and Clinical Risk Management* 13 (Sept. 2017), pp. 1223–1231. ISSN: 1178203X. DOI: 10.2147/TCRM.S143809.

[169] Henry Humberto León-Ariza, Daniel Alfonso Botero-Rosas, and Aura Catalina Zea-Robles. “HEART RATE VARIABILITY AND BODY COMPOSITION AS VO2MAX DETERMINANTS”. In: *Revista Brasileira de Medicina do Esporte* 23.4 (May 2017), pp. 317–321. ISSN: 1517-8692. DOI: 10.1590/1517-869220172304152157.

[170] Catharina C. Grant et al. “A comparison between heart rate and heart rate variability as indicators of cardiac health and fitness”. In: *Frontiers in Physiology* 4 NOV (Nov. 2013), p. 65494. ISSN: 1664042X. DOI: 10.3389/FPHYS.2013.00337/BIBTEX.

[171] Vladan Vukomanovic et al. “Association between functional capacity and heart rate variability in patients with uncomplicated type 2 diabetes”. In: *Blood pressure* 28.3 (May 2019), pp. 184–190. ISSN: 1651-1999. DOI: 10.1080/08037051.2019.1586431.

[172] Foroohar Foroozan, Madhan Mohan, and Jian Shu Wu. “Robust Beat-To-Beat Detection Algorithm for Pulse Rate Variability Analysis from Wrist Photoplethysmography Signals”. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2018-April (Sept. 2018), pp. 2136–2140. ISSN: 15206149. DOI: 10.1109/ICASSP.2018.8462286.

[173] Ayca Aygun, Hassan Ghasemzadeh, and Roozbeh Jafari. “Robust interbeat interval and heart rate variability estimation method from various morphological features using wearable sensors”. In: *IEEE Journal of Biomedical and Health Informatics* 24.8 (Aug. 2020), pp. 2238–2250. ISSN: 21682208. DOI: 10.1109/JBHI.2019.2962627.

[174] Garmin. *Understanding HRV Status on Your Garmin Device*. Aug. 2023. URL: https://www.garmin.com/en-GB/garmin-technology/health-science/hrv-status/.

[175] FITBIT. *FITBIT INSPIRE 2*. 2023. URL: https://www.fitbit.com/global/uk/products/trackers/inspire2?utm_medium=shopping&utm_source=google&utm_campaign=UK_PF_ROAS&gclid=Cj0KCQjwtJKqBhCaARIsAN_yS_k81185toDqMAZ5aQV3Ja7jLdRXZ4mlRqEW--Zp0ofuZv3MmKHDKBcaAqsUEALw_wcB&gclsrc=aw.ds.

[176] Gari D. Clifford and Lionel Tarassenko. “Quantifying errors in spectral estimates of HRV due to beat replacement and resampling”. In: *IEEE transactions on bio-medical engineering* 52.4 (Apr. 2005), pp. 630–638. ISSN: 0018-9294. DOI: 10.1109/TBME.2005.844028.

[177] Shanhu Qiu et al. “Heart Rate Recovery and Risk of Cardiovascular Events and All-Cause Mortality: A Meta-Analysis of Prospective Cohort Studies”. In: *Journal of the American Heart Association* 6.5 (May 2017). ISSN: 20479980. DOI: 10.1161/JAHA.117.005505.

[178] Peter H. Charlton et al. “Breathing Rate Estimation From the Electrocardiogram and Photoplethysmogram: A Review”. In: *IEEE reviews in biomedical engineering* 11 (Jan. 2018), p. 2. ISSN: 19411189. DOI: 10.1109/RBME.2017.2763681.

[179] Haipeng Liu et al. “Recent development of respiratory rate measurement technologies”. In: *Physiological measurement* 40.7 (Aug. 2019). ISSN: 1361-6579. DOI: 10.1088/1361-6579/AB299E.

[180] Ruisheng Lei et al. “Estimation of Heart Rate and Respiratory Rate from PPG Signal Using Complementary Ensemble Empirical Mode Decomposition with both Independent Component Analysis and Non-Negative Matrix Factorization”. In: *Sensors (Basel, Switzerland)* 20.11 (May 2020), pp. 1–13. ISSN: 14248220. DOI: 10.3390/S20113238.

[181] Hamparsum Bozdogan. “Model selection and Akaike’s Information Criterion (AIC): The general theory and its analytical extensions”. In: *Psychometrika* 52.3 (Sept. 1987), pp. 345–370. ISSN: 00333123. DOI: 10.1007/BF02294361/METRICS.

[182] Patrick Schober and Lothar A. Schwarte. “Correlation coefficients: Appropriate use and interpretation”. In: *Anesthesia and Analgesia* 126.5 (May 2018), pp. 1763–1768. ISSN: 15267598. DOI: 10.1213/ANE.0000000000002864.

[183] James A. Nichols, Hsien W. Herbert Chan, and Matthew A.B. Baker. “Machine learning: applications of artificial intelligence to imaging and diagnosis”. In: *Biophysical Reviews* 11.1 (Feb. 2018), p. 111. ISSN: 18672469. DOI: 10.1007/S12551-018-0449-9.

[184] Maren E. Shipe et al. “Developing prediction models for clinical use using logistic regression: an overview”. In: *Journal of Thoracic Disease* 11.Suppl 4 (2019), S574. ISSN: 20776624. DOI: 10.21037/JTD.2019.01.25.

[185] Christel Rushing et al. “A leave-one-out cross-validation SAS macro for the identification of markers associated with survival”. In: *Computers in Biology and Medicine* 57 (Feb. 2015), pp. 123–129. ISSN: 0010-4825. DOI: 10.1016/J.COMPBIOMED.2014.11.015.

[186] Marco Altini et al. “Cardiorespiratory fitness estimation using wearable sensors: Laboratory and free-living analysis of context-specific submaximal heart rates”. In: *J Appl Physiol* 120 (2016), pp. 1082–1096. DOI: 10.1152/japplphysiol.00519.2015.-In.

[187] Yuankai Zhang et al. “Association of Smartwatch-Based Heart Rate and Physical Activity With Cardiorespiratory Fitness Measures in the Community: Cohort Study”. In: *J Med Internet Res 2024;26:e56676 https://www.jmir.org/2024/1/e56676* 26.1 (Aug. 2024), e56676. ISSN: 1438-8871. DOI: 10.2196/56676.

[188] Corinna Cortes, Vladimir Vapnik, and Lorenza Saitta. “Support-vector networks”. In: *Machine Learning 1995 20:3* 20.3 (Sept. 1995), pp. 273–297. ISSN: 1573-0565. DOI: 10.1007/BF00994018.

[189] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-August-2016 (Mar. 2016), pp. 785–794. DOI: 10.1145/2939672.2939785.

[190] Dermot Sheridan et al. “Estimating oxygen uptake in simulated team sports using machine learning models and wearable sensor data: A pilot study”. In: *PLOS ONE* 20.4 (Apr. 2025), e0319760. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0319760.

[191] Panagiotis Pintelas and Ioannis E. Livieris. “Special Issue on Ensemble Learning and Applications”. In: *Algorithms 2020, Vol. 13, Page 140* 13.6 (June 2020), p. 140. ISSN: 1999-4893. DOI: 10.3390/A13060140.

[192] Shaohua Wan and Hua Yang. “Comparison among methods of ensemble learning”. In: *Proceedings - 2013 International Symposium on Biometrics and Security Technologies, ISBAST 2013* (2013), pp. 286–290. DOI: 10.1109/ISBAST.2013.50.

[193] Mariana Belgiu and Lucian Drăgu. “Random forest in remote sensing: A review of applications and future directions”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 114 (Apr. 2016), pp. 24–31. ISSN: 0924-2716. DOI: 10.1016/J.ISPRSJPRS.2016.01.011.

[194] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *Nature 2015 521:7553* 521.7553 (May 2015), pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539.

[195] Sarah M Powell and Ann V Rowlands. “Intermonitor Variability of the RT3 Accelerometer during Typical Physical Activities”. In: *Medicine and Science in Sports and Exercise* 36.2 (May 2004), pp. 324–330. ISSN: 01959131. DOI: 10.1249/01.MSS.0000113743.68789.36.

[196] Gareth J. Williams et al. “Wearable technology and the cardiovascular system: the future of patient assessment”. In: *The Lancet Digital Health* 5.7 (July 2023), e467–e476. ISSN: 25897500. DOI: 10.1016/S2589-7500(23)00087-0.

[197] Mahsa Sadat Afzali Arani, Diego Elias Costa, and Emad Shihab. “Human Activity Recognition: A Comparative Study to Assess the Contribution Level of Accelerometer, ECG, and PPG Signals”. In: *Sensors 2021, Vol. 21, Page 6997* 21.21 (Oct. 2021), p. 6997. ISSN: 1424-8220. DOI: 10.3390/S21216997.

[198] Davide Morelli et al. “Profiling the propagation of error from PPG to HRV features in a wearable physiological-monitoring device”. In: *Healthcare Technology Letters* 5.2 (Apr. 2018), pp. 59–64. ISSN: 2053-3713. DOI: 10.1049/HTL.2017.0039.

[199] Christoph Hoog Antink et al. “Accuracy of heart rate variability estimated with reflective wrist-PPG in elderly vascular patients”. In: *Scientific Reports 2021 11:1* 11.1 (Apr. 2021), pp. 1–12. ISSN: 2045-2322. DOI: 10.1038/s41598-021-87489-0.

[200] Steven A. Lubitz et al. “Detection of Atrial Fibrillation in a Large Population Using Wearable Devices: The Fitbit Heart Study”. In: *Circulation* 146.19 (Nov. 2022), pp. 1415–1424. ISSN: 15244539. DOI: 10.1161/CIRCULATIONAHA.122.060291.

[201] Juan Pablo Martínez et al. “Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances”. In: (2022). DOI: 10.3390/s22041476.

[202] Nasim Talebi, Cory Hallam, and Gianluca Zanella. “The new wave of privacy concerns in the wearable devices era”. In: *PICMET 2016 - Portland International Conference on Management of Engineering and Technology: Technology Management For Social Innovation, Proceedings* (May 2017), pp. 3208–3214. DOI: 10.1109/PICMET.2016.7806826.

[203] Varda Mone and Fayazullaeva Shakhlo. “Health Data on the Go: Navigating Privacy Concerns with Wearable Technologies”. In: *Legal Information Management* 23.3 (May 2023), pp. 179–188. ISSN: 1472-6696. DOI: 10.1017/S1472669623000427.

[204] Lisa A. Simpson et al. “Clinicians’ perceptions of a potential wearable device for capturing upper limb activity post-stroke: a qualitative focus group study”. In: *Journal of NeuroEngineering and Rehabilitation* 18.1 (Dec. 2021), pp. 1–10. ISSN: 17430003. DOI: 10.1186/S12984-021-00927-Y/FIGURES/1.

[205] Foof and Drug Administration. *General Wellness: Policy for Low Risk Devices*. Sept. 2019. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/general-wellness-policy-low-risk-devices.

[206] Cheng Yang et al. “Cost, time savings and effectiveness of wearable devices for remote monitoring of patient rehabilitation after total knee arthroplasty: study protocol for a randomized controlled trial”. In: *Journal of Orthopaedic Surgery and Research* 18.1 (Dec. 2023). ISSN: 1749799X. DOI: 10.1186/S13018-023-03898-Z.

[207] Andrew Pritchard et al. “ARTP statement on cardiopulmonary exercise testing 2021”. In: *BMJ Open Respiratory Research* 8.1 (Nov. 2021), p. 1121. ISSN: 2052-4439. DOI: 10.1136/BMJRESP-2021-001121.

[208] Soon Bin Kwon et al. “Estimating maximal oxygen uptake from daily activity data measured by a watch-type fitness tracker: Cross-sectional study”. In: *JMIR mHealth and uHealth* 7.6 (2019). ISSN: 22915222. DOI: 10.2196/13327.

[209] *Schedule of Procedures*. Dec. 2022. URL: https://codes.bupa.co.uk/home.

[210] Candela Diaz-Canestro et al. “Differences in Cardiac Output and Aerobic Capacity Between Sexes Are Explained by Blood Volume and Oxygen Carrying Capacity”. In: *Frontiers in Physiology* 13 (Mar. 2022), p. 747903. ISSN: 1664042X. DOI: 10.3389/FPHYS.2022.747903/BIBTEX.

[211] Jon Magne Letnes, Bjarne M. Nes, and Ulrik Wisløff. “Age-related decline in peak oxygen uptake: Cross-sectional vs. longitudinal findings. A review”. In: *International Journal of Cardiology Cardiovascular Risk and Prevention* 16 (Mar. 2023), p. 200171. ISSN: 2772-4875. DOI: 10.1016/J.IJCRP.2023.200171.

[212] Jonas Van Der Donckt et al. “Mitigating data quality challenges in ambulatory wrist-worn wearable monitoring through analytical and practical approaches”. In: *Scientific Reports* — 14 (123), p. 17545. DOI: 10.1038/s41598-024-67767-3.

[213] Teresa L. Hart et al. “How many days of monitoring predict physical activity and sedentary behaviour in older adults?” In: *The International Journal of Behavioral Nutrition and Physical Activity* 8 (May 2011), p. 62. ISSN: 14795868. DOI: 10.1186/1479-5868-8-62.

[214] Dominique Makowski et al. “NeuroKit2: A Python toolbox for neurophysiological signal processing”. In: *Behavior Research Methods* 53.4 (Aug. 2021), pp. 1689–1696. ISSN: 15543528. DOI: 10.3758/S13428-020-01516-Y/TABLES/3.

[215] Gareth J. Williams et al. “Wearable technology and the cardiovascular system: the future of patient assessment”. In: *The Lancet Digital Health* 5.7 (July 2023), e467–e476. ISSN: 25897500. DOI: 10.1016/S2589-7500(23)00087-0/ASSET/996AB19D-2ECD-4DB0-8B8B-8FBDD0773B30/MAIN.ASSETS/GR5.JPG.

[216] Ikaro Silva, George Moody, and Leo Anthony Celi. *Improving the Quality of ECGs Collected using Mobile Phones: The PhysioNet/Computing in Cardiology Challenge 2011*. 2011.

[217] Fotsing Kuetche et al. “Simple, efficient, and generalized ECG signal quality assessment method for telemedicine applications”. In: *Informatics in Medicine Unlocked* 42 (Jan. 2023), p. 101375. ISSN: 2352-9148. DOI: 10.1016/J.IMU.2023.101375.

[218] Saifur Rahman et al. “Robustness of electrocardiogram signal quality indices”. In: *Journal of the Royal Society Interface* 19.189 (2022). ISSN: 17425662. DOI: 10.1098/RSIF.2022.0012.

[219] Chengyu Liu and Jianqing Li. “Feature engineering and computational intelligence in ECG monitoring”. In: *Feature Engineering and Computational Intelligence in ECG Monitoring* (Jan. 2020), pp. 1–268. DOI: 10.1007/978-981-15-3824-7/COVER.

[220] Beatrice Zanchi et al. “Synthetic ECG signals generation: A scoping review”. In: *Computers in Biology and Medicine* 184 (Jan. 2025), p. 109453. ISSN: 0010-4825. DOI: 10.1016/J.COMPBIOMED.2024.109453.

[221] Kirina Van Der Bijl, Mohamed Elgendi, and Carlo Menon. “Automatic ECG Quality Assessment Techniques: A Systematic Review”. In: (2022). DOI: 10.3390/diagnostics12112578.

[222] Katri Karhinoja et al. *Flexible framework for generating synthetic electrocardiograms and photoplethysmograms*. Aug. 2024. URL: https://arxiv.org/abs/2408.16291v1.

[223] Premysl Jiruska et al. “Reference noise method of removing powerline noise from recorded signals”. In: *Journal of Neuroscience Methods* 184.1 (Oct. 2009), pp. 110–114. ISSN: 0165-0270. DOI: 10.1016/J.JNEUMETH.2009.07.003.

[224] Christina Orphanidou et al. “Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring”. In: *IEEE journal of biomedical and health informatics* 19.3 (May 2015), pp. 832–838. ISSN: 2168-2208. DOI: 10.1109/JBHI.2014.2338351.

[225] Peter Charlton. *Signal Processing and Learning for Wearables*. 2023. URL: https://peterhcharlton.github.io/bsp-book/tutorial/notebooks/signal-quality-assessment.html.

[226] Zhidong Zhao and Yefei Zhang. “SQI quality evaluation mechanism of single-lead ECG signal based on simple heuristic fusion and fuzzy comprehensive evaluation”. In: *Frontiers in Physiology* 9.JUN (May 2018), p. 361542. ISSN: 1664042X. DOI: 10.3389/FPHYS.2018.00727/BIBTEX.

[227] Linus Kramer, Carlo Menon, and Mohamed Elgendi. “ECGAssess: A Python-Based Toolbox to Assess ECG Lead Signal Quality”. In: *Frontiers in Digital Health* 4 (May 2022), p. 847555. ISSN: 2673253X. DOI: 10.3389/FDGTH.2022.847555/BIBTEX.

[228] Mohamed Elgendi, Kirina van der Bijl, and Carlo Menon. “An Open-Source Graphical User Interface-Embedded Automated Electrocardiogram Quality Assessment: A Balanced Class Representation Approach”. In: *Diagnostics (Basel, Switzerland)* 13.22 (Nov. 2023). ISSN: 2075-4418. DOI: 10.3390/DIAGNOSTICS13223479.

[229] Jesus Lazaro et al. “Wearable Armband Device for Daily Life Electrocardiogram Monitoring”. In: *IEEE Transactions on Biomedical Engineering* 67.12 (Dec. 2020), pp. 3464–3473. ISSN: 15582531. DOI: 10.1109/TBME.2020.2987759.

[230] Ping Lu et al. “Improving Classification of Tetanus Severity for Patients in Low-Middle Income Countries Wearing ECG Sensors by Using a CNN-Transformer Network”. In: *IEEE transactions on bio-medical engineering* 70.4 (Apr. 2023), pp. 1340–1350. ISSN: 1558-2531. DOI: 10.1109/TBME.2022.3216383.

[231] Alexander Neshitov et al. “Estimation of cardiorespiratory fitness using heart rate and step count data”. In: *Scientific Reports 2023 13:1* 13.1 (May 2023), pp. 1–13. ISSN: 2045-2322. DOI: 10.1038/s41598-023-43024-x.

[232] Maria Cecília Moraes Frade et al. “Toward characterizing cardiovascular fitness using machine learning based on unobtrusive data”. In: *PLOS ONE* 18.3 (May 2023). ISSN: 19326203. DOI: 10.1371/JOURNAL.PONE.0282398.

[233] Atiqa Ashfaq, Neil Cronin, and Philipp Müller. “Recent advances in machine learning for maximal oxygen uptake (VO2 max) prediction: A review”. In: *Informatics in Medicine Unlocked* 28 (Jan. 2022), p. 100863. ISSN: 2352-9148. DOI: 10.1016/J.IMU.2022.100863.

[234] Polona Caserman et al. “Assessing the Accuracy of Smartwatch-Based Estimation of Maximum Oxygen Uptake Using the Apple Watch Series 7: Validation Study”. In: *JMIR Biomedical Engineering* 9 (2024), e59459. ISSN: 25613278. DOI: 10.2196/59459.

[235] Marcin Straczkiewicz et al. “Open-Source, Step-Counting Algorithm for Smartphone Data Collected in Clinical and Nonclinical Settings: Algorithm Development and Validation Study”. In: *JMIR cancer* 9.1 (May 2023). ISSN: 2369-1999. DOI: 10.2196/47646.

[236] Jim Luckhurst, Cara Hughes, and Benjamin Shelley. “Classifying physical activity levels using Mean Amplitude Deviation in adults using a chest worn accelerometer: validation of the Vivalink ECG Patch”. In: *BMC sports science, medicine & rehabilitation* 16.1 (Dec. 2024). ISSN: 2052-1847. DOI: 10.1186/S13102-024-00991-6.

[237] Franziska Beck et al. “Determination of cut-off points for the Move4 accelerometer in children aged 8–13 years”. In: *BMC Sports Science, Medicine and Rehabilitation* 15.1 (Dec. 2023), p. 163. ISSN: 20521847. DOI: 10.1186/S13102-023-00775-4.

[238] Rosemary Walmsley et al. “Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease”. In: *British Journal of Sports Medicine* 56.18 (Sept. 2022), pp. 1008–1017. ISSN: 0306-3674. DOI: 10.1136/BJSPORTS-2021-104050.

[239] Catrine Tudor-Locke et al. “How fast is fast enough? Walking cadence (steps/min) as a practical estimate of intensity in adults: a narrative review”. In: *British Journal of Sports Medicine* 52.12 (June 2018), pp. 776–788. ISSN: 0306-3674. DOI: 10.1136/BJSPORTS-2017-097628.

[240] British Heart Foundation. *Your Heart Rate*. July 2022. URL: https://www.bhf.org.uk/informationsupport/how-a-healthy-heart-works/your-heart-rate.

[241] Hirofumi Tanaka, K D Monahan, and D R Seals. “Age-predicted maximal heart rate revisited”. In: *Journal of the American College of Cardiology* 37.1 (2001), pp. 153–156.

[242] Fred Shaffer and J. P. Ginsberg. “An Overview of Heart Rate Variability Metrics and Norms”. In: *Frontiers in public health* 5 (Sept. 2017). ISSN: 2296-2565. DOI: 10.3389/FPUBH.2017.00258.

[243] Davide Morelli et al. “SDNN24 Estimation from Semi-Continuous HR Measures”. In: *Sensors 2021, Vol. 21, Page 1463* 21.4 (May 2021), p. 1463. ISSN: 1424-8220. DOI: 10.3390/S21041463.

[244] Marc N. Jarczok et al. “Heart rate variability in the prediction of mortality: A systematic review and meta-analysis of healthy and patient populations”. In: *Neuroscience & Biobehavioral Reviews* 143 (Dec. 2022), p. 104907. ISSN: 0149-7634. DOI: 10.1016/J.NEUBIOREV.2022.104907.

[245] Hisako Tsuji et al. “Impact of reduced heart rate variability on risk for cardiac events: The Framingham Heart Study”. In: *Circulation* 94.11 (1996), pp. 2850–2855. ISSN: 00097322. DOI: 10.1161/01.CIR.94.11.2850,.

[246] Yasuhiko Kubota et al. “Heart Rate Variability and Lifetime Risk of Cardiovascular Disease: the Atherosclerosis Risk in Communities Study”. In: *Annals of epidemiology* 27.10 (Oct. 2017), p. 619. ISSN: 18732585. DOI: 10.1016/J.ANNEPIDEM.2017.08.024.

[247] James S. Mulcahy et al. “Heart rate variability as a biomarker in health and affective disorders: A perspective on neuroimaging studies”. In: *NeuroImage* 202 (Nov. 2019). ISSN: 10959572. DOI: 10.1016/j.neuroimage.2019.116072.

[248] Giampaolo Perna et al. “Heart rate variability: Can it serve as a marker of mental health resilience?: Special Section on “Translational and Neuroscience Studies in Affective Disorders” Section Editor, Maria Nobile MD, PhD”. In: *Journal of Affective Disorders* 263 (Feb. 2020), pp. 754–761. ISSN: 15732517. DOI: 10.1016/j.jad.2019.10.017.

[249] Marek Malik et al. “Heart rate variability: Standards of measurement, physiological interpretation, and clinical use”. In: *European Heart Journal* 17.3 (May 1996), pp. 354–381.

[250] Aviv A. Rosenberg et al. “Signatures of the autonomic nervous system and the heart’s pacemaker cells in canine electrocardiograms and their applications to humans”. In: *Scientific Reports 2020 10:1* 10.1 (June 2020), pp. 1–15. ISSN: 2045-2322. DOI: 10.1038/s41598-020-66709-z.

[251] Lü Fei et al. “Short- and long-term assessment of heart rate variability for risk stratification after acute myocardial infarction”. In: *The American journal of cardiology* 77.9 (May 1996), pp. 681–684. ISSN: 0002-9149. DOI: 10.1016/S0002-9149(97)89199-0.

[252] Robert E. Kleiger et al. “Decreased heart rate variability and its association with increased mortality after acute myocardial infarction”. In: *The American Journal of Cardiology* 59.4 (Feb. 1987), pp. 256–262. ISSN: 00029149. DOI: 10.1016/0002-9149(87)90795-8,.

[253] L. Gao et al. “Value of DC and DRs in prediction of cardiovascular events in acute myocardial infarction patients”. In: *Zhonghua yi xue za zhi* 96.19 (May 2016), pp. 1519–1522. ISSN: 03762491. DOI: 10.3760/CMA.J.ISSN.0376-2491.2016.19.012,.

[254] Maciej Karcz et al. “Prognostic significance of heart rate variability in dilated cardiomyopathy”. In: *International Journal of Cardiology* 87.1 (Jan. 2003), pp. 75–81. ISSN: 01675273. DOI: 10.1016/S0167-5273(02)00207-3.

[255] Stefanie Hillebrand et al. “Heart rate variability and first cardiovascular event in populations without known cardiovascular disease: Meta-analysis and dose-response meta-regression”. In: *Europace* 15.5 (May 2013), pp. 742–749. ISSN: 10995129. DOI: 10.1093/EUROPACE/EUS341,.

[256] Wollner Materko et al. “Maximum Oxygen Uptake Prediction Model Basedon Heart Rate Variability Parameters for YoungHealthy Adult Males at Rest”. In: *Open Access Biostatistics and Informatics* 2 (2018), pp. 1–7.

[257] Saimi Zaki et al. “View of Association between heart rate variability and cardiorespiratory fitness in individuals with type 2 diabetes mellitus”. In: *Journal of Human Sport and Exercise* (2024).

[258] David Herzig et al. “Reproducibility of heart rate variability is parameter and sleep stage dependent”. In: *Frontiers in Physiology* 8.JAN (Jan. 2018), p. 302881. ISSN: 1664042X. DOI: 10.3389/FPHYS.2017.01100/BIBTEX.

[259] Benjamin Israel et al. “Short-Term Stability of Sleep and Heart Rate Variability in Good Sleepers and Patients with Insomnia: For Some Measures, One Night is Enough”. In: *Sleep* 35.9 (May 2012), pp. 1285–1291. ISSN: 0161-8105. DOI: 10.5665/SLEEP.2088.

[260] Lina Zhao et al. “Influence of Ectopic Beats on Heart Rate Variability Analysis”. In: *Entropy 2021, Vol. 23, Page 648* 23.6 (May 2021), p. 648. ISSN: 1099-4300. DOI: 10.3390/E23060648.

[261] Hope Davis-Wilson et al. “Effects of Missing Data on Heart Rate Variability Measured From A Smartwatch: Exploratory Observational Study.” In: *JMIR formative research* 9.1 (Feb. 2025), e53645. ISSN: 2561-326X. DOI: 10.2196/53645.

[262] Menelaos Pavlou et al. “How to develop a more accurate risk prediction model when there are few events”. In: *The BMJ* 351 (Aug. 2015), h3868. ISSN: 17561833. DOI: 10.1136/BMJ.H3868.

[263] Ulrike Grömping. “Variable importance in regression models”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 7.2 (May 2015), pp. 137–152. ISSN: 1939-0068. DOI: 10.1002/WICS.1346.

[264] Scott M Lundberg and Su In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems* 2017-December (May 2017), pp. 4766–4775. ISSN: 10495258.

[265] Mucahid Yilmaz, Hidayet Kayancicek, and Yusuf Cekici. “Heart rate variability: Highlights from hidden signals”. In: *Journal of Integrative Cardiology* 4.5 (2018). DOI: 10.15761/JIC.1000258.

[266] Ben Ewald, M McEvoy, and J Attia. “Pedometer counts superior to physical activity scale for identifying health markers in older adults”. In: *British Journal of Sports Medicine* 44.10 (May 2010), pp. 756–761. ISSN: 0306-3674. DOI: 10.1136/BJSM.2008.048827.

[267] Roy J Shephard. “Maximal oxygen intake and independence in old age”. In: *British journal of sports medicine* 43.5 (May 2009), pp. 342–346. ISSN: 1473-0480. DOI: 10.1136/BJSM.2007.044800.

[268] Udit Satija, Barathram Ramkumar, and M. Sabarimalai Manikandan. “A Review of Signal Processing Techniques for Electrocardiogram Signal Quality Assessment”. In: *IEEE reviews in biomedical engineering* 11 (Feb. 2018), pp. 36–52. ISSN: 1941-1189. DOI: 10.1109/RBME.2018.2810957.

[269] Fotsing Kuetche et al. “Signal quality indices evaluation for robust ECG signal quality assessment systems”. In: *Biomedical Physics and Engineering Express* 9.5 (Sept. 2023). ISSN: 20571976. DOI: 10.1088/2057-1976/ACE9E0,.

[270] Larisa G. Tereshchenko and Mark E. Josephson. “Frequency Content and Characteristics of Ventricular Conduction”. In: *Journal of electrocardiology* 48.6 (2015), p. 933. ISSN: 15328430. DOI: 10.1016/J.JELECTROCARD.2015.08.034.

[271] Sean Bae et al. “Prospective validation of smartphone-based heart rate and respiratory rate measurement algorithms”. In: *Communications Medicine 2022 2:1* 2.1 (Apr. 2022), pp. 1–10. ISSN: 2730-664X. DOI: 10.1038/s43856-022-00102-x.

[272] George B. Moody and R. G. Mark. “The impact of the MIT-BIH arrhythmia database”. In: *IEEE Engineering in Medicine and Biology Magazine* 20.3 (2001), pp. 45–50. ISSN: 07395175. DOI: 10.1109/51.932724,.

[273] George Moody, B. Moody, and I Silva. *Robust Detection of Heart Beats in Multimodal Data: The PhysioNet/Computing in Cardiology Challenge 2014 v1.0.0*. 2014. URL: https://physionet.org/content/challenge-2014/1.0.0/.

[274] Heba Khamis et al. “TELE ECG Database: 250 telehealth ECG records (collected using dry metal electrodes) with annotated QRS and artifact masks, and MATLAB code for the UNSW artifact detection and UNSW QRS detection algorithms”. In: (2016).

[275] Md Moklesur Rahman et al. “A Systematic Survey of Data Augmentation of ECG Signals for AI Applications”. In: *Sensors 2023, Vol. 23, Page 5237* 23.11 (May 2023), p. 5237. ISSN: 1424-8220. DOI: 10.3390/S23115237.

[276] Awni Y. Hannun et al. “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network”. In: *Nature Medicine* 25.1 (Jan. 2019), pp. 65–69. ISSN: 1546170X. DOI: 10.1038/S41591-018-0268-3,.

[277] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-December (Dec. 2015), pp. 770–778. ISSN: 10636919. DOI: 10.1109/CVPR.2016.90.

[278] Enbiao Jing et al. “ECG Heartbeat Classification Based on an Improved ResNet-18 Model”. In: *Computational and Mathematical Methods in Medicine* 2021.1 (Jan. 2021), p. 6649970. ISSN: 1748-6718. DOI: 10.1155/2021/6649970.

[279] Zhibin Zhao et al. “Analysis of an adaptive lead weighted ResNet for multi-class classification of 12-lead ECGs”. In: *Physiological Measurement* 43.3 (Apr. 2022), p. 034001. ISSN: 0967-3334. DOI: 10.1088/1361-6579/AC5B4A.

[280] Qiao Li, R. G. Mark, and G. D. Clifford. “Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter”. In: *Physiological Measurement* 29.1 (Dec. 2007), p. 15. ISSN: 0967-3334. DOI: 10.1088/0967-3334/29/1/002.

[281] Nicholas M. Beltz et al. “Graded Exercise Testing Protocols for the Determination of VO2max: Historical Perspectives, Progress, and Future Considerations”. In: *Journal of Sports Medicine* 2016 (2016), p. 3968393. ISSN: 2356-7651. DOI: 10.1155/2016/3968393.

[282] Simon Nolte, Robert Rein, and Oliver Jan Quittmann. “Data Processing Strategies to Determine Maximum Oxygen Uptake: A Systematic Scoping Review and Experimental Comparison with Guidelines for Reporting”. In: *Sports Medicine* 53.12 (Dec. 2023), pp. 2463–2475. ISSN: 11792035. DOI: 10.1007/S40279-023-01903-3/TABLES/5.

[283] Vickers RR. *Measurement Error in Maximal Oxygen Uptake Tests — Enhanced Reader*. May 2003.

[284] Nisha Pradhan et al. “Attitudes about use of preoperative risk assessment tools: a survey of surgeons and surgical residents in an academic health system”. In: *Patient Safety in Surgery* 16.1 (May 2022), pp. 1–9. ISSN: 17549493. DOI: 10.1186/S13037-022-00320-1/TABLES/4.

[285] Finlay A. McAlister et al. “A comparison of four risk models for the prediction of cardiovascular complications in patients with a history of atrial fibrillation undergoing non-cardiac surgery”. In: *Anaesthesia* 75.1 (Jan. 2020), pp. 27–36. ISSN: 1365-2044. DOI: 10.1111/ANAE.14777.

[286] Suwen Lin et al. “Filling missing values on wearable-sensory time series data”. In: *Proceedings of the 2020 SIAM International Conference on Data Mining* (2020), pp. 46–54. DOI: 10.1137/1.9781611976236.6.

[287] Shweta Chakrabarti et al. “Binned Data Provide Better Imputation of Missing Time Series Data from Wearables”. In: *Sensors* 23.3 (Feb. 2023), p. 1454. ISSN: 14248220. DOI: 10.3390/S23031454/S1.

[288] Queen Mary et al. “A Survey Of Feature Selection And Feature Extraction Techniques In Machine Learning,SAI,2014 Enhanced Framework for recognizing indoor daily life activities View project Shamila Nasreen A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning”. In: (2014).

[289] Julian Leube et al. “Reconstruction of the respiratory signal through ECG and wrist accelerometer data”. In: *Scientific Reports 2020 10:1* 10.1 (Sept. 2020), pp. 1–12. ISSN: 2045-2322. DOI: 10.1038/s41598-020-71539-0.

[290] Farrokh Mohammadzadeh, Chang S. Nam, and Edgar Lobaton. “Prediction of Physiological Response over Varying Forecast Lengths with a Wearable Health Monitoring Platform”. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* 2018-July (Oct. 2018), pp. 437–440. ISSN: 1557170X. DOI: 10.1109/EMBC.2018.8512276.

[291] Ian Smith et al. “Respiratory rate measurement: a comparison of methods”. In: *https://doi.org/10.12968/bjha.2011.5.1.18* 5.1 (Aug. 2013), pp. 18–23. ISSN: 1753-1586. DOI: 10.12968/BJHA.2011.5.1.18.

[292] Bouchaib Zazoum, Khalid Mujasam Batoo, and Muhammad Azhar Ali Khan. “Recent Advances in Flexible Sensors and Their Applications”. In: *Sensors 2022, Vol. 22, Page 4653* 22.12 (June 2022), p. 4653. ISSN: 1424-8220. DOI: 10.3390/S22124653.

[293] Ying Huang et al. “Resistive pressure sensor for high-sensitivity e-skin based on porous sponge dip-coated CB/MWCNTs/SR conductive composites”. In: *Materials Research Express* 5.6 (June 2018), p. 065701. ISSN: 2053-1591. DOI: 10.1088/2053-1591/AAC8C0.

[294] Tibor Stracina et al. “Golden Standard or Obsolete Method? Review of ECG Applications in Clinical and Experimental Context”. In: *Frontiers in Physiology* 13 (Apr. 2022), p. 867033. ISSN: 1664042X. DOI: 10.3389/FPHYS.2022.867033.

Appendix A

Supplementary Material for Chapter 2

Appendix Table A.1. Summary of reviewed studies

This table provides a breakdown of the samples used in each of the key research papers included in this review for analysis. Within each column, the sample size is reported alongside participant demographics including: Average age, gender split, ethnicity and health outcomes.

Appendix Table A.2. Overview of feature extraction methods

This table provides a comprehensive breakdown of the most relevant papers included in this review. The columns present the data extracted from each paper including the Outcome variable, sensor modality, pre-processing methods, features extracted and the model of analysis.

Supplementary Table S1.

Title	Sample size	Average age	Gender split (% male)	Ethnicity	Health condition
Predicting Outcomes in Patients Undergoing Pancreatectomy Using Wearable Technology and Machine Learning: Prospective Cohort Study	48	63.2	40%	White: 95% Non-white: 5%	Patients undergoing pancreatectomy
Predicting Post-Operative Complications with Wearables: A Case Study with Patients Undergoing Pancreatic Surgery	61	64.4	69%	White/Caucasian: 92% Black or African American: 0.5% Unknown: 3%	Patients undergoing pancreatectomy
Objectively measured preoperative physical activity is associated with time to functional recovery after hepato-pancreato-biliary cancer surgery: a pilot study	31	66	58%	N/A	Hepato-pancreato-biliary cancer
Preoperative physical activity levels and postoperative pulmonary complications post-esophagectomy	37	61	78%	N/A	Esophageal cancer
Feasibility and patient's experiences of perioperative telemonitoring in major abdominal surgery: an observational pilot study	42	68	70%	N/A	Patients undergoing major abdominal surgery
Wearable Technology in the Perioperative Period: Predicting Risk of Postoperative Complications in Patients Undergoing Elective Colorectal Surgery	99	55	48.4%	White: 89%	Patients undergoing elective colorectal surgery
Fitbit Data to Assess Functional Capacity in Patients Before Elective Surgery: Pilot Prospective Observational Study	31	76	29%	N/A	Patients under consideration for major non-cardiac surgeries
Wearable Health Technology for Preoperative Risk Assessment in	31	76	29%	N/A	Patients under consideration for major non-cardiac surgeries

Supplementary Table 1.

Elderly Patients: The WELCOME Study					
Wearable technology and the association of perioperative activity level with 30-day readmission among patients undergoing major colorectal surgery	94	54	48%	White: 89%	Patients undergoing major elective colorectal surgery
Wireless Monitoring Program of Patient-Centered Outcomes and Recovery Before and After Major Abdominal Cancer Surgery	20	55	25%	White: 65%	Patients scheduled to undergo curative resection for hepatobiliary and GI cancers
Predicting post-discharge cancer surgery complications via telemonitoring of patient-reported outcomes and patient-generated health data	52	N/A	N/A	N/A	Patients with Gastrointestinal or Lung cancer
The association between low pre-operative step count and adverse post-operative outcomes in older patients undergoing colorectal cancer surgery	85	76	49%	White: 92% Maori: 3.5% Other: 4.7%	Patients undergoing surgery for major colorectal cancer surgery
How Many Steps Per Day are Necessary to Prevent Postoperative Complications Following Hepato-Pancreato-Biliary Surgeries for Malignancy?	78	71	68%	N/A	Patients scheduled to undergo open abdominal surgeries for HPB malignancies
Preoperative Physical Activity Predicts Surgical Outcomes Following Lung Cancer Resection	78	70	45%	N/A	Patients undergoing anatomical Lung resection
Modeling Biobehavioral Rhythms with Passive Sensing in the Wild: A Case Study to Predict Readmission Risk after Pancreatic Surgery	53	65	47%	White: 94%	Patients undergoing surgery for pancreatic cancer or benign conditions (e.g. pancreatic cysts).

Supplementary Table 1.

Value of the average basal daily walked distance measured using a pedometer to predict maximum oxygen consumption per minute in patients undergoing lung resection	38	63	79%	N/A	Patients referred to major lung resection for lung cancer
Prediction of Physiological Response over Varying Forecast Lengths with a Wearable Health Monitoring Platform	N/A	N/A	N/A	N/A	N/A
Learning Individualized Cardiovascular Responses from Large-scale Wearable Sensors Data	80,137	31	18%	N/A	N/A
Self-supervised transfer learning of physiological representations from free-living wearable data	2,100	N/A	N/A	N/A	N/A
Wearable sensors enable personalized predictions of clinical laboratory measurements	54	57	44%	European: 74% Asian: 15% African American:7% Hispanic: 4%	N/A
Turning silver into Gold:Domain adaptation with noise labels for wearable cardio-respiratory fitness prediction	12,425	N/A	N/A	N/A	N/A
Cardiorespiratory fitness estimation in free-living using wearable sensors	46	25	84%	N/A	N/A
Cardiorespiratory fitness estimation using wearable sensors: Laboratory andfree-living analysis of context-specific submaximal heart rates	51	25	88%	N/A	N/A
Prediction of oxygen uptake dynamics by machine learning	16	27	100%	N/A	N/A

Supplementary Table 1.

analysis of wearable sensors during activities of daily living					
Longitudinal cardio-respiratory fitness prediction through wearables in free-living environments	11,059	N/A	48%	N/A	N/A
The association of pre-operative home accelerometry with cardiopulmonary exercise variables	48	71	88%	N/A	Patients attending the pre-operative cardiopulmonary exercise clinic
Can wearable technology be used to approximate cardiopulmonary exercise testing metrics?	49	65	65%	N/A	Patients scheduled for major elective intra-abdominal surgery

Supplementary Table S1. This table provides a breakdown of the samples used in each of the key research papers included in this review for analysis. Within each column, the sample size is reported alongside participant demographics including: Average age, gender split, ethnicity and health outcomes.

Supplementary Table S2.

Title	Authors	Year	Outcome	Sensor & Sensor Modality	Pre-processing Techniques	Features Extracted	Analysis Models
Predicting Outcomes in Patients Undergoing Pancreatectomy Using Wearable Technology and Machine Learning: Prospective Cohort Study	Cos et al.	2021	Complications post pancreatectomy	Fitbit Inspire HR: - PPG - Accelerometer	Detrended Fluctuation Analysis (DFA) 2 level imputation method for missing data	Daily Step Count Heart rate Sleep time-series	RF GBT KNN SVM Logistic Regression
Predicting Post-Operative Complications with Wearables: A Case Study with Patients Undergoing Pancreatic Surgery	Zhang et al.	2022	Complications post pancreatectomy	Fitbit Inspire HR: - PPG - Accelerometer	Daily feature extraction threshold > 8hours. 2 level imputation method for missing data (Figure 7): Short term imputation (<10mins) Robust singular spectrum analysis to impute missing daily features	Daily Feature extraction: Step Sleep HR time series High level feature extraction: Singular spectrum analysis	RF XGBoost KNN SVM Lasso Regressions Ridge Regression
Objectively measured preoperative physical activity is associated with time to functional recovery after hepato-pancreato-biliary cancer surgery: a pilot study	Mylius et al.	2021	Time to functional recovery post hepato-pancreato-biliary cancer surgery	Actigraph wGT3X- BT+: - Accelerometer	Extraction Threshold of 6 days+ of data	Time spent in MVPA: daily median of total accumulated minutes, daily median minutes accumulate in >10 minute bouts	Univariate and multivariate robust regression
Preoperative physical activity levels and postoperative pulmonary complications post-esophagectomy	Feeney et al.	2011	Postoperative pulmonary complications post-esophagectomy	RT3 Accelerometer: - Accelerometer	N/A	Physical activity intensities: inactive, light, moderate and vigorous	Independent t-test
Feasibility and patient's experiences of perioperative telemonitoring in major abdominal surgery: an observational pilot study	Haveman et al.	2022	Compliance and satisfaction with wearable sensor	Everion Biosensor: - PPG - Accelerometer	N/A	N/A	N/A
Wearable Technology in the Perioperative Period: Predicting Risk of Postoperative Complications in Patients Undergoing Elective Colorectal Surgery	Hedrick et al.	2020	Postoperative complications	Fitbit charge 2: - PPG - Accelerometer	N/A	Daily steps and stratification into two groups (active:>5000 dailys steps, inactive<5000) Daily HR	Chi-squared Wilcoxon rank-sum Kruskal-Wallis Multivariable regression models

Supplementary Table 2.

Fitbit Data to Assess Functional Capacity in Patients Before Elective Surgery: Pilot Prospective Observational Study	Angelucci et al.	2023	6MWT	Fitbit Inspire 2: - PPG - Accelerometer	Missing PPG data imputed with HR data from hospital visit.	HR RHR (24hr average) HR Zones during exercise Daily steps Distance walked Physical activity intensities HROS NET-F	Pearson correlation Wilcoxon Rank Sum Test
Wearable Health Technology for Preoperative Risk Assessment in Elderly Patients: The WELCOME Study	Greco et al.	2023	6MWT Preoperative scales	Fitbit Inspire 2: - PPG - Accelerometer	N/A	Average daily steps VO2max (as processed by the Fitbit device) HR data activity intensity energy expenditure calories.	Correlation analyses
Wearable technology and the association of perioperative activity level with 30-day readmission among patients undergoing major colorectal surgery	Kane et al.	2022	30-day readmission	Fitbit Charge 2: - PPG - Accelerometer	N/A	Daily Step count Average HR	Chi-squeared/Fishers exact test (categorical) Wilcoxon Rank Sum/Kruskal Wallis (continuous)
Wireless Monitoring Program of Patient-Centered Outcomes and Recovery Before and After Major Abdominal Cancer Surgery	Sun et al.	2017	Adherence with WS Satisfaction with monitoring	Garmin Vivofit 2: - Accelerometer	N/A	Patients Daily Steps	Correlation
Predicting post-discharge cancer surgery complications via telemonitoring of patient-reported outcomes and patient-generated health data	Rossi et al.	2021	Complications up to 30 days post discharge	Garmin Vivofit: - Accelerometer	N/A	Daily steps: maximum, minimum, SD, medium, slope and intercept of linear interpolation, differences compared to baseline	Logistic Regression
The association between low pre-operative step count and adverse post-operative outcomes in older patients undergoing colorectal cancer surgery	Richards et al.	2020	Length of hospital stay Rate of postoperative complications Mortality	Garmin Vivofit 3: - Accelerometer	N/A	Patients stratified into two groups from step count: <2500:low, >2500:normal	Univariate variables: Chi-sqaured and Kruskal-Wallis Negative Binomial Regressions Multivariable Logistic Regression

Supplementary Table 2.

How Many Steps Per Day are Necessary to Prevent Postoperative Complications Following Hepato-Pancreato-Biliary Surgeries for Malignancy?	Nakajima et al.	2020	Rate of major complications with Clavien-Dindo complications Rate of infectious complications Length of hospital stay	Kenx LifecoderGZ: - Accelerometer	N/A	Patients stratified into two groups from step count: <5000:poor, ≥ 5000:good	Chi-sqaured and Fisher's exact test and Mann-Whitney *U* test Spearman's rank correlation coefficient Multivariate logistic regression
Preoperative Physical Activity Predicts Surgical Outcomes Following Lung Cancer Resection	Billé et al.	2021	Respiratory and cardiac complications 30-day readmission rate	3D Trisport: - Accelerometer	N/A	Patients split into 4 groups based on daily step count based on median and 25% quartiles (not pre-defined quarters).	Chi-squared or 2-tailed T-tests
Modeling Biobehavioral Rhythms with Passive Sensing in the Wild: A Case Study to Predict Readmission Risk after Pancreatic Surgery	Doryab at el.	2019	Re-admission within 90-days of discharge.	Fitbit Charge 2: - PPG - Accelerometer		From step count: Number, length and number of steps in active bouts Number and length of sedentary bouts From HR: The minimum, maximum and mean of positive, negative and absolute change in HR Detection of rhythmicity by building individual's cosinors using data from each patient before build population level cosinors from readmitted vs non-readmitted patients. Visual features were also extracted from autocorrelation and periodograms	RF Logistic Regression SVM Bayesian Network Boosted Logistic Regression.
Value of the average basal daily walked distance measured using a pedometer to predict maximum oxygen consumption per minute in patients undergoing lung resection	Novoa et al.	2010	VO_{2Max}	OMROM walking style pedometer PRO: - Accelerometer	N/A	Daily steps Daily aerobic steps (aerobic steps are calculated after 10 mins of walking at >60steps per min) Daily time spent in aerobic activity (minutes) Daily distance measured (km)	Linear Regression Models Plotting of correlation index between models using Bland and Altman method.

Supplementary Table 2.

Prediction of Physiological Response over Varying Forecast Lengths with a Wearable Health Monitoring Platform	Mohammadzadeh et al.	2018	Breathing rate	Bioharness Zephyr: - ECG - Accelerometer Empatica E4: - PPG - Accelerometer	N/A	Respiratory rate HR HRV	SVM
Learning Individualized Cardiovascular Responses from Large-scale Wearable Sensors Data	Hallgrimson et al.	2018	Cardiovascular response Age BMI*	Achievement reward platform (Fitbit or Apple watch): - Accelerometer - PPG	Minute level from step count and heart rate was scaled to measurements between (0,1). Missing data was imputed as mean HR of activity at waking hours.	HR Step count Two sleep stages: Asleep or restless asleep	HR autoencoder that is trained on physical activity and sleep stages. The signature encoder learns the participants 'signature' from the HR responses to activity and the decoder employs the learned signature predict HR from PA. XGBoost models were used to compare performance against the encoder.
Self-supervised transfer learning of physiological representations from free-living wearable data	Spathis et al.	2021	HR Response VO_2 Max BMI Resting HR	Actiheart Chest ECG wearable monitor (2-lead) - ECG Wrist device triaxial accelerometer - Accelerometer	Extraction threshold of >72 hours for inclusion. Magnitude of acceleration from accelerometer data was calculated through Euclidean Norm Minus One and high passed filtered vector magnitude. Both accelerometer and HR data was filtered to a time resolution of one sample per 15 seconds. Temporal features were encoded into timestamps using cyclical temporal features.	HR Acceleration	Proposal of a multimodal self-supervised model for feature extraction from wearable data. The 'Step2Heart' model receives high-dimensional activity inputs to predict HR response. It stacks CNN and RNN layers where the CNN learns spacial features and the RNN learns temporal features of the data. This is compared against several models: Convolutional autoencoder XGBoost

Supplementary Table 2.

Wearable sensors enable personalized predictions of clinical laboratory measurements	Dunn et al.	2021	Clinical laboratory measurements including: Hematocrit Hemoglobin Red blood cell counts Absolute monocyte counts HbA1c (average blood glucose)	Intel Basis Smartwatch: - PPG - Accelerometer - Skin Temperature EDA	Implementation of same method for pre-processing clinical records: removal of outliers defined as values >3 S.D. from the mean for that laboratory.	153 features from the continuous wearable sensor data.	RF Two-sided Wilcoxon signed Rank test.
Turning silver into Gold:Domain adaptation with noise labels for wearable cardio-respiratory fitness prediction	Wu et al.	2022	VO_2Max	Actiheart Chest ECG wearable monitor (2-lead) - ECG Wrist device triaxial accelerometer - Accelerometer	Non-wear periods were removed through pre-processing algorithm that identified periods of non-physical heart rate and no movement. Movement intensities were converted into standard metabolic equivalent units (METs). Signals were down sampled to a frequency of 15 minutes.	HR Acceleration	Proposal of UDAMA : Unsupervised Domain Adaptation and Multi-Discriminator Adversarial Training that uses the noisy data that is labelled 'silver standard' to improve the modelling of gold standard data.
Cardiorespiratory fitness estimation in free-living using wearable sensors	Altini et al.	2016	VO_2Max	Holst ECG Necklace: - ECG - Accelerometer	Accelerometer data was band-passed between 0.1 and 10Hz to isolate dynamic components. HR was extracted from RR intervals and averaged over 15s.	From accelerometer: Mean of absolute signal Interquartile range Median, variance Low frequency band signal power. HR	Hierarchical Bayesian models for cardio-respiratory fitness estimation.
Cardiorespiratory fitness estimation using wearable sensors: Laboratory andfree-living analysis of context-specific submaximal heart rates	Altini et al.	2016	VO_2Max	Holst ECG Necklace: - ECG - Accelerometer	For the detection of activities recognised as walking, the accelerometer signal was segmented to 5s and filtered by two separate filters. HR was averaged over 15s.	From accelerometer: Mean of absolute signal Interquartile range Median, variance Low frequency band signal power. HR	Multiple Linear Regressions models using LOPO cross validation.
Prediction of oxygen uptake dynamics by machine learning	Beltrame et al.		VO_2	Hexoskin smartshirt: - ECG - Accelerometer	The HR difference variable was calculated by finding the difference	HR Difference in HR Total hip accereleration	RF

Supplementary Table 2.

analysis of wearable sensors during activities of daily living				- Respiration Band	between the current HR value and the previous value. Features were low-pass filtered 0.01Hz.	Minute ventilation Breathing frequency Walking cadence	
Longitudinal cardio-respiratory fitness prediction through wearables in free-living environments	Spathis et al.	2022	VO_2Max	Actiheart Chest ECG wearable monitor (2-lead) - ECG Wrist device triaxial accelerometer - Accelerometer	Non-wear periods were removed through pre-processing algorithm that identified periods of non-physical heart rate and no movement. Movement intensities were converted into standard metabolic equivalent units (METs). Signals were down sampled to a frequency of 15 minutes. Temporal factors were encoded into sensor timestamps.	Statistical features were extracted from signals and every participant week was represented as a row in the feature vector. Features included: Acceleration HR HRV Acceleration derived Euclidean Norm Minus One Acceleration Derived METs.	Deep Neural Network, a network of 2 densely connected feed-forward layers with 128 units.
The association of pre-operative home accelerometry with cardiopulmonary exercise variables	Cui et al.	2017	VO_2Max Anaerobic Threshold (AT)	AX3 Axivity: - Accelerometer	Employed a fast Fourier transformation to integrate the frequency between 1Hz and 10Hz of the power spectrum. The extraction threshold was set at above 23hrs per day. Each 10s period was categorised as active, stationary or lying.	Activity score Acceleration (mean, S.D., lateral axis mean, vertical axis mean, frequency).	Multiple linear regression analysis.
Can wearable technology be used to approximate cardiopulmonary exercise testing metrics?	Jones et al.	2021	VO_{2peak} Ventilatory equivalent for CO_2 AT Peak work	Garmin Vivosmart HR: - Accelerometer - PPG	Features were averaged across the 7-day wear period. Total METs was calculated by summing METs from across the week.	HR Average HR, Maximum HR Total steps Floors climbed Number of intense minutes of exercise Total calories Total distance travelled	Linear Regression Correlation between fitted and observed values.

Supplementary Table S2. This table provides a comprehensive breakdown of the most relevant papers included in this review. The columns present the data extracted from each paper including the Outcome variable, sensor modality, pre-processing methods, features extracted and the model of analysis.

Appendix B

Supplementary Material for Chapter 3

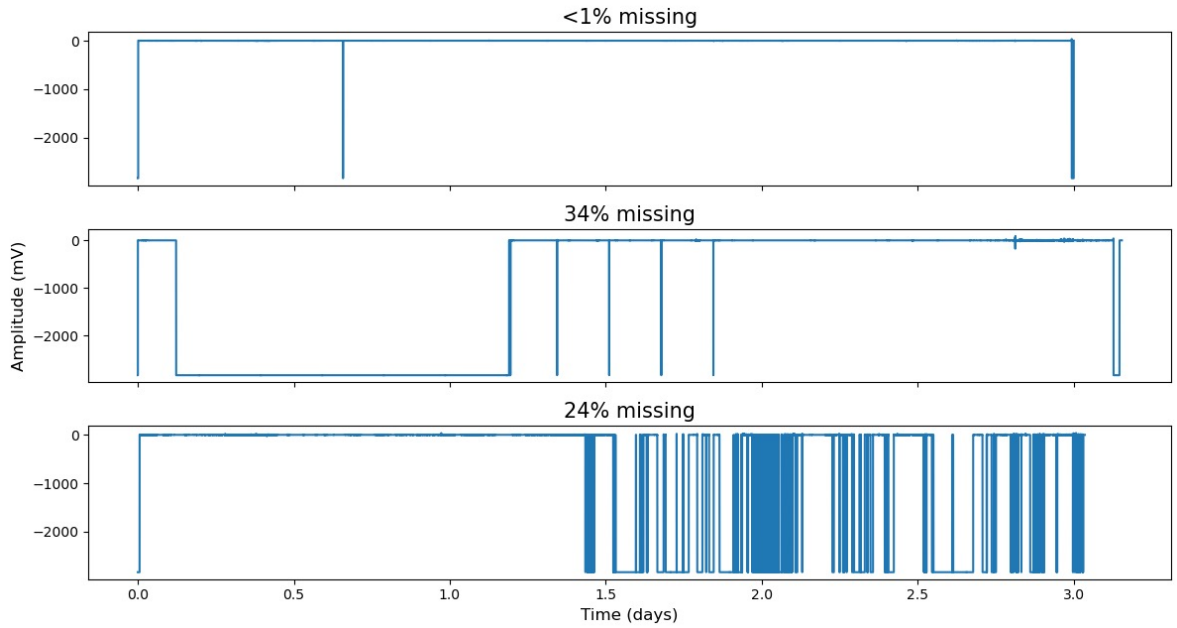


Figure B.1: Examples of ECG signals with varying volumes and patterns of missing data. The first example ($<$1% missing) shows an almost complete recording with only two very short dropouts. The second example (34% missing) illustrates a long initial disconnect resulting in substantial data loss. The third example (24% missing) displays frequent short dropouts throughout the recording, demonstrating a different characteristic of missingness despite similar overall volume.

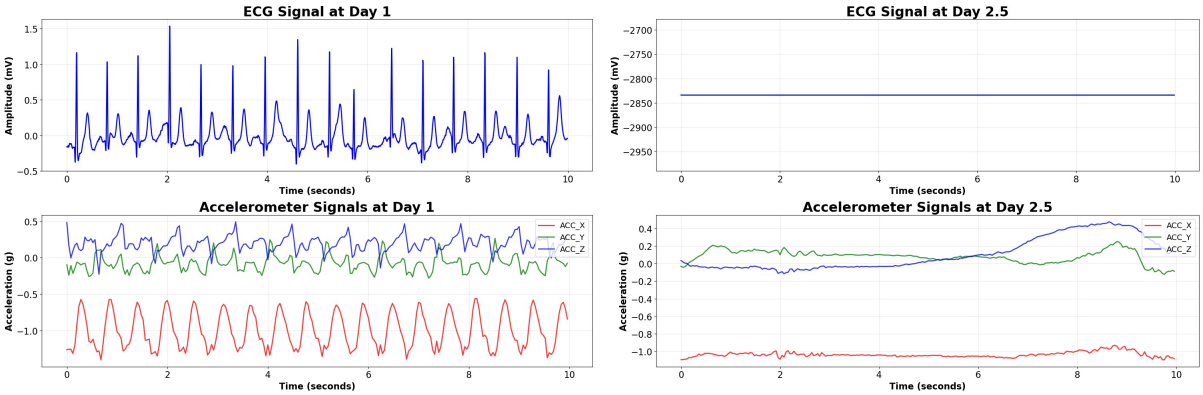


Figure B.2: Examples of missing segments in wearable sensor signals. The two panels on the left show 10-second segments collected simultaneously, with the upper panel displaying ECG signals and the lower panel showing tri-axial accelerometer signals—both without missing data. The two panels on the right show 10-second segments recorded two days later at the same time of day. The upper panel illustrates a period of completely missing ECG data, whereas the lower panel shows continuous tri-axial accelerometer data with no missing periods.

Appendix C

Supplementary Material for Chapter 4

Appendix Figure: Computing in Cardiology (CinC) 2025 Poster

This poster was presented at *Computing in Cardiology (CinC) 2025* and summarises the Chapter 4 work on assessing and implementing ECG signal quality indices (SQIs) and comparing them using sythetic ECG data.

Assessment of ECG Signal Quality Index Algorithms using synthetic ECG data

Aron Berger Syversen[1], Zhiqiang Zhang[1], Jonathan A Batty[1], Matti Kaisti[2], David Jayne[3], David Wong[1]

1. University of Leeds, Leeds, United Kingdom
2. University of Turku, Turku, Finland
3. Leeds Teaching Hospital NHS Trust, Leeds, United Kingdom

Engineering and Physical Sciences Research Council

UNIVERSITY OF LEEDS

Background

Electrocardiographs (ECGs) are widely implemented across healthcare to assess cardiac health. ECG signals can suffer from noise caused by:

- Electrode motion artifacts.
- Electrical interreference.
- Electromyogram noise.

ECG signals in wearable sensors are particularly susceptible to noise.

Why is signal quality important?

The presence of **noise** in an ECG recording can impact on its interpretation. Significant noise can make it difficult to:

- Identify the exact locations of waves.
- Distinguish important changes to the morphology of waves.

What is a Signal Quality Index (SQI)?

SQI algorithms are tools that can *automate the classification of signal quality*. Methods vary from the application of feasibility rules on extracted features, to machine learning models employed on raw signals. Our aim:

To assess the outputs of several publicly available SQI tools against various forms of noise in an ECG signal.

Results

The results in table 1 present the thresholds at which further increase in noise would lead to a <0.5 proportion of signals being labelled by the SQIs as 'acceptable'.

Table 1. Table showing the thresholds from each SQI in each category of noise

Noise Type	SQI1	SQI2	SQI3	SQI4
Heart Rate (bpm)	155	925	495	255
White Noise (dB)	4.75	1.32	N/A	1.41
Power Line (dB)	4.13	10.71	0.70	0.70
Motion Artifacts	walking	N/A	N/A	N/A

SQI comparison with clinical expert:

Table 2. Table showing number of criteria that agreed with each SQI label .

	SQI1	SQI2	SQI3	SQI4
Criterion 1	12	9	12	15
Criterion 2	10	11	12	13
Criterion 3	10	7	4	7
Criterion 4	10	7	4	7

- Four *SQIs* were inconsistent with each other.
- *SQIs* frequently disagreed with cardiologist assessment. The *SQIs* had the lowest agreement when asked if the ECG was '*clinically useful*' (criterion 4) - between 4/16 and 10/16.

Pipeline for ECG synthesis & assessment

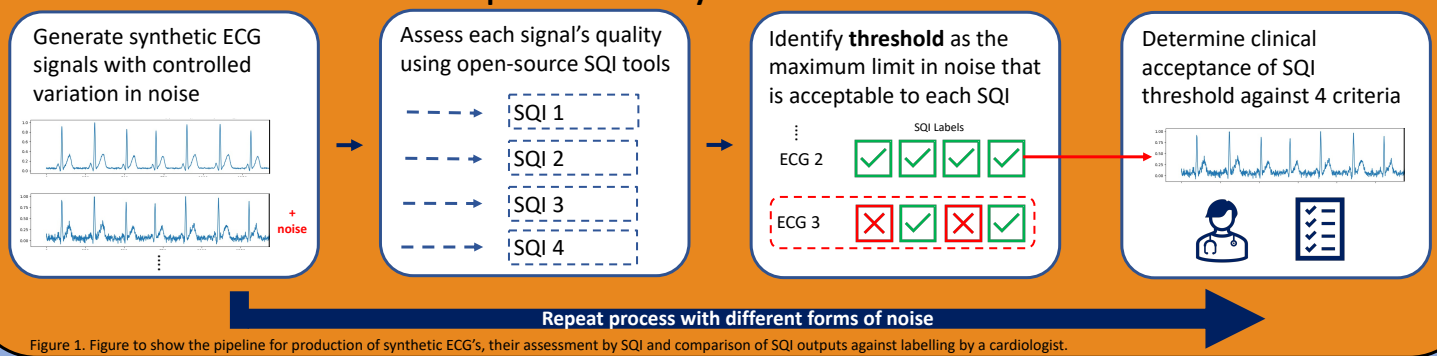


Figure 1. Figure to show the pipeline for production of synthetic ECG's, their assessment by SQI and comparison of SQI outputs against labelling by a cardiologist.

Methods

Synthetic ECG generator

The framework and code for developing synthetic ECG signals is taken from [1]. Variation was added into the segments in 4 forms:

- **Heart Rate (bpm)**
- **White Noise (dB)**
- **Power Line Interference (dB)**
- **Motion Artifacts**

Pipeline

These forms of noise were gradually increased whilst other parameters kept stable. At each increment in noise, 100x ECG signals were produced and assessed with 4 *SQIs*. The proportion of 100 signals labelled by each *SQI* as acceptable was calculated. The signals are also assessed by a cardiologist against 4 criteria (below) and compared against *SQIs* outputs:

1. *Can you estimate a plausible HR?*
2. *Can you located all QRS complexes?*
3. *Can you locate all P & T waves?*
4. *Is the signal clinically useful?*

SQI tools:

SQI1. Orphanidou et al. (2015). 3 feasibility rules & an adaptive template matching threshold [2].

SQI2. Zhao & Zhang. (2018). Combines simple heuristic fusion and fuzzy comprehensive evaluation [3].

SQI3. Kramer et al. (2022). ECGAssess : 3 stage process [4].

SQI4. Elgendi et al. (2023). SQI uses the same 3 rules as SQI3 alongside a **CNN classifier** [5].

Discussion

Evaluation of outcomes:

This study investigated the performance of four publicly-available *SQIs* against cardiologist assessments on synthetically-generated ECGs. The experiment yielded several key findings:

- *SQI1* and *SQI4* had the highest agreement with the cardiologist assessment.
- For criteria relating to HR extraction (*criteria 1 & 2*), *SQIs* showed moderately good agreement with the cardiologist.
- For the final two criteria relating to other clinical features of the ECG (identification & morphology of P & T-waves), *SQIs* had low agreement with the cardiologist.

Conclusion:

The majority of SQI tools were unable to detect serious degradation in signals and often disagreed with labelling from a cardiologist.

This suggests the limited suitability of these SQI for clinical applications beyond extracting heart rate from the ECG.

References

[1] Karhinoja K, Vasankari A, Jukka-Pekka S, Airola A, Wong D, Kaisti M. Flexible framework for generating synthetic electrocardiograms and photoplethysmograms. arXiv:2408.16291. 2024.
[2] Orphanidou C, Bonnici T, Charlton P, Clifton D, Vallance D, Tarassenko L. Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring. IEEE J Biomed Health Inform. 2015 May;19(3):832-8. doi: 10.1109/JBHI.2014.2338351. Epub 2014 Jul 23. PMID: 25069129.
[3] Zhao Z, Zhang Y. SQI quality evaluation mechanism of single-lead ECG signal based on simple heuristic fusion and fuzzy comprehensive evaluation. Frontiers in physiology. 2018 Jun 14;9:361542.
[4] Kramer L, Menon C, Elgendi M. ECGAssess: A Python-Based Toolbox to Assess ECG Lead Signal Quality. Front Digit Health. 2022 May 6;4:847555. doi: 10.3389/fdgth.2022.847555. PMID: 35601886; PMCID: PMC9120362.
[5] Elgendi M, van der Bijl K, Menon C. An Open-Source Graphical User Interface-Embedded Automated Electrocardiogram Quality Assessment: A Balanced Class Representation Approach. *Diagnostics*. 2023; 13(22):3479. https://doi.org/10.3390/diagnostics13223479

Appendix D

Supplementary Material for Chapter 5

Appendix Figure: IEEE EMBC 2025 Poster Presentation

This poster was presented at the *IEEE Engineering in Medicine and Biology Conference (EMBC) 2025*, summarising the work presented in Chapter 5 on machine learning approaches for predicting VO_2max from wearable sensor data. It highlights the comparison of models, and the overall methodological pipeline.

Machine Learning for VO_2max Predictions: A Comparison of Methods using Wearable Sensor Data

Aron Berger Syversen[1], Alexios Dosis[2], Zhiqiang Zhang[1], David Jayne[2], David Wong[1]

1. University of Leeds, Leeds, United Kingdom
2. Leeds Teaching Hospital NHS Trust, Leeds, United Kingdom

UNIVERSITY OF LEEDS

Engineering and Physical Sciences Research Council

Background

Why should we predict VO_2?

Cardiorespiratory fitness (CRF), commonly measured as **VO_2max**, is a strong predictor of health outcomes and surgical risk [1].
The gold standard for assessing VO_2max is **cardiopulmonary exercise testing (CPET)**. While accurate, CPET is **expensive, resource-intensive**, and **not scalable** for large or high-risk populations.
→ **Wearable sensors** could provide an alternative.

Application of ML models

To our knowledge, no study has directly compared multiple machine-learning approaches for VO_2max estimation in a clinical pre-operative cohort using free-living wearable sensor data.

Aims: To directly compare the performance of 5 machine learning models for estimating VO_2max from wearable sensor data in a preoperative cohort.

Methods

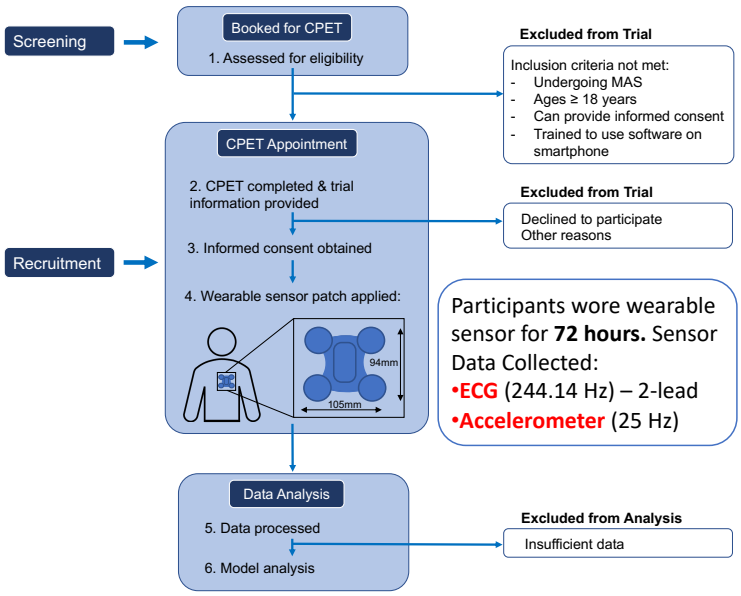


Figure 1. Participant flow through the study

Signal Processing & Feature Extraction

- Raw ECG signals were processed using a validated signal quality index (SQI) to identify clean 10-second windows [2].
- Accelerometer data were processed to classify daily activity (e.g. sleep, sedentary, MVPA) and calculate step counts per minute using open-source tools[3].

Model Development

1. **Support Vector Regression.**
2. **Random Forest**
3. **XGBoost**
4. **Mult—layer Perceptron**
5. **Multiple Linear Regression**

Model Validation

Validation: 5-fold cross-validation

Results

A total of 169 participants were included after exclusions for poor data quality.

Table 1. Participant characteristics

Variable	Men (n=125)	Women (n=44)
Age (years)	68.78 ± 10.17	67.16 ± 13.37
BMI (kg/m^2)	28.26 ± 5.05	30.27 ± 7.49
Collected data (hours)	72.51 ± 14.34	73.73 ± 18.28
VO_2max (ml/kg/min)	18.73 ± 4.72	15.25 ± 3.70
Average Daily Step Count	3836± 3043	2778 ± 2215

Model Performance

- **MLR** achieved the best overall fit (RMSE = 3.35; R^2 = 0.46; SEE = 3.95; Pearson r = 0.68), outperforming all other models.
- **SVR** matched MLR on Pearson correlation (r = 0.68).
- **XGBoost** showed the lowest performance across every metric (highest RMSE, lowest R^2 and SEE).

Table 2. Comparison of metrics across 5 ML models.

Model	RMSE	R^2	Correlation	SEE
MLR	**3.35 ± 0.32**	**0.46 ± 0.13**	**0.68 ± 0.09**	**3.95 ± 0.39**
RF	3.82 ± 0.33	0.31 ± 0.13	0.61 ± 0.13	7.94 ± 0.61
XGBoost	3.86 ± 0.42	0.30 ± 0.14	0.59 ± 0.12	8.04 ± 0.82
SVR	3.40 ± 0.23	0.45 ± 0.11	**0.68 ± 0.08**	4.01 ± 0.27
MLP	3.69 ± 0.34	0.32 ± 0.05	0.63 ± 0.11	7.68 ± 0.76

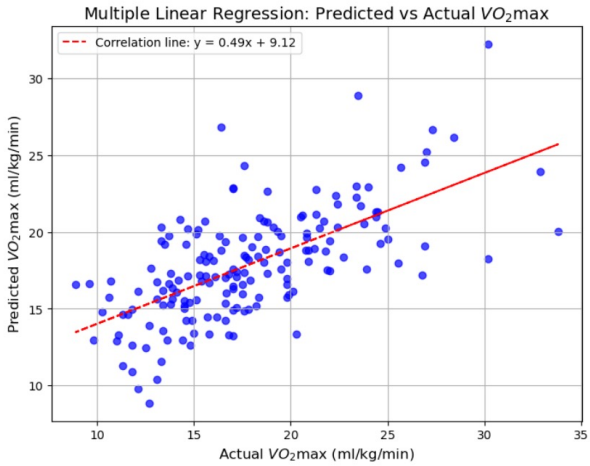


Figure 2. Scatter plot to show predicted VO_2max values from MLR against actual values.

Conclusions

- **Linear regression outperforms ML models** in this pre-operative cohort, achieving the highest R^2 (0.46 ± 0.13), lowest RMSE (3.35 ± 0.32).
- **Model simplicity may suit clinical heterogeneity**: the strong linear relationships between age, sex and step count with VO_2 in a diverse patient group favours interpretable linear models over non-linear ones.
- **Performance lags exercise-derived models**: unlike studies using sub-maximal or exercise-test variables (R > 0.8), purely free-living wearable data yield lower accuracy. However, they offer a scalable, low-burden alternative for high-risk surgical populations.
- **Future work** should investigate including clinical variables alongside wearable features, refining task-specific SQI frameworks for ECG pre-processing, and further validate HRV's role in VO_2max estimation.

References

[1] Ross R, Blair SN, Arena R, Church TS, Després JP, Franklin BA, et al. Importance of Assessing Cardiorespiratory Fitness in Clinical Practice: A Case for Fitness as a Clinical Vital Sign: A Scientific Statement from the American Heart Association. Circulation [Internet]. 2016 Dec 13 [cited 2024 Nov 12];134(24):e653–99. Available from: https://www.ahajournals.org/doi/10.1161/CIR.0000000000000461
[2] Orphanidou C, Bonnici T, Charlton P, Clifton D, Vallance D, Tarassenko L. Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring. IEEE J Biomed Health Inform [Internet]. 2015 May 1 [cited 2023 Jun 29];19(3):832–8. Available from: https://pubmed.ncbi.nlm.nih.gov/25069129/
[3] Walmsley R, Chan S, Smith-Byrne K, Ramakrishnan R, Woodward M, Rahimi K, Dwyer T, Bennett D, Doherty A. Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease. British journal of sports medicine. 2022 Sep 1;56(18):1008-17.

Acknowledgements

This work uses data provided by patients and collected by the NHS as part of their care and support.
Icon by Minh Do, from Freepik.

Appendix E

Supplementary Material for Chapter 7

Table E.1: Comparison of mean (SD) improvements from baseline (non-HRV) to HRV models using the original SQI versus the HR-specific SQI approach.

Metric	**Original SQI**	**HR-specific SQI**
R^2	+0.05 (0.04)	+0.07 (0.04)
Correlation (r)	+0.04 (0.02)	+0.05 (0.03)
MAE (ml·kg^{-1}·min^{-1})	-0.14 (0.16)	-0.16 (0.14)
RMSE (ml·kg^{-1}·min^{-1})	-0.17 (0.13)	-0.23 (0.16)
APE (%)	-0.68 (0.97)	-0.97 (0.83)

List of Figures

List of Tables