POLITECNICO DI TORINO

Master Degree in Biomedical Engineering

Master Thesis

Comparative evaluation of PPG-based heart-rate estimation across three wearable devices



Supervisor
Dr. Luigi Borzì
Dr. Irene Rechichi

Candidate
Sofía Escobar Londoño

Academic Year 2025/2026

Summary

Wearable photoplethysmography (PPG) is widely used to estimate heart rate outside the clinical setting, where movement and posture are not controlled. This thesis compares three wearable PPG devices recording simultaneously at three body sites during activities of daily living.

Twenty-two participants were recorded in the laboratory while wearing all three devices: the Persimmon prototype on the forearm, a Polar Verity Sense on the proximal forearm, and a Shimmer3 GSR+ at the earlobe. The protocol included resting while seated, three tasks performed while seated, resting while standing without support, walking, and postural transitions. The optical signals from all three devices were processed using the same heart-rate estimation pipeline, reducing differences attributable to device-specific heart-rate algorithms. Agreement was assessed using the concordance correlation coefficient (CCC), the intraclass correlation coefficient (ICC), the mean absolute error (MAE), and Bland-Altman limits of agreement, with CIs accounting for within-participant dependence.

The two comparison devices, Polar and Shimmer, showed greater agreement with each other (CCC = 0.87) than either did with the prototype (CCC = 0.72-0.74), while the ICC across all three devices was 0.79. Agreement varied widely across recordings and depended strongly on measurement condition. Seated rest produced the lowest inter-device MAE (2.7 bpm), whereas walking produced the highest among sustained activities (12.9 bpm). The seated tasks and unsupported standing rest fell between these conditions (MAE = 4.8-6.5 bpm). Standing rest produced more than twice the MAE observed during seated rest, although both were resting conditions. During walking and at high measured movement levels, agreement between Polar and Shimmer also deteriorated, indicating that disagreement under these conditions could not be attributed to the prototype alone. Cross-group analyses did not identify significant differences in agreement associated with sex or skin phototype.

The main limitations are the absence of an electrocardiographic reference, the lack of dedicated motion-artefact suppression, a cohort predominantly composed of young adults, and skin phototypes restricted to Fitzpatrick types I-IV. Overall, the study identifies the conditions under which the three devices provide consistent heart-rate estimates and highlights robustness to postural and movement-related artefacts as a key target for future development before broader use in ambulatory monitoring.

Contents

List of Tables 4

List of Figures 5

1 Introduction 6
- 1.1 Heart rate as a physiological parameter . . . 6
- 1.2 Principles of heart-rate measurement . . . 7
- 1.3 PPG in wearable devices . . . 9
- 1.4 Challenges of wearable PPG during physical activity . . . 10
- 1.5 Validation and agreement . . . 11
- 1.6 Objective . . . 12

2 Related work 13
- 2.1 Validation of wearable PPG for heart-rate monitoring . . . 13
- 2.2 Performance during activity and motion . . . 15
- 2.3 Conventional signal conditioning . . . 16
- 2.4 Signal processing and motion-artefact mitigation . . . 17
- 2.5 Effects of measurement conditions and participant characteristics . . . 18
- 2.6 Prior validation of the comparison devices . . . 19
- 2.7 Research gap and study motivation . . . 20

3 Materials and Methods 22
- 3.1 Data collection . . . 22
  - 3.1.1 Participants . . . 22
  - 3.1.2 Wearable devices . . . 24
  - 3.1.3 Experimental protocol . . . 28
  - 3.1.4 Generated data . . . 33
- 3.2 Preprocessing . . . 34
  - 3.2.1 Data standardisation and organisation . . . 34
  - 3.2.2 Synchronisation and resampling . . . 35
  - 3.2.3 Activity annotation . . . 39

3.2.4 Filtering . . . 40
3.2.5 Beat detection . . . 42
3.2.6 Heart-rate estimation . . . 45
3.2.7 Data-quality review and exclusions . . . 47
3.3 Statistical analysis . . . 48
3.3.1 Agreement metrics . . . 48
3.3.2 Pooled analysis . . . 51
3.3.3 Per-recording analysis . . . 51
3.3.4 Subgroup analyses . . . 51

4 Results 54
4.1 Data availability . . . 54
4.2 Choice of the bandpass filter . . . 55
4.3 Overall agreement . . . 55
4.4 Agreement per recording . . . 58
4.5 Agreement by activity . . . 60
4.6 Agreement by motion level . . . 63
4.7 Agreement by sex . . . 66
4.8 Agreement by skin phototype . . . 67
4.9 Agreement against the practical benchmarks . . . 69

5 Discussion 70
5.1 Interpretation of the overall agreement . . . 70
5.2 Effects of activity and movement . . . 71
5.3 Effects of participant characteristics . . . 73
5.4 Practical implications . . . 73
5.5 Methodological strengths and considerations . . . 74
5.6 Limitations . . . 75

6 Conclusion and future work 77

List of Tables

2.1 Studies most comparable to this work 15
3.1 Participant information 23
3.2 The Fitzpatrick skin phototype scale 23
3.3 Self-reported physical activity levels 24
3.4 Summary of the four devices 27
3.5 Activity labels and nominal durations 33
3.6 Raw data generated during acquisition 34
3.7 Recording-level exclusions and data-quality notes 48
3.8 Pairwise agreement and difference metrics 50
4.1 Usable-window availability by sensor 54
4.2 Comparison of the two candidate filters 55
4.3 Pooled concordance between devices 56
4.4 Pairwise difference and association summaries 57
4.5 Agreement by activity 62
4.6 Between-participant subgroup comparisons 68

List of Figures

1.1 The ECG and the RR interval . . . 8
1.2 The components of the PPG signal . . . 8
3.1 The Persimmon prototype . . . 25
3.2 The Polar Verity Sense and its logging application . . . 26
3.3 The Shimmer3 unit and its ear-clip probe . . . 26
3.4 The Nordic Thingy:52 and the data-collection application . . . 27
3.5 Device placement on the forearm . . . 28
3.6 Equipment used in a recording session . . . 29
3.7 The acquisition station . . . 30
3.8 A recording in progress . . . 31
3.9 Objects used by the activity protocol . . . 31
3.10 The activity protocol . . . 32
3.11 Acceleration over the whole recording, before alignment . . . 36
3.12 Acceleration before synchronisation . . . 37
3.13 Acceleration after synchronisation . . . 38
3.14 Residual drift check . . . 38
3.15 The activity tagging interface . . . 39
3.16 The bandpass filter on one interval . . . 42
3.17 The peak-detection steps . . . 44
3.18 Cleaning and peak detection . . . 45
3.19 Heart rate per window . . . 47
4.1 Bland-Altman agreement . . . 57
4.2 Agreement scatter plots . . . 58
4.3 Pairwise agreement in each recording . . . 59
4.4 Three-device ICC by recording . . . 60
4.5 Concordance by activity . . . 63
4.6 Optical signals at three motion levels . . . 65
4.7 Concordance by motion level . . . 65
4.8 Mean absolute difference by motion level . . . 66
4.9 Concordance by skin phototype . . . 68

Chapter 1

Introduction

Heart rate is one of the few cardiovascular parameters that can be monitored continuously outside clinical environments using a body-worn device without requiring active intervention from the wearer. Optical sensing techniques, particularly photoplethysmography (PPG), have enabled such non-invasive monitoring. However, the reliability of heart-rate measurements derived from optical signals can deteriorate substantially during physical movement. This limitation is particularly relevant for wearable devices, which are intended to operate under real-world conditions that extend beyond controlled resting states. Consequently, performance demonstrated under stationary or resting conditions alone is insufficient to establish the suitability of a wearable device for its intended application.

This chapter provides the background necessary to formulate and contextualise this problem. It first introduces heart rate as a physiological parameter and discusses the rationale for its measurement. It then describes the electrical and optical principles underlying heart-rate measurement, followed by an examination of the effects of physical movement on optical measurements. Finally, the requirements for validating wearable heart-rate monitoring devices are discussed, and the objective of this thesis is presented.

1.1 Heart rate as a physiological parameter

Heart rate is defined as the number of ventricular contractions per unit time and is conventionally expressed in beats per minute (bpm). Cardiac rhythm is initiated by the sinoatrial node, whose pacemaker cells undergo spontaneous depolarisation without external stimulation. Their intrinsic firing rate is continuously modulated by the autonomic nervous system: parasympathetic activity decreases the sinoatrial firing rate, whereas sympathetic activity increases it [1]. As autonomic balance changes in response to physiological demands, heart rate varies over short

timescales with factors such as metabolic demand, posture, and autonomic state.

Resting heart rate also has established prognostic relevance. Cohort studies have shown that an elevated resting heart rate is associated with increased cardiovascular and all-cause mortality. This association remains significant after adjustment for age, blood pressure, smoking, and other established cardiovascular risk factors, indicating that resting heart rate provides prognostic information beyond that captured by conventional risk factors [2].

However, resting heart rate represents only one aspect of cardiac response. In exercise physiology, training intensity is commonly prescribed relative to an age-predicted maximal heart rate [3], making the heart rate reached during exertion relevant for assessing exercise intensity. Outside these controlled contexts, heart rate varies continuously throughout daily life in response to changes in physical activity, posture, and physiological demand. Consequently, a single measurement obtained under resting conditions cannot characterise these temporal variations.

Monitoring heart rate outside the clinical environment therefore requires devices that can be worn for extended periods and continue to provide reliable measurements during movement and changes in posture. For wearable heart-rate monitors, performance under these conditions is therefore essential when assessing their suitability for real-world use.

1.2 Principles of heart-rate measurement

Electrocardiography (ECG) is the established method for measuring cardiac timing. It records the electrical activity of the heart through electrodes placed on the skin. Within each cardiac cycle, the R wave typically appears as a prominent deflection whose temporal position can be identified with high precision. Automated methods for R-wave detection have been developed for several decades, with the algorithm proposed by Pan and Tompkins [4] representing a seminal contribution in this field. The temporal interval between successive R waves defines the RR interval, from which heart rate can be derived on a beat-to-beat basis (Figure 1.1). Owing to its established temporal precision and reliability for cardiac timing, ECG is commonly used as the reference method for evaluating alternative heart-rate measurement techniques.

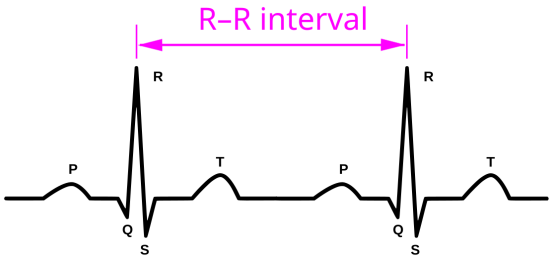


Figure 1.1. Two consecutive cardiac cycles in an ECG signal. Heart rate can be derived from the interval between successive R waves [5].

Photoplethysmography (PPG), in contrast, detects peripheral changes in blood volume associated with the cardiac cycle. A light source illuminates the tissue, while a photodetector measures either the light transmitted through the tissue, as in transmittance PPG, or the light returning from the tissue surface, as in reflectance PPG. Transmittance measurements are commonly performed across thin anatomical sites such as the fingertip or earlobe, whereas reflectance configurations are more suitable for wearable sensors placed on the skin surface. The detected light intensity varies with the optical properties of the illuminated tissue and with changes in local blood volume. The resulting PPG signal comprises a large slowly varying component associated with tissue, venous blood, and mean arterial blood volume, together with a smaller pulsatile component associated with arterial blood-volume changes produced by each cardiac cycle [6]. The interval between successive pulses in this pulsatile component can therefore be used to estimate pulse rate (Figure 1.2).

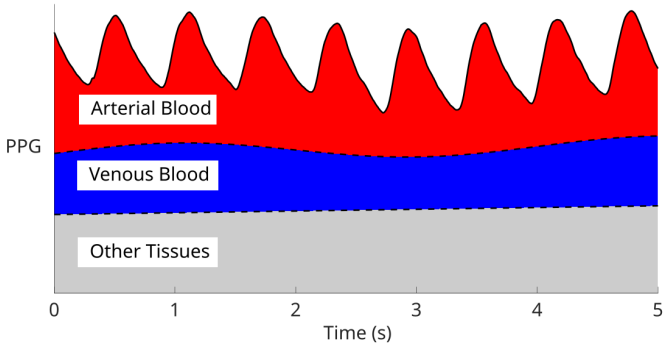


Figure 1.2. Schematic representation of the main components contributing to the PPG signal. Tissue, venous blood, and mean arterial blood volume contribute predominantly to the slowly varying component, whereas pulsatile arterial blood-volume changes produce the smaller cardiac-synchronous component [7].

ECG and PPG therefore provide information about different physiological events. ECG measures the electrical activity associated with cardiac depolarisation, whereas PPG detects peripheral haemodynamic changes that occur after ventricular ejection and propagation of the resulting pulse wave through the arterial system. The quantity derived directly from PPG is therefore pulse rate rather than heart rate. Under conditions in which each cardiac cycle produces a detectable peripheral pulse, the two rates are expected to correspond closely. For consistency with common terminology in wearable monitoring, the term heart rate is used throughout this work for estimates derived from the optical signals. The morphology and timing of the peripheral PPG pulse can additionally be influenced by vascular properties, local tissue composition and perfusion, the anatomical measurement site, and the optical and mechanical coupling between the sensor and the skin. These factors can modify the recorded PPG waveform independently of changes in cardiac timing.

1.3 PPG in wearable devices

PPG is particularly suitable for integration into wearable devices because it can be implemented using relatively simple and compact optical hardware. A typical configuration consists of one or more light-emitting diodes and a photodetector, components that can be incorporated into small, low-power sensing systems. In reflectance mode, both the light source and detector are positioned on the same side of the tissue, allowing measurements to be obtained from a single skin surface without the electrodes or conductive interfaces required for electrocardiographic recordings. This configuration facilitates integration into devices such as watches, rings, and armbands. As a result, PPG has become widely used for heart-rate monitoring in consumer and research wearable devices [8].

The characteristics of the recorded PPG signal depend strongly on the anatomical measurement site. In a comparison of six sites within the same participants, the fingertip produced the largest pulse amplitude, while the earlobe was among the sites with the highest proportion of analysable waveforms. In contrast, the dorsal wrist, a common location for consumer wearable devices, exhibited the lowest pulse amplitude among the sites examined [9]. Anatomical placement can also influence the performance of heart-rate estimation during physical activity. In a comparison of measurements obtained at the wrist, forearm, and upper arm against a chest-strap reference, the upper arm showed better agreement than the wrist across all tested exercise intensities, while the forearm also performed better than the wrist in all conditions except a brief bout of burpees [10]. These findings indicate that measurement site is an important factor in wearable PPG performance.

Heart rate must be estimated from the acquired PPG waveform rather than

measured directly. A typical processing pipeline includes signal acquisition, preprocessing to suppress unwanted components and noise, identification of the cardiac-related information, and conversion of this information into a heart-rate estimate. In time-domain approaches, individual pulse events are detected and the intervals between successive pulses are used to derive beat-to-beat or windowed heart-rate estimates. Frequency-domain approaches instead estimate the dominant periodic component associated with cardiac activity over a defined time window. Although these approaches differ in their implementation, both infer heart rate from the optical waveform. Consequently, the performance of a wearable PPG system depends not only on the sensing hardware and measurement site, but also on the signal-processing and heart-rate estimation algorithms applied to the recorded signal.

1.4 Challenges of wearable PPG during physical activity

The pulsatile component of the PPG signal is small relative to the total optical signal. Consequently, changes in the optical coupling between the light source, tissue, and photodetector can introduce signal variations of comparable or greater magnitude than the cardiac-related component. Under controlled resting conditions, these interactions are relatively stable; during everyday activities, however, they are continuously affected by movement and changes in sensor-skin contact.

Motion artefact is one of the principal sources of degradation in wearable PPG measurements [11, 12]. Movement can produce relative displacement between the sensor and the skin, modify the contact pressure applied to the tissue, and alter local blood volume within the illuminated region. The challenge is not determined solely by the magnitude of these disturbances. During repetitive physical activity, motion can introduce periodic or quasi-periodic components within the frequency range occupied by the cardiac signal. When the spectral content of motion artefact overlaps with the cardiac-frequency range, conventional frequency-selective filtering cannot reliably separate the two components [8, 11]. The practical impact can be substantial: across wearable PPG recordings, the proportion of data reported as being of sufficient quality for pulse-rate monitoring has ranged from 14% to 56% [8].

Posture and sensor position relative to the heart can also influence PPG signal quality, even in the absence of substantial physical activity. In a study involving more than eleven hundred participants, body posture and the height of the sensor relative to the heart were identified as important determinants of wrist PPG signal quality. Measurements obtained while participants were seated with the

arm resting in the lap showed a markedly higher signal-to-noise ratio than measurements obtained while standing with the arm hanging alongside the body [13]. Resting measurements should therefore not be considered a single homogeneous condition, because differences in posture and arm position can substantially affect the characteristics of the recorded signal.

Wearer-specific physiological and anatomical characteristics may further influence PPG measurements. The amplitude of the pulsatile component depends partly on local blood perfusion, and reduced peripheral blood flow is a recognised challenge for wearable PPG sensing [8]. Skin pigmentation has received particular attention because melanin absorbs light within wavelengths commonly used for PPG and may therefore affect the detected optical signal. Evidence that this effect translates into systematic bias in estimated pulse rate, however, is less conclusive. A meta-analysis including more than one hundred and forty thousand paired PPG and ECG observations found no statistically or clinically significant pigmentation-related bias in pulse-rate estimates, although the 95% limits of agreement were not contained within the ±5 bpm criterion in the light-, medium-, and dark-pigmentation groups [14]. These findings relate to the resulting pulse-rate estimates and therefore do not exclude the possibility that skin pigmentation influences the underlying optical signal.

Performance obtained under resting conditions does not necessarily predict performance during physical activity. In a free-living validation of a consumer wearable against a chest-strap reference, overall agreement deteriorated substantially during walking: the concordance correlation coefficient decreased from 0.796 across the complete recording period to 0.164 during walking, and the wearable underestimated heart rate in this condition [15]. Wearable heart-rate monitoring systems should therefore be evaluated across the range of postural and movement conditions expected during their intended use.

1.5 Validation and agreement

A new wearable sensor requires formal validation rather than visual inspection of its recorded signals. Despite the large number of devices available on the market, formal validation remains limited [16]. For heart-rate monitoring, performance criteria have been defined using the mean absolute percentage error (MAPE) [17].

Validation also requires the use of statistical methods appropriate to the question being addressed. Two devices may be strongly correlated while still showing poor agreement, because correlation quantifies association rather than agreement [18]. Agreement should therefore be assessed using methods developed specifically for this purpose. Bland-Altman analysis examines the differences between two measurement methods as a function of their mean and summarises them in terms

of bias and limits of agreement. The concordance correlation coefficient (CCC) accounts for both random variability and systematic differences [19], while the intraclass correlation coefficient (ICC) can be used to quantify agreement when an appropriate formulation is selected [20]. Error metrics provide complementary information on the magnitude of the discrepancy between measurements.

A further distinction concerns the interpretation of the comparison. An ECG can provide an independent reference against which a wearable device is evaluated. A comparison among PPG devices alone, however, establishes agreement rather than absolute accuracy. When the devices disagree, such a comparison cannot determine which estimate is closer to the true cardiac rate. Agreement and accuracy should therefore be treated as distinct concepts in the absence of an ECG reference.

1.6 Objective

The MOR4.3 research prototype (hereafter referred to as *Persimmon*) was developed within the PERSIMMON project, a European initiative on wearable health monitoring. It is a body-worn optical research device and is not a commercial product. The prototype has not yet been evaluated against established devices, while the project specifies target heart-rate accuracy values of within ±5 bpm at rest and ±10 bpm during exercise [21]. Its performance under the intended conditions of use, including activities of daily living rather than controlled rest alone, has not yet been characterised.

This thesis addresses this gap by comparing the prototype with two established optical devices during activities of daily living. Two aspects of the study design determine how the results can be interpreted. First, the same heart-rate estimation pipeline is applied to the optical signals from all three devices, thereby reducing differences attributable to device-specific heart-rate algorithms. Second, no electrocardiographic reference was recorded; consequently, the analysis assesses agreement between devices rather than accuracy relative to a reference cardiac rate. Chapter 3 describes the devices, protocol, and signal processing procedures.

The objective is to determine how closely the heart-rate estimates obtained from the prototype agree with those obtained simultaneously from the two comparison devices, and to identify the conditions under which this agreement is maintained or deteriorates. Agreement is examined according to activity, measured movement level, and participant characteristics, sex and Fitzpatrick skin phototype [22]. The two comparison devices are also compared with each other throughout the analysis, allowing disagreement involving the prototype to be interpreted in relation to the consistency between the comparison devices under the same conditions.

Chapter 2

Related work

Wearable PPG has been validated in numerous previous studies. This chapter reviews methods and findings from the related literature, identifies current limitations, and positions the present work in relation to them.

2.1 Validation of wearable PPG for heart-rate monitoring

An umbrella review of consumer wearable technologies, covering 24 systematic reviews, 249 validation studies, and 430,465 participants, reported a mean bias of approximately ±3% for heart rate [16]. The same review found that only about 11% of commercially available devices had been validated for at least one biometric outcome. Validation was concentrated on a limited number of devices, while most had not been evaluated.

Many wearable PPG validation studies follow a similar design. A wearable device is used simultaneously with an electrical reference, and the heart rate reported by the wearable is compared with that reference across a protocol of increasing intensity. The reference is commonly an ECG-based chest strap [10, 15, 23, 24], and occasionally a clinical-grade ECG patch [25] or continuous ECG telemetry [26]. Results are typically summarised using a concordance or intraclass correlation coefficient, Bland-Altman analysis, and percentage error. A MAPE below 10% has been used as one of the acceptance criteria in validation procedures based on Consumer Technology Association standards [17].

Measurement condition is an important determinant of the observed performance. Sixteen participants wore two Polar devices during nine activities ranging from lying down to interval training. The Verity Sense worn on the upper arm produced a mean absolute percentage error of 1.35%, whereas a wrist-worn watch used by the same participants produced an error of approximately 6-7% [23]. Five

arm-worn monitors compared against a chest strap in 28 adults all reached a CCC above 0.99 at rest, whereas values fell below 0.50 during thirty seconds of burpees [10]. Six consumer- and research-grade devices worn by 53 participants and compared with an ECG patch showed errors that were, on average, 30% higher during activity than at rest [25].

Free-living activity can reveal this dependence more clearly than graded exercise. A consumer watch worn for twelve hours of free living showed a CCC of 0.796 and a MAPE of 5.1% across the complete recording. The recording was classified into eleven categories of everyday activity. During walking, agreement with the reference was substantially lower, and the watch underestimated heart rate by 6.6 bpm [15]. Aggregate performance metrics can therefore mask activity-specific deterioration in performance.

Research prototypes have also been evaluated under controlled exercise. One prototype, tested in 20 adults against a chest strap during an incremental shuttle walk test, reached an ICC of 0.96 at rest and during recovery, and 0.78 at maximal heart rate [24]. Agreement was therefore lower at maximal effort than at rest and during recovery, consistent with the condition-dependent performance reported for commercial devices.

A recent study used a design more closely related to the present work by recording raw PPG simultaneously from four wearable devices positioned at the ear, finger, wrist, and neck. The same heart-rate estimation methods were applied to the signals from each device and evaluated against an electrocardiographic reference [27]. In 20 participants recorded over several days of free living, mean absolute error varied across measurement sites, from 2.30 bpm at the ear to 8.68 bpm at the neck. This design allows differences between measurement sites to be examined without introducing different device-specific heart-rate algorithms, although the recordings were not obtained under a structured activity protocol.

Table 2.1 summarises the studies most comparable to the present work.

Almost all of this literature uses an electrical reference. Comparisons among optical devices alone are less common, and when they are performed, one device is often treated as the laboratory reference. Electrodermal and optical signals from 12 participants were recorded simultaneously using a Fitbit Sense, an Empatica E4, and a Shimmer GSR3+, with the Shimmer serving as the laboratory device [28]. Such a design addresses a different question: none of the measurements is known to be closer to the true cardiac rate than the others, so the comparison establishes agreement rather than determining which device is more accurate.

Table 2.1: Studies most directly comparable to the present work, ordered by year. *Rate from* indicates whether the compared heart rate was the value reported directly by the device or obtained using a common processing pipeline.

Study	Devices/site	N	Reference	Protocol	Rate from	Main finding
Shcherbina et al. [26]	7 consumer, wrist	60	ECG telemetry	sitting, walking, running, cycling	device	error lowest cycling, highest walking
Bent et al. [25]	6 consumer and research-grade; wrist and arm	53	ECG patch	rest, walking, typing	device	no difference across skin tones; error 30% higher in activity
Giggins et al. [15]	Withings ScanWatch, wrist	7	Polar H10	12 h free living	device	CCC 0.796 overall, lower during walking
Ronca et al. [28]	2 wrist devices and a Shimmer GSR3+	12	none; Shimmer as laboratory device	rest	not reported	PPG features correlated; consumer device least reliable
Loro et al. [24]	research prototype	20	Polar H10	incremental shuttle walk	not reported	ICC 0.96 at rest, 0.78 at maximal rate
Schweizer and Gilgen-Ammann [23]	Verity Sense, arm and forearm; watch, wrist	16	Polar H10	9 activities, lying to interval training	device	MAPE 1.35% arm band, 6-7% watch
Moghaddam et al. [10]	5 monitors, wrist to upper arm	28	Polar H10	rest, warm-up, burpees, treadmill	device	CCC $>$ 0.99 at rest, $<$ 0.50 during burpees
Shao et al. [27]	4 wearables; ear, finger, wrist, neck	20	Polar H10 ECG	several days of free living	common pipeline	MAE 2.30 bpm ear, 8.68 bpm neck

2.2 Performance during activity and motion

Agreement was lower during movement across these studies, although the ordering of conditions did not correspond directly to cardiovascular demand. Percentage error remained low at rest and during graded treadmill walking, but increased by an order of magnitude during thirty seconds of burpees [10]. Walking showed the weakest agreement among the eleven activity categories examined in the free-living study [15]. Across seven wrist-worn devices tested during sitting, walking, running, and cycling, error was lowest during cycling and highest during walking [26].

The conditions associated with the poorest agreement were therefore not necessarily those involving the greatest cardiovascular demand. Instead, they were also conditions involving substantial movement of the arm carrying the sensor. This pattern suggests that cardiovascular demand alone does not explain the observed variation and that arm movement may contribute independently. However, these studies did not vary arm movement and cardiovascular intensity independently, so their designs cannot separate the effects of the two factors.

2.3 Conventional signal conditioning

Before individual beats are detected, the recorded optical signal is typically conditioned. Sources of interference addressed during this stage can be broadly classified as slowly varying baseline components, high-frequency noise, and motion artefact [29]. The first two can generally be attenuated by frequency-selective filtering because their dominant spectral content lies outside the frequency range of interest. Motion artefact is more difficult to separate and is discussed in Section 2.4.

The pulsatile component of a photoplethysmogram is superimposed on a slowly varying baseline influenced by respiration, sympathetic activity, and thermoregulation [6], together with a large quasi-static component associated with the illuminated tissue. Bandpass filtering is commonly used to attenuate these low-frequency components and high-frequency noise. The lower cut-off frequency is often placed near 0.5 Hz, below the typical resting heart-rate range, whereas the choice of upper cut-off depends on the intended analysis. Applications concerned primarily with beat timing can use a narrower frequency range, while analyses of pulse morphology, such as those used for blood-pressure estimation or assessment of arterial stiffness, require a wider band to preserve higher-frequency waveform components [29].

Several filter designs have been used for PPG preprocessing. Butterworth, Chebyshev, and finite impulse response (FIR) filters are among the commonly reported frequency-selective approaches [29]. Other methods include moving-average and median filtering, Savitzky-Golay smoothing, empirical mode decomposition, and wavelet-based processing [29, 30]. The choice of filter depends on the characteristics of the signal and the intended analysis, and no single design has been established as universally optimal for PPG processing.

Comparative studies illustrate this variability. Liang and colleagues evaluated nine filter families across ten filter orders using 219 short PPG recordings graded by experts and reported that a fourth-order Chebyshev II filter achieved the highest performance according to a skewness-based signal-quality index [30]. However, this result was obtained from short segments assessed according to waveform quality and does not establish an optimal filter for all PPG applications.

No standard preprocessing pipeline has been established. A review of 118 studies covering PPG waveform analysis, noise, preprocessing, peak detection, and signal quality concluded that no standardised preprocessing sequence exists [29]. Filter type, order, and frequency band therefore remain study-specific choices and should be considered when results obtained using different processing pipelines are compared.

A further consideration is the phase response of the filter. Forward filtering introduces a temporal delay that can depend on frequency unless a linear-phase design is used. In offline processing, phase distortion can be avoided by applying the filter in the forward and reverse directions, resulting in zero-phase filtering. This procedure, however, requires access to samples occurring after the time point being processed and is therefore not applicable to real-time operation [31, Ch. 5].

After conditioning, individual pulse events must be identified from the PPG signal. Several approaches have been proposed, including zero-crossing methods, local maximum or minimum detection, adaptive thresholding, and machine-learning techniques [29]. Local-extrema methods identify candidate systolic peaks within predefined windows or according to amplitude and temporal constraints, whereas adaptive approaches modify the detection threshold in response to changes in signal amplitude. More elaborate methods may incorporate slope information, transformed representations, or multiple time scales to improve robustness to waveform variability and noise.

One example is the method proposed by Elgendi et al. [32], which uses two moving averages computed over different time scales to identify candidate pulse regions before locating the systolic peak. Signal-quality indices may additionally be used to determine whether individual segments are suitable for further analysis [29].

These conventional preprocessing and pulse-detection methods are most effective when unwanted signal components can be distinguished from the cardiac component by their frequency or waveform characteristics. Motion artefact presents a greater challenge because its spectral content can overlap with that of the cardiac signal.

2.4 Signal processing and motion-artefact mitigation

Motion artefact presents a particular challenge for wearable PPG because its spectral content can overlap with that of the cardiac signal. Conventional frequency-selective filtering is therefore often insufficient to separate the two components,

particularly during periodic activities such as walking or running [11]. Methods developed specifically for motion-artefact mitigation consequently use additional signal characteristics or information from other sensors to distinguish motion-related components from the cardiac pulse.

Accelerometry is commonly used for this purpose because movement recorded by an inertial sensor can provide information about the origin and frequency content of motion-related disturbances in the PPG signal. The TROIKA framework, for example, uses accelerometer signals recorded alongside the optical signal to identify spectral components associated with movement. In recordings from 12 participants running at speeds of up to 15 km/h, the method achieved an average absolute error of 2.34 bpm relative to the reference heart rate [33].

Other approaches to motion-artefact mitigation include adaptive filtering, independent or component-based decomposition, signal reconstruction, and frequency-domain methods that exploit information from simultaneously recorded motion signals. These approaches differ in how the motion-related component is identified and removed, and their performance depends on the characteristics of the activity and the degree of overlap between motion and cardiac frequencies [11].

Studies of this type evaluate the ability of a signal-processing method to recover cardiac information from motion-contaminated PPG. This differs from conventional device-validation studies, in which heart rate is typically generated by a proprietary algorithm. In such studies, the performance of individual processing stages cannot be assessed separately.

2.5 Effects of measurement conditions and participant characteristics

Measurement site and posture have often been examined in dedicated studies rather than as isolated factors within broader validation protocols. Measurement site was compared across six anatomical locations within the same participants [9], and across the wrist, forearm, and upper arm against a chest-strap reference. In the latter study, the upper arm showed better performance than the wrist at every tested intensity, while the forearm performed better than the wrist in all conditions except a brief bout of burpees [10]. A recent free-living study provides further evidence of site-dependent performance. Raw PPG signals recorded simultaneously at the ear, finger, wrist, and neck were processed using the same heart-rate estimation methods and compared against an electrocardiographic reference [27]. Mean absolute error varied across sites, from 2.30 bpm at the ear to 8.68 bpm at the neck. These findings indicate that anatomical placement can influence wearable PPG performance even when the processing method is held constant. However, the observed differences depend on the devices, sites, and protocols used and therefore

do not establish a general ranking of measurement locations.

Posture has been investigated primarily in studies of PPG signal quality rather than in device-validation studies. In more than eleven hundred participants, body posture and the height of the sensor relative to the heart were identified as important determinants of wrist PPG signal quality [13]. Graded-exercise validation protocols may not isolate these effects because posture, arm support, and physical activity can change simultaneously without being analysed as separate factors.

Participant characteristics have also been investigated as potential determinants of wearable heart-rate performance. In a diverse cohort wearing seven wrist-worn devices, measurement error was associated with sex, body mass index, and skin tone, with greater errors reported in men, participants with higher body mass index, and those with darker skin [26]. However, findings regarding skin pigmentation have not been consistent across studies. A later study specifically designed to examine skin tone included 53 participants distributed across the Fitzpatrick scale and assessed pigmentation both visually and by spectrophotometry. It found no significant overall difference in heart-rate accuracy across skin tones, although performance differed significantly between devices and activities [25]. This is consistent with a subsequent pulse-rate meta-analysis, which found no statistically or clinically significant pigmentation-related bias [14]. These findings refer to the resulting heart-rate estimates and do not exclude the possibility that pigmentation influences the underlying optical signal from which they are derived.

2.6 Prior validation of the comparison devices

The two devices used for comparison in the present study have not been validated to the same extent.

The Polar Verity Sense has been evaluated against an ECG-based chest strap in two previous studies. In one study, 16 participants wore the device on the upper arm and forearm during nine activities; the Verity Sense achieved a MAPE of 1.35% and a CCC of 1.00 on the upper arm, and a CCC of 0.98 on the forearm [23]. In the second study, 28 participants wore the device on the forearm alongside five other heart-rate monitors. The Verity Sense showed the lowest MAPE during seated rest and the highest CCC during graded treadmill exercise, while its CCC fell below 0.50 during burpees [10].

No peer-reviewed validation study directly comparing heart-rate estimates from the Shimmer3 GSR+ optical pulse ear-clip with an ECG reference was identified in the literature reviewed here. Instead, the device has been used as a laboratory reference in comparisons with consumer wearables. In the study by Ronca et al. [28], for example, the Shimmer served as the reference device rather than the device under validation, and its own heart-rate accuracy was not independently assessed.

In the present study, both devices are therefore treated as comparison devices rather than as reference standards. They provide two additional optical measurements acquired simultaneously from the same participant, and their agreement with each other is evaluated alongside their agreement with the prototype.

2.7 Research gap and study motivation

Several limitations recur in the literature reviewed above. Performance depends strongly on measurement condition, so a single metric calculated over an entire protocol can conceal poor agreement during specific activities [15]. In most validation studies, the quantity compared with the reference is the heart rate reported directly by the device. Sensor characteristics, measurement site, and proprietary processing are therefore evaluated together, making it difficult to determine which component contributes to disagreement. Movement is also commonly represented indirectly by activity category rather than quantified directly.

Methodological limitations of wearable heart-rate validation have also been identified in previous work [34]. Reference devices are themselves subject to measurement error and are not always independently verified. The authors recommend the use of an additional electrical reference implemented through a different measurement configuration, for example an ECG patch alongside a chest strap, so that periods of disagreement between the reference measurements can be identified before evaluating the device under test. They also highlight the inappropriate use of correlation when agreement is the quantity of interest, the evaluation of devices under conditions that may not reflect their intended use, and incomplete reporting of factors such as sensor placement, strap tightness, and skin type.

The present study addresses several of these limitations through the experimental design and analysis described in Chapter 3. Its main contributions are summarised below.

- **Three optical devices recording simultaneously.** The prototype and two comparison devices are worn simultaneously by the same participant, so each comparison is based on the same underlying cardiac activity. The two comparison devices are also compared with each other, allowing disagreement involving the prototype to be interpreted in relation to the agreement between the comparison devices under the same conditions.

- **Activities labelled from video.** A synchronised video recording provides an independent record of the activity performed by the participant. Agreement can therefore be evaluated within individual conditions rather than only across the complete recording, reducing the possibility that poor performance during a specific activity is obscured by aggregate results.

- **Acceleration recorded alongside PPG on every device.** Movement is quantified continuously, providing an additional way to group the analysis windows independently of predefined activity categories. Activity labels describe what the participant was doing, whereas the movement index quantifies the level of motion measured by the sensor during that activity. These complementary representations allow activities with similar overall physical demands but different levels of local sensor movement to be distinguished.
- **Participant characteristics recorded and used for group comparisons.** Age, sex, skin phototype, and body measurements are recorded for each participant. Agreement is additionally examined according to sex and skin phototype rather than only at the cohort level.
- **One processing pipeline for all three optical signals.** The same filtering, beat-detection, and heart-rate estimation procedures are applied to all three devices, eliminating differences attributable to device-specific heart-rate estimation algorithms. Differences in sensing hardware, anatomical site, and upstream acquisition characteristics nevertheless remain.

Chapter 3

Materials and Methods

Simultaneous PPG recordings were acquired from three wearable devices worn by each participant during a protocol of activities of daily living. Heart-rate estimates were derived from all three optical signals using the same processing pipeline. Persimmon, the prototype evaluated in this study, was compared with the two comparison devices, which were also compared with each other.

This chapter describes the experimental acquisition, signal processing, and statistical analysis. Section 3.1 describes the wearable devices, their positioning on the body, and the experimental protocol. Section 3.2 describes the processing pipeline, including device synchronisation, video-based activity labelling, segmentation, filtering, pulse detection, and heart-rate estimation. Section 3.3 describes the statistical methods used to quantify agreement and estimate its uncertainty.

3.1 Data collection

This section describes the enrolled sample, the wearable devices and their positioning on the body, the experimental protocol, and the generated data.

3.1.1 Participants

Twenty-two healthy volunteers were enrolled, each completing one recording session. Data collection was conducted between May and July 2026 at the Politecnico di Torino facilities. Table 3.1 summarises the demographic and anthropometric characteristics of the participants, Table 3.2 presents the Fitzpatrick skin phototype classification, and Table 3.3 describes the self-reported physical activity levels used in the study.

Table 3.1: Continuous variables are reported as mean ± standard deviation (range). BMI denotes body mass index.

Characteristic	Value
Age (years)	30.6 ± 9.7 (23-63)
Sex	Female: 13, Male: 9
Height (cm)	168.7 ± 8.9 (155-184)
Weight (kg)	67.0 ± 11.7 (48-90)
BMI (kg/m^2)	23.5 ± 3.5 (17.0-31.2)
Arm length (cm)	53.0 ± 3.9 (47-59)
Forearm circumference (cm)	25.2 ± 2.2 (21-30)
Fitzpatrick skin phototype	I: 3, II: 7, III: 5, IV: 7
Physical activity level	Sedentary: 1, Lightly active: 14, Moderately active: 6, Very active: 1
Hand dominance	Right: 21, Left: 1
Smoking	No: 19, Yes: 3
Nationality	Italian: 11, Colombian: 9, Ecuadorian: 1, Iranian: 1

Table 3.2: The six levels of the Fitzpatrick skin phototype scale [22], worded as presented in the questionnaire.

Type	Description
I	Very fair skin, often with freckles; always burns, never tans
II	Fair skin; usually burns, tans minimally
III	Medium skin; sometimes burns, tans uniformly
IV	Olive skin; rarely burns, tans easily
V	Brown skin; very rarely burns, tans very easily
VI	Dark brown or black skin; never burns, tans very easily

Table 3.3: The four physical activity levels included in the questionnaire, worded as presented to participants. MET denotes metabolic equivalent of task.

Level	Description
Sedentary	Minimal movement beyond daily tasks such as washing dishes or light household chores; typically prolonged sitting or desk work
Lightly active	About three hours of exercise per week in addition to daily tasks, such as walking, yoga or dancing
Moderately active	Moderate-intensity activity most days, such as brisk walking, doubles tennis or raking the yard, requiring 3.0-5.9 METs
Very active	Vigorous-intensity activity performed regularly, such as running, jogging or strenuous fitness classes, exceeding 6 METs

The study was approved by the Research Ethics Committee of Politecnico di Torino (CER-Polito; protocol 3441/2026, approved on 20 January 2026). Before each recording session, participants received the study information sheet and provided written informed consent. Participation was voluntary, and participants could withdraw from the study at any time. The study was conducted in accordance with the principles of the Declaration of Helsinki.

3.1.2 Wearable devices

Four wearable devices were used simultaneously during each recording session (Table 3.4; Figure 3.5). Three devices provided the PPG signals used for heart-rate estimation: the prototype sensor (Persimmon), the Polar Verity Sense, and the Shimmer3 GSR+. The fourth device, a Nordic Thingy:52, recorded acceleration and provided the common acquisition reference used for subsequent temporal synchronisation and video alignment.

Only the signals required for the present analysis were considered. PPG and acceleration were acquired from Persimmon, Polar, and Shimmer, while the Nordic Thingy:52 provided acceleration only. The acceleration signals were subsequently used for device synchronisation, transfer of video-derived activity labels, and quantification of movement, as described in Section 3.2.

Persimmon

Persimmon was developed by STMicroelectronics in collaboration with other partners within the PERSIMMON project. The electronics are encapsulated in a translucent silicone body approximately 9 cm in length (Figure 3.1). In the present study, the device was positioned on the forearm, as shown in Figure 3.5.

The prototype can sense multiple inertial and physiological signals. Only a subset of these signals was enabled for the present study, including the optical channels and inertial sensors, all sampled at 128 Hz. Three PPG channels (green, red, and infrared) were recorded, and only the green PPG signal was used in the subsequent analysis.

Persimmon was operated in streaming mode, transmitting data via Bluetooth to Bio2Bit WinApp, a Windows application running on the acquisition laptop. The application was used to enable the required sensors and configure the acquisition settings.



Figure 3.1. The Persimmon prototype, shown against a centimetre ruler for scale.

Polar Verity Sense

The Polar Verity Sense (hereafter referred to as *Polar*) is a commercially available optical heart-rate sensor. In the present study, it was positioned on the proximal forearm and used to record PPG and acceleration at an approximate sampling rate of 55 Hz.

The device transmits data over Bluetooth to the Polar Sensor Logger mobile application, installed on a smartphone [35] (Figure 3.2). The optical stream exposed through the Polar software development kit is documented as raw sensor data expressed on the manufacturer-specific scale, without on-device processing [36].



Figure 3.2. Left: the Polar Verity Sense armband. Source: Polar Electro Oy [37]. Right: the icon of Polar Sensor Logger, the third-party application used to record the device's PPG and acceleration streams [35].

Shimmer3 GSR+

The Shimmer3 GSR+ unit (hereafter referred to as *Shimmer*) is a research-grade sensor platform composed of a main sensing unit and an external optical pulse probe. The main unit, containing the inertial sensors, was worn on the wrist, while the PPG signal was acquired from the earlobe using an optical clip connected to the unit by a 1 m cable (Figure 3.3). PPG and acceleration were sampled at 128 Hz.

The device was configured using the manufacturer's Consensys software, which was also used to download the recorded data after each session. The unit was charged between recording sessions using a multi-slot docking station. The complete experimental equipment is shown in Figure 3.6.



Figure 3.3. The Shimmer3 sensor unit with its optical pulse ear-clip probe. Source: Shimmer Research Ltd. [38].

Nordic Thingy:52

The Nordic Thingy:52 [39] (hereafter referred to as *Nordic*) is a compact multi-sensor development kit (Figure 3.4). In the present study, it was positioned on

the forearm alongside the other devices. Only the triaxial acceleration signal was recorded, at a sampling rate of 64 Hz.

The device streamed acceleration data via Bluetooth Low Energy to the IoT4C-BAN Android application running on a smartphone. The application acts as the data aggregator of the wireless body area network described by Demrozi et al. [40]. During acquisition, the same smartphone simultaneously recorded video of the participant using its integrated camera. The Nordic acceleration stream and the video therefore shared the same acquisition clock.

The Nordic acceleration stream was subsequently used to establish the temporal relationship between the reference video and the synchronised device signals, as described in Section 3.2.2.

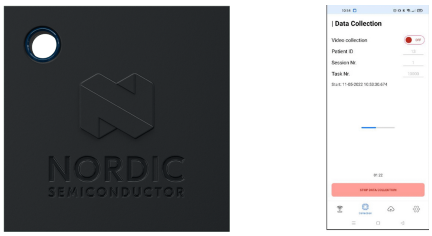


Figure 3.4. Left: the Nordic Thingy:52 in its protective cover [39]. Right: the data-collection screen of the IoT4C-BAN application, which receives the Thingy:52 stream and records the synchronised video. Reproduced from Demrozi et al. [40].

Table 3.4: The four devices used for data collection. The short names in the first column are used throughout the thesis and correspond to those used in the recorded data files. Only the signals used in the present analysis are listed: photoplethysmography (PPG) and acceleration (ACC).

Device	Model	Role	Location	Signals	Sampling
Persimmon	Custom prototype	Prototype	Forearm	PPG, ACC	128 Hz
Polar	Verity Sense	Comparison	Proximal forearm	PPG, ACC	55 Hz
Shimmer	Shimmer3 GSR+	Comparison	Wrist (unit), earlobe (PPG)	PPG, ACC	128 Hz
Nordic	Thingy:52	Synchronisation	Forearm	ACC	64 Hz

3.1.3 Experimental protocol

Each recording session followed a predefined sequence of activities of daily living and lasted approximately 5-6 minutes.

Sensor positioning

The four device units were positioned along the same arm, from the proximal forearm to the wrist. Polar was placed most proximally, near the elbow, and the Shimmer main unit most distally, at the wrist, with Persimmon and Nordic positioned between them. The Shimmer optical probe was the only sensing element located outside the arm and was attached to the earlobe through a 1 m cable connected to the main unit (Figure 3.5). The same general arrangement was maintained for all participants, although exact positions varied slightly according to arm length and strap adjustment.

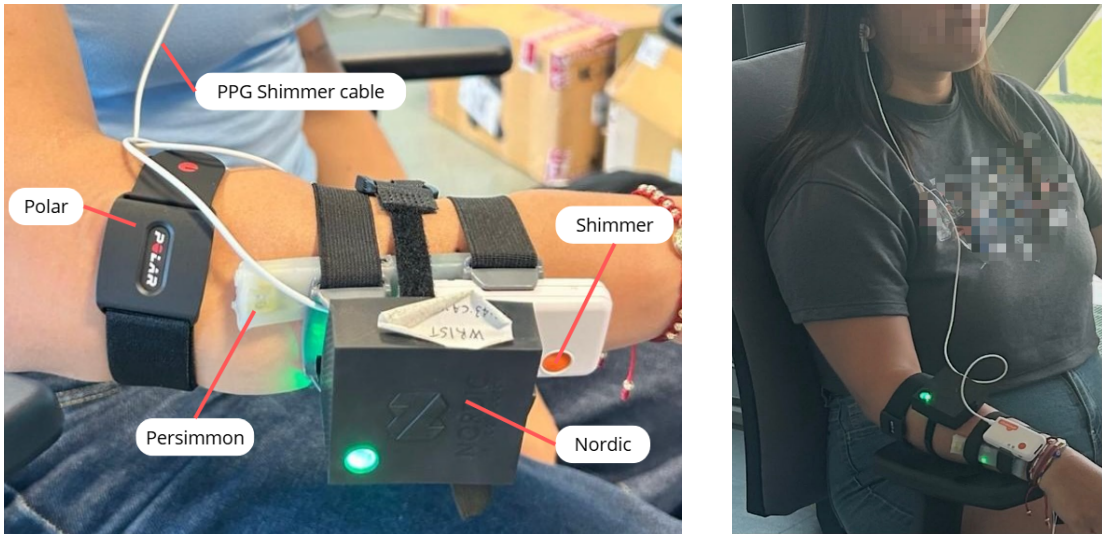


Figure 3.5. Left: placement of the four device units along the arm, from Polar near the elbow to the Shimmer unit at the wrist. Right: the same arrangement on a seated participant, with the Shimmer optical probe attached to the earlobe.

Figure 3.6 shows the complete set of equipment used during data collection. Two laptops were used to configure and record data from Persimmon and Shimmer, while two smartphones were used to record Polar data and the reference video. The setup also included the Shimmer charging dock, the wearable devices and their straps, the optical ear-clip probe, and the tripod supporting the video-recording smartphone.



Figure 3.6. The complete set of equipment used during a recording session.

Recording stations

Figure 3.7 shows the acquisition station used during data collection. Persimmon streamed data to the Bio2Bit WinApp, which displayed the optical and inertial channels in real time and allowed the operator to verify signal acquisition before the protocol began. A second laptop ran the Shimmer Consensys software, which was used to configure the device and monitor its recorded signals. Polar PPG and acceleration were recorded using the Polar Sensor Logger application. Nordic acceleration was streamed via Bluetooth to the smartphone mounted on the tripod, which simultaneously recorded the reference video.

The files generated at the end of each recording session are described in Section 3.1.4.



Figure 3.7. The acquisition station during a recording session. The software used for each device is indicated: Bio2Bit WinApp for Persimmon, Polar Sensor Logger for Polar, and Consensys for Shimmer.

Recording environment and video reference

Recordings were performed indoors under controlled laboratory conditions. The participant was seated at a desk facing the acquisition station (Figure 3.8). A smartphone mounted on a tripod in front of the participant recorded the session through the IoT4C-BAN application, which associated each recording with the participant identifier, session number, and task number. The resulting video provided the reference for subsequent activity labelling, as described in Section 3.2.3.



Figure 3.8. A recording session in progress. The participant is seated at the desk wearing all four devices, with the Shimmer ear-clip cable connected to the earlobe. The smartphone mounted on the tripod records the reference video used for subsequent activity labelling.

Three everyday objects were used during the seated tasks (Figure 3.9): a laptop, a smartphone, and a water bottle. These objects correspond to the `using the computer`, `using the phone`, and `drinking water` activities, respectively.



Figure 3.9. The laptop, smartphone, and water bottle used during the seated activity tasks.

Activity sequence

The protocol alternated short activities of daily living with periods of rest, such that each seated activity was preceded and followed by a resting interval. Synchronisation gestures were performed at the beginning and end of the recording, and a block of postural transitions was included between the seated activities

and the walking task. Figure 3.10 shows the complete sequence, while Table 3.5 summarises the activity labels and nominal durations.

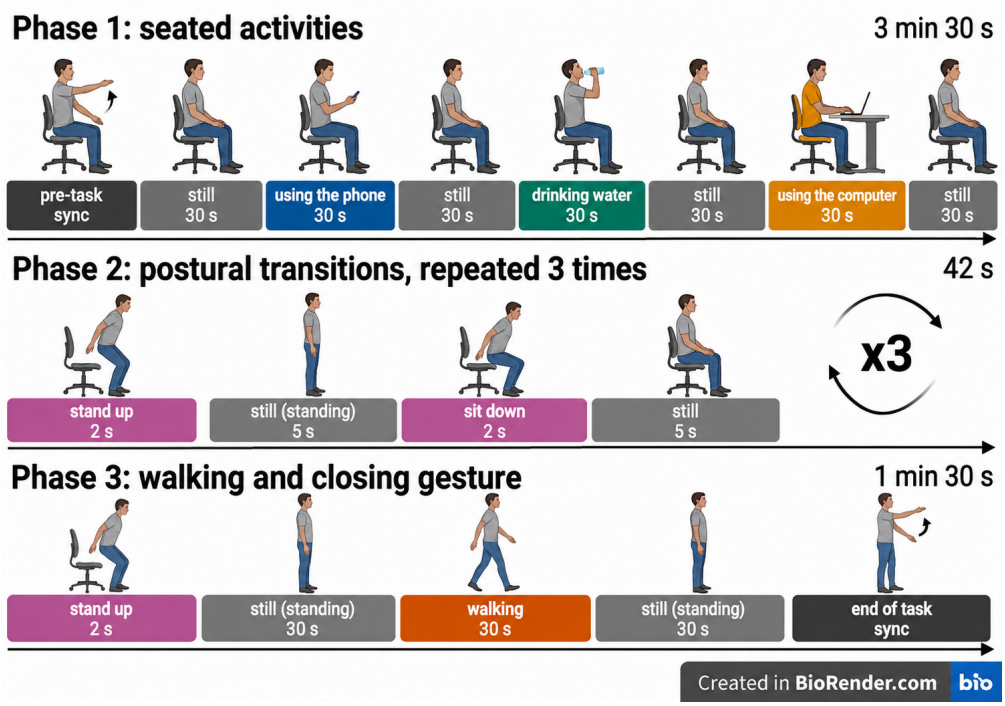


Figure 3.10. The activity protocol divided into three phases. Each step indicates the posture, the label used in the recorded data, and its nominal duration. Grey steps represent resting periods, coloured steps represent active tasks, and the two dark steps represent the synchronisation gestures.

The protocol began with the `pre-task` synchronisation gesture. The participant extended one arm forward, raised and lowered it three times, paused, and then repeated the same sequence. This movement produced a distinct acceleration event recorded by all four devices and later used for temporal alignment (Section 3.2.2). A seated rest period followed, after which three seated tasks were performed, each separated by an additional rest period: using a smartphone, opening a bottle and miming drinking from it, and using a laptop.

A block of postural transitions was then performed. The participant stood up, remained standing briefly, returned to the seated position, and rested; this cycle was repeated three times. The participant subsequently stood once more, remained standing, walked at a self-selected pace, and finished the walking task in the standing position. The recording ended with the `end of task` synchronisation gesture, consisting of the same arm movement performed at the beginning of the session. The two synchronisation events were used subsequently to assess temporal

alignment and clock drift, as described in Section 3.2.2. The `pre-task` and `end of task` intervals were used exclusively for synchronisation and drift assessment and were excluded from all heart-rate agreement analyses.

Table 3.5: Activity labels and nominal durations corresponding to Figure 3.10. The synchronisation gestures had no predefined duration.

Label	Nominal duration	Description
`pre-task`	-	Arm extended forward, raised and lowered three times, followed by a pause and repetition of the same sequence
`still`	30 s	Seated at rest between tasks
`using the phone`	30 s	Seated while using a smartphone
`drinking water`	30 s	Seated while opening a bottle and miming drinking
`using the computer`	30 s	Seated while using a laptop placed in front of the participant
`stand up`	2 s	Transition from seated to standing
`sit down`	2 s	Transition from standing to seated
`still (standing)`	5 / 30 s	Standing at rest, briefly within each stand-up/sit-down cycle and for a longer period before and after walking
`walking`	30 s	Walking at a self-selected pace
`end of task`	-	Arm extended forward and raised and lowered three times for the final synchronisation event

3.1.4 Generated data

Each recording session generated data from the four wearable devices together with one synchronised video recording. Because the acquisition systems operated independently, the resulting files differed in format, structure, and time base. Table 3.6 summarises the raw output generated by each acquisition source. All subsequent processing and quality-control procedures are described in Section 3.2.

Table 3.6: Raw output data format from each device and the respective standardization steps.

Source	Recorded output	Format
Persimmon	One binary file per recorded channel, generated by Bio2Bit WinApp	Binary
Polar	Separate files containing the optical and acceleration streams, generated by Polar Sensor Logger	Text
Shimmer	One tabular recording exported through Consensys	CSV
Nordic	Sequential files containing the recorded acceleration stream	Text
Video	One video recording acquired by the same smartphone application used for the Nordic data	MP4

3.2 Preprocessing

The four devices generated different file formats, operated on independent clocks, and contained no activity annotations. Several processing steps were therefore required before heart-rate estimation and statistical analysis. The raw files were first standardised and organised into a common data structure. The independent device clocks were then synchronised, and activity labels were assigned using the reference video. Finally, the optical signals were filtered, systolic peaks were detected, and heart-rate estimates were generated.

3.2.1 Data standardisation and organisation

The raw recordings generated by the four acquisition systems required different preparation procedures before they could be processed jointly.

The Shimmer recording required no format conversion. The Consensys software exported a CSV file in which all recorded signals already shared a common time base.

Persimmon streamed data to the Bio2Bit WinApp, which stored each recorded channel in a separate binary file. Because the individual channels did not share exactly the same time base, each channel was read using its corresponding timestamps and resampled onto a common 128 Hz grid. Only the temporal interval shared by all required channels was retained. The inertial channels were converted to physical units, while the optical channels were retained on their original

dimensionless scale.

Polar data were recorded by the Polar Sensor Logger application as separate text files containing one row per sample. The optical and acceleration signals were stored independently. During standardisation, the PPG channel used for analysis was selected, and the optical and inertial signals were organised using a common timestamp so that the Polar streams shared a single device-specific time base.

For the Nordic Thingy:52, the receiving application generated a new text file approximately every second. These files were concatenated, and a continuous time base was reconstructed from the recorded sample timestamps.

Acceleration units differed among devices. Persimmon raw acceleration was converted to m/s^2, Shimmer exported acceleration directly in m/s^2, and Polar reported acceleration in milli-g. These differences did not require cross-device amplitude standardisation because acceleration was used to identify the timing of the synchronisation gesture and, for Persimmon, to construct the motion index. Absolute acceleration amplitudes were therefore not compared across devices.

The optical channels were also retained on their device-specific scales because no common optical amplitude unit was available. Shimmer exported its optical channel in millivolts, whereas Persimmon and Polar reported detector values on their respective internal scales. Optical amplitude additionally depends on device-specific characteristics such as detector gain and emitter current and was therefore not compared across devices. This did not affect heart-rate estimation, which was based on the temporal position of the detected systolic peaks rather than on comparison of absolute PPG amplitudes. Figures containing optical signals from more than one device were therefore normalised for visualisation.

At the end of the standardisation stage, each device was represented by one table per recording containing the signals of interest referenced to that device's own time base. The four device clocks remained independent and were subsequently aligned as described below.

3.2.2 Synchronisation and resampling

Four independent device clocks operated during each recording session, and the recordings were started manually several seconds apart. Temporal alignment was therefore required before the signals could be analysed jointly.

The arm movement performed at the beginning of the protocol provided the primary synchronisation event. Because all four device units were attached to the same arm, the movement produced a distinct acceleration transient in each device at the same physical instant. The timing of this event was used to estimate the offset between device clocks.

Figure 3.11 shows the acceleration norm calculated for the four devices over an entire recording and highlights the intervals containing the initial and final arm

movements. Figure 3.12 shows the initial gesture in greater detail.

Persimmon was used as the reference time base. For each of the remaining devices, the time of the selected acceleration peak was compared with the corresponding peak in Persimmon. The resulting temporal offset was applied to the complete timeline of that device. Figure 3.13 shows the acceleration traces after this alignment. The recordings were then restricted to the temporal interval for which data from all required devices were available.

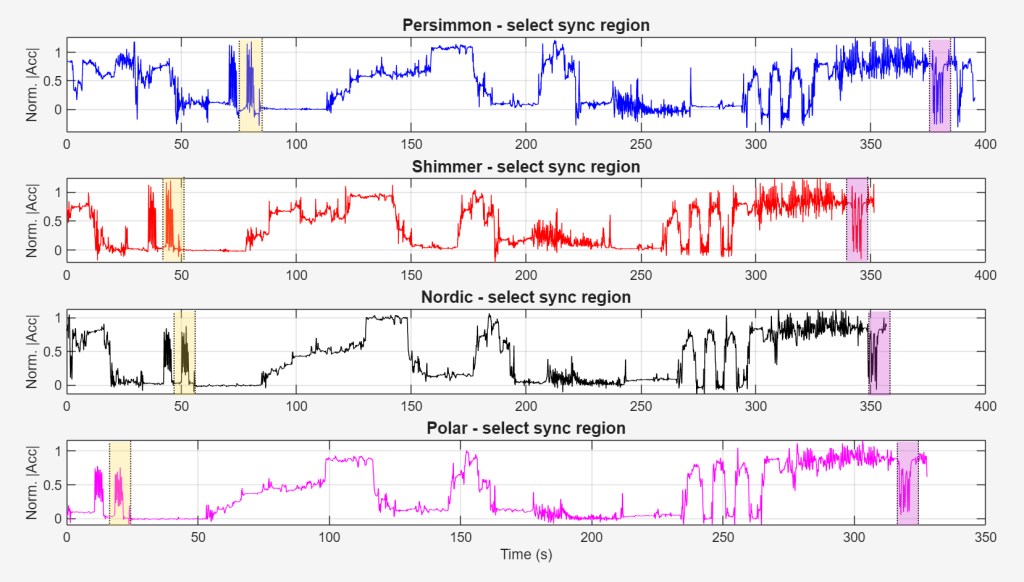


Figure 3.11. Acceleration norm from each device over an entire recording before temporal alignment. The opening synchronisation gesture is shaded in yellow and the closing gesture in purple. The latter was used for the residual drift assessment shown in Figure 3.14.

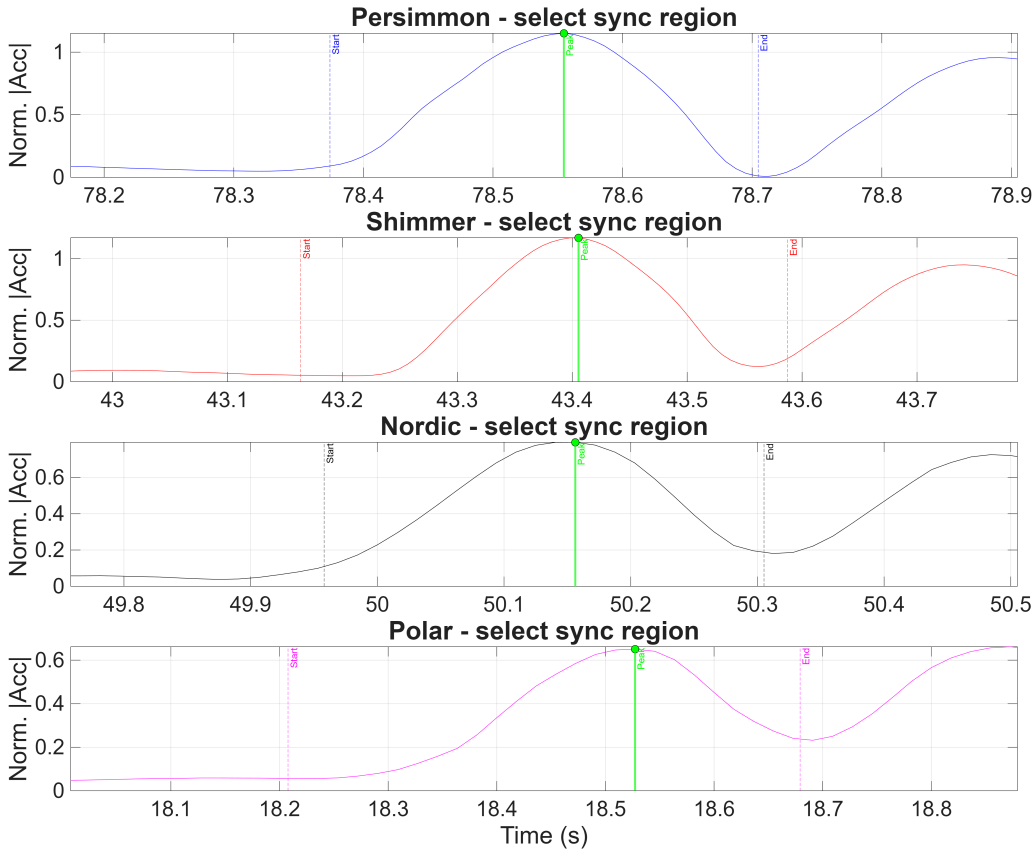


Figure 3.12. Acceleration from the four devices during the opening synchronisation gesture in one recording. The region marked Start and End defines the gesture interval, and the largest sample within this interval, marked Peak, was selected as the synchronisation instant.

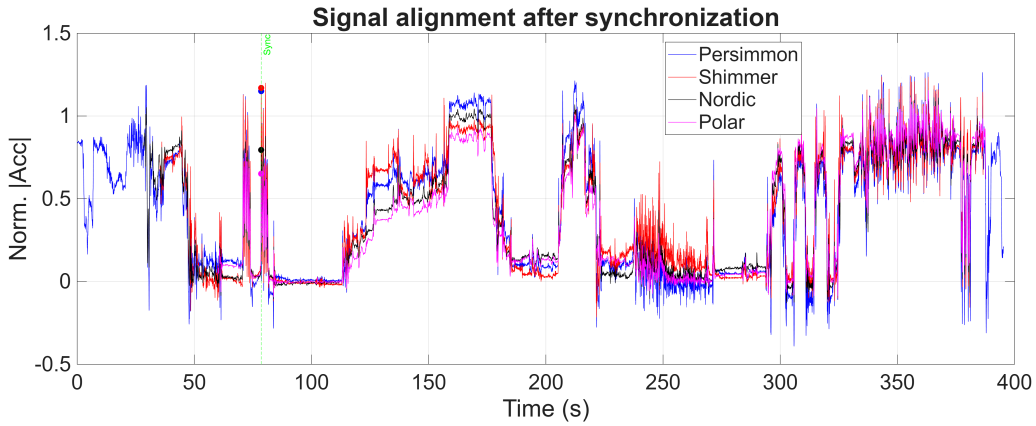


Figure 3.13. Acceleration from the four devices after temporal alignment to the Persimmon time base. The green marker indicates the synchronisation instant.

Correction of a constant temporal offset does not account for possible clock drift caused by differences in the effective clock rates of the devices. The final arm movement was therefore used to quantify residual temporal drift at the end of each recording (Figure 3.14). The processing script flagged any recording in which the residual difference exceeded 0.10 s.

Because heart rate was subsequently estimated in consecutive 12 s windows, a residual drift of 0.10 s corresponds to less than 1% of the analysis-window duration and was considered negligible relative to the temporal resolution of the activity-specific analysis.

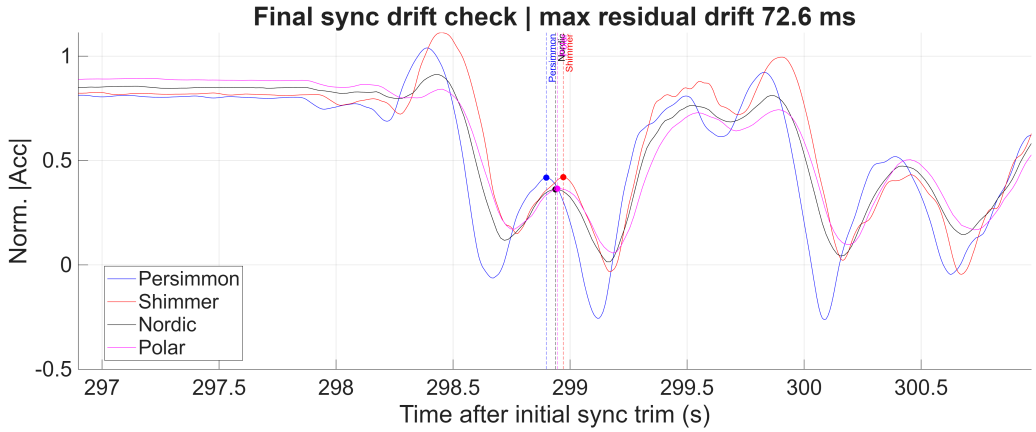


Figure 3.14. Residual temporal drift assessed using the closing synchronisation gesture. Any remaining difference in the timing of this event represents drift accumulated after alignment using the opening gesture.

After temporal alignment, the devices shared a common time reference but retained their original sampling rates, ranging from approximately 55 Hz for Polar to 128 Hz for Persimmon and Shimmer. All non-Persimmon signals were therefore linearly interpolated onto the Persimmon post-alignment time grid. The resulting synchronised dataset contained one common temporal sample grid for all devices. No extrapolation beyond the original recorded range of any device was performed.

3.2.3 Activity annotation

Following device synchronisation, the video recording was aligned with the common time base through the Nordic acceleration stream. Because the Nordic data and the video were recorded by the same smartphone application, they shared a common acquisition clock. The synchronisation event previously identified in the Nordic acceleration therefore provided the temporal link between the video and the synchronised device signals.

An operator reviewed each recording using a custom graphical user interface (Figure 3.15) and marked the beginning and end of each activity or task using two key presses. All samples between the selected start and end times were assigned the corresponding activity label. Intervals were annotated sequentially until the protocol had been covered, while samples outside the annotated intervals remained unlabelled. Each recording was restricted to the interval between the initial and final synchronisation gestures.

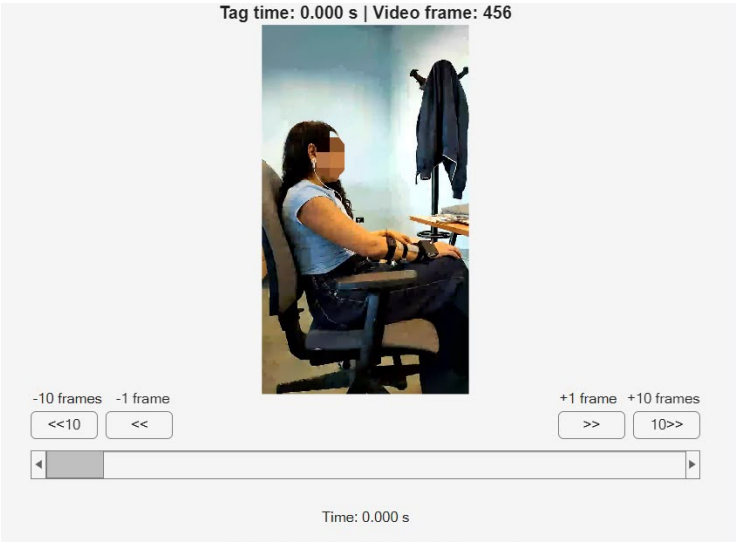


Figure 3.15. Custom graphical user interface used to assign activity labels from the reference video.

3.2.4 Filtering

The optical signals were converted into a sequence of systolic pulse times in two stages: bandpass filtering, described in this section, followed by peak detection, described in Section 3.2.5. The same processing parameters were applied to the PPG signals from all three devices.

A PPG waveform consists of a pulsatile component synchronous with the cardiac cycle superimposed on a slowly varying baseline influenced by respiration, sympathetic activity, and thermoregulation [6]. Because the present analysis required pulse timing rather than detailed waveform morphology, the signals were bandpass filtered between 0.5 and 3.5 Hz, corresponding to 30-210 beats per minute.

A linear-phase finite impulse response (FIR) filter with 513 coefficients was used. Each output sample was calculated from the current input sample and the preceding 512 samples. Because the ideal impulse response of the selected bandpass is infinitely long, it was truncated to 513 coefficients. A Hamming window was applied to taper the truncated impulse response and reduce ripple in the resulting frequency response [31, Ch. 7].

The symmetric coefficients of the FIR design provide a linear phase response within the passband. Consequently, all frequency components within the passband experience the same temporal delay, preserving the temporal shape of the pulse apart from the fixed delay.

The passband was selected according to the range of heart rates expected during the experimental protocol rather than according to preservation of detailed PPG morphology. The lower cut-off frequency of 0.5 Hz lies below the expected resting heart rates in the study cohort and attenuates low-frequency baseline variation. A lower cut-off near 0.5 Hz is also commonly used in PPG preprocessing to attenuate the direct-current component and slow baseline fluctuations [29]. The upper cut-off of 3.5 Hz corresponds to 210 bpm and lies above the age-predicted maximum heart rate of every participant, approximately 192 bpm for the youngest participant [3]. Comparable frequency ranges have been used in previous studies estimating heart rate from wrist PPG [33, 41].

This bandwidth attenuates higher-frequency components associated with detailed pulse morphology, which may extend into the third and fourth harmonics of the fundamental cardiac frequency [29]. These components were not required for the present analysis because heart rate was estimated from pulse timing rather than waveform morphology. Restricting the passband also reduced the amount of out-of-band noise reaching the peak detector.

The FIR filter was applied causally using only current and past input samples, rather than using a forward-backward zero-phase implementation. Zero-phase

forward-backward filtering requires samples occurring after the time being processed and is therefore not directly applicable to real-time operation [31, Ch. 5]. The causal implementation was selected to reproduce the conditions under which the filtering stage could operate on a wearable device receiving samples sequentially.

Filtering was performed in blocks containing 2 s of new samples, while the internal filter state was retained between consecutive blocks. The filter therefore operated continuously across the complete recording rather than being reinitialised for each block.

The filter length was determined primarily by the lower edge of the passband. A frequency of 0.5 Hz has a period of 2 s, and the filter used here spans approximately two cycles, corresponding to 4 s. At 128 Hz, this resulted in 513 coefficients. Each output sample was therefore determined by approximately the preceding 4 s of input data.

The symmetric FIR coefficients produce a group delay of 256 samples. At 128 Hz, this corresponds to 2 s. Because the phase response is linear, the same delay is applied to all frequencies within the passband [31, Ch. 5].

The filtered output was shifted by 256 samples to compensate for this known group delay and restore each pulse to its original timestamp. This correction does not remove the corresponding real-time latency: the filtered value associated with a given time instant becomes available only after the required samples have been acquired. A real-time implementation of this stage would therefore report a detected pulse approximately 2 s after its occurrence while retaining the correct timestamp.

Heart-rate estimation was itself based on a retrospective 12 s interval (Section 3.2.6). The filtering latency therefore represents an additional 2 s delay in the availability of an estimate, although it does not alter the temporal interval to which the estimate refers.

Before filtering, a constant offset estimated from the first 5 s of each optical channel was subtracted. The optical signals contained baseline levels substantially larger than the pulsatile component. Although the bandpass filter strongly attenuates a constant component, residual offset can affect the subsequent detector because its input is clipped at zero. Initial offset removal also reduced the transient produced when filtering began from a zero initial state. The same procedure was applied to every recording and every optical channel, and only samples available at the beginning of the recording were used.

Figure 3.16 shows the combined effect of offset removal and bandpass filtering for one interval of a Persimmon recording.

An alternative Butterworth bandpass filter was also evaluated. Butterworth filters are commonly used for PPG preprocessing and were used in the study that introduced the peak detector adopted here [29, 32]. Unlike the FIR design,

the Butterworth filter is an infinite impulse response (IIR) filter whose outputs depend on previous inputs and previous outputs. It can achieve a given frequency selectivity using substantially fewer coefficients, but its phase response is nonlinear.

The FIR and Butterworth filters were compared over the complete dataset using the same passband, peak detector, and heart-rate estimation procedure. Their frequency responses differed substantially immediately above the passband. At 4.5 Hz, a fourth-order-per-edge Butterworth filter retained approximately 28% of the input amplitude, whereas the FIR filter used in the final pipeline retained approximately 0.09%. Consequently, more spectral energy immediately above the selected PPG band reached the detector when the Butterworth filter was used.

The linear-phase FIR filter produced higher inter-device agreement and lower inter-device MAE and was therefore selected for the final processing pipeline. The comparison is reported in Section 4.2, and all subsequent results were obtained using the FIR filter.

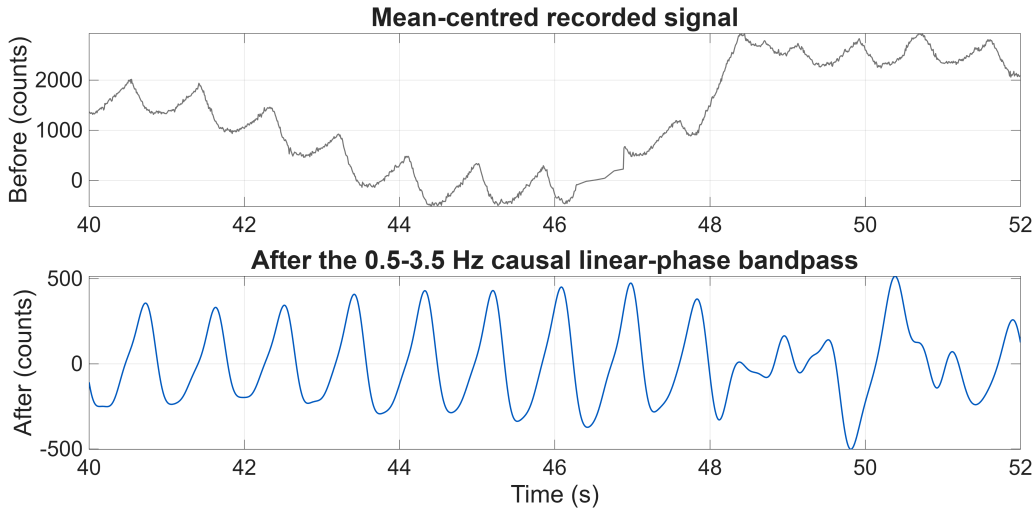


Figure 3.16. One interval of a Persimmon recording before and after filtering. Top: the mean-centred recorded signal, containing the pulsatile component and baseline variation. Bottom: the bandpass-filtered signal used for subsequent systolic peak detection.

3.2.5 Beat detection

Systolic peaks were identified from the filtered PPG signal using the method proposed by Elgendi et al. [32]. The signal was first clipped to its positive component and squared, suppressing negative portions of the waveform and emphasising positive peaks.

Two moving averages were then computed from the squared signal using windows of 111 ms and 667 ms. The shorter window approximately corresponds to the duration of an individual systolic peak, whereas the longer window represents a time scale comparable to a cardiac cycle. Figure 3.17 illustrates the successive processing stages over 10 s of one recording.

The short moving average increases in the presence of a systolic peak, whereas the longer moving average represents the local signal level. A candidate region began when the short moving average exceeded the long moving average plus an offset and ended when it fell below this threshold. The offset was defined as 2% of the mean squared-signal amplitude calculated over the preceding 2 s. This additional threshold reduced the probability that low-amplitude fluctuations between pulses generated candidate regions.

Candidate regions shorter than 111 ms were discarded. For each remaining region, the systolic peak was assigned to the largest sample of the filtered PPG signal within that region, rather than to the squared signal used to define the candidate region.

The use of a trailing 2 s interval to determine the threshold offset represents one modification of the method described by Elgendi et al. [32]. In the original method, the offset is calculated from the mean squared-signal amplitude over the complete recording. In the present protocol, signal amplitude can change substantially between seated rest and walking. A threshold based on the complete recording can therefore be influenced by higher-amplitude portions of the signal and become disproportionately large during lower-amplitude intervals. The same 2% scaling factor was retained, but the mean was calculated over the preceding 2 s so that the threshold adapted to the local signal amplitude using only previously acquired samples.

Both moving averages were centred on the sample being evaluated. They therefore required samples extending forward by half of their respective window lengths. The longer moving average determined the principal detector delay, approximately one third of a second. Additional delay was introduced because a candidate region had to close before the largest sample within that region could be identified.

A short settling period was excluded from peak detection. The FIR filter contains 513 coefficients, corresponding to approximately 4 s at 128 Hz, and its internal state is initially filled progressively as samples become available. The trailing 2 s interval used to calculate the threshold offset undergoes a similar initial settling period. The first 5 s of filtered output were therefore treated as settling time, and peaks detected within this interval were discarded for all recordings and all sensors. The filter itself operated from the first available sample and maintained its state between blocks, so this settling period occurred only once per recording.

Figure 3.18 shows the resulting processed signals and detected peaks over the opening 30 s of one recording.

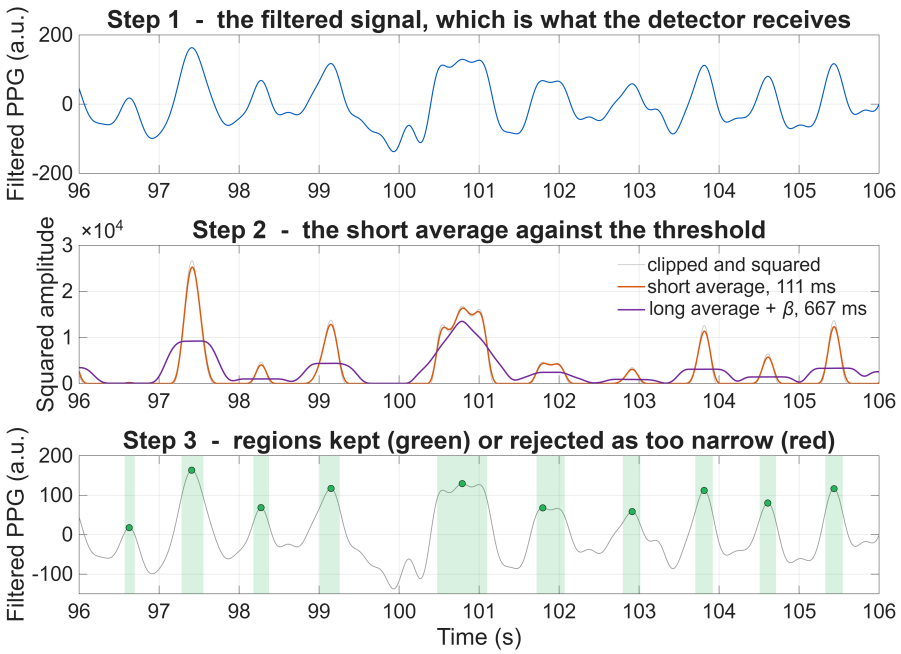


Figure 3.17. Application of the peak-detection method of Elgendi et al. [32] to 10 s of one Persimmon recording. Top: filtered PPG signal. Middle: squared signal together with the short moving average and the long moving average plus offset. Bottom: candidate regions and accepted systolic peaks.

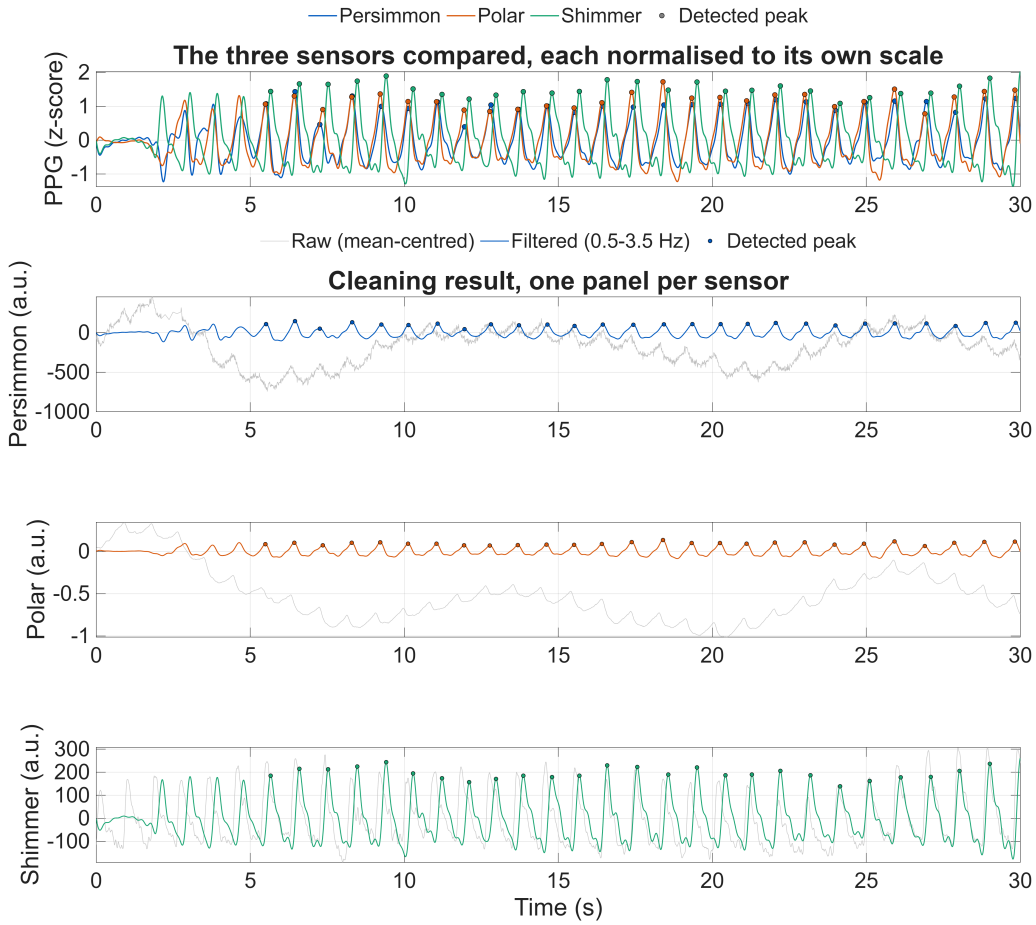


Figure 3.18. Cleaning and peak detection over the opening 30 s of one recording. Top: filtered signals from the three PPG devices, each normalised to its own scale, with detected peaks indicated. Bottom: one device per panel, with the mean-centred raw signal shown behind the corresponding filtered signal.

3.2.6 Heart-rate estimation

Heart rate was estimated using sliding windows of 12 s with a step of 2 s, corresponding to approximately 83% overlap between consecutive windows. Within each window, the interval between every pair of consecutive systolic peaks defined an inter-beat interval (IBI). Each IBI was converted to an instantaneous heart-rate estimate as 60/IBI, with the IBI expressed in seconds. Intervals corresponding to heart-rate values outside the accepted physiological range were discarded, and the heart rate assigned to each window was defined as the median of the remaining instantaneous estimates. Deriving pulse-to-pulse timing from consecutive systolic peaks is an established approach for estimating heart rate from PPG [29].

A missed peak generally produces an IBI longer than expected, whereas an additional detected peak produces an abnormally short IBI. The corresponding instantaneous heart-rate estimates can therefore become substantially lower or higher than the underlying pulse rate. The median is less sensitive than the mean to a small number of such outliers and was therefore used to represent each window.

The accepted range was 30-210 beats per minute. Values outside this range could still occur after bandpass filtering because filtering constrains the frequency content of the waveform rather than the temporal spacing of the detected peaks. For example, a missed pulse can produce an abnormally long IBI, whereas a distorted waveform can generate more than one detected peak within a cardiac cycle and produce an abnormally short IBI. Under the present protocol, values outside the selected range were considered more likely to represent detection errors than physiological heart rates. The range was intentionally broad enough to include the expected heart rates of the study cohort. Physiological range filtering therefore removed only clearly implausible intervals, while the median reduced the influence of remaining outlying estimates.

The 12 s window length was selected according to two requirements. First, each window had to contain enough intervals for the median estimate to remain stable in the presence of a small number of peak-detection errors. Second, the duration had to remain sufficiently short relative to the approximately 30 s activity periods used in the protocol. At a typical resting heart rate, a 12 s window contains approximately a dozen cardiac cycles. The 2 s step generated a new estimate every 2 s and provided sufficient temporal resolution to characterise changes during the protocol.

A heart-rate estimate was reported only when at least three valid IBIs remained after physiological range filtering. Otherwise, the corresponding window was left without a heart-rate estimate. This criterion represented a minimum data requirement rather than a signal-quality classifier. In the exported tables, this condition was represented by the sensor-specific `Usable` flag; a usable window therefore corresponded to a window for which that sensor produced a heart-rate estimate.

Alternative procedures for obtaining one heart-rate estimate from each window were evaluated over the complete dataset while keeping all other processing stages unchanged. These alternatives included removing the highest and lowest instantaneous heart-rate values before averaging, rejecting IBIs that differed by more than 30% from the median IBI of the window, and requiring six valid intervals instead of three. None provided an improvement in the overall balance between inter-device agreement and data coverage, and the median-based procedure with a minimum of three valid intervals was therefore retained.

Figure 3.19 shows the resulting heart-rate series for one recording, together with the corresponding activity labels.

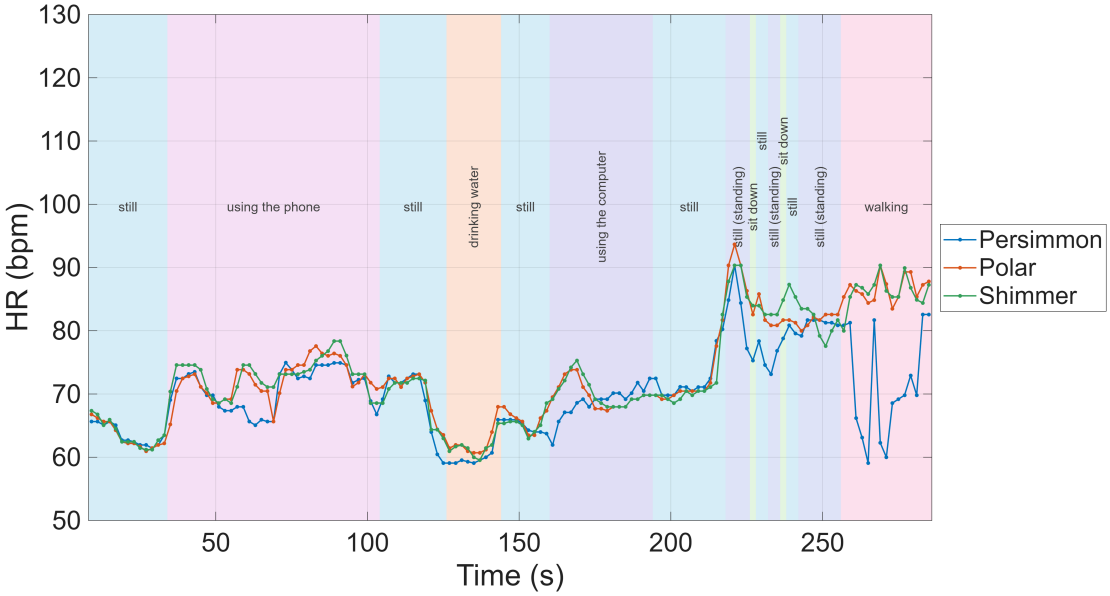


Figure 3.19. Heart-rate estimates obtained for one recording, with one series per PPG device and the corresponding activity intervals shown in the background.

3.2.7 Data-quality review and exclusions

Four recordings were associated with acquisition or data-quality issues that affected their use in the subsequent agreement analyses. Sensor-specific exclusions were applied only to the affected device within a recording; data from the remaining devices were retained when applicable.

In `s7`, the Shimmer ear-clip probe was incorrectly positioned during acquisition, and the Shimmer data from this recording were excluded. In `s14`, inspection of the Polar recording indicated incorrect sensor positioning together with an abnormal noise pattern, and the corresponding Polar data were excluded. In `s18`, the Persimmon strap was visibly loose throughout the session video, and the Persimmon data from this recording were excluded.

In `s22`, communication errors affected the Nordic recording and prevented reliable alignment of the activity labels with the physiological signals. The recording was therefore excluded only from activity-specific analyses and was retained for analyses that did not require activity labels.

Table 3.7: Recording-level acquisition and data-quality issues identified before the agreement analyses. Windows denotes the number of 12 s heart-rate windows affected in each recording. Sensor-specific exclusions apply only to the affected device, whereas the `s22_t1` exclusion applies only to analyses requiring activity labels.

Recording	Affected data	Reason	Windows
`s7_t1`	Shimmer	Ear-clip probe incorrectly positioned during acquisition	156
`s14_t1`	Polar	Incorrect sensor positioning associated with an abnormal noise pattern in the recorded signal	150
`s18_t1`	Persimmon	Strap visibly loose throughout the session video	150
`s22_t1`	Activity labels	Nordic communication errors prevented reliable activity labelling; excluded only from activity-specific analyses	157

3.3 Statistical analysis

Statistical analysis was performed on heart-rate estimates obtained from the three PPG devices over the same analysis windows. Because no electrocardiographic reference was available, the analysis quantified inter-device agreement rather than accuracy relative to a reference heart rate.

Agreement was evaluated at three levels: across all usable windows pooled across participants, within individual recordings, and within subgroups defined by activity, sex, Fitzpatrick skin phototype, and motion level. Sex and skin phototype were characteristics of the participants and were additionally examined using between-group statistical comparisons.

The usable-window criterion was applied separately to each analysis. For a pairwise comparison, a window was included only when both devices in the pair provided a `Usable` heart-rate estimate. For the three-device ICC, a window was included only when all three devices provided `Usable` estimates. A window could therefore contribute to one comparison but not to another, depending on the estimates available after preprocessing and sensor-specific exclusions.

3.3.1 Agreement metrics

Pairwise agreement between devices was quantified using Lin's concordance correlation coefficient (CCC) [19]. The CCC was calculated as

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \tag{3.1}$$

where s_x^2 and s_y^2 are the variances of the two heart-rate series, s_{xy} is their covariance, and $\bar{x}$ and $\bar{y}$ are their respective means. Unlike Pearson's correlation coefficient, the CCC accounts for both covariation and departure from the line of identity and was therefore used as the principal measure of pairwise agreement.

For descriptive interpretation, concordance values were classified as poor below 0.50, moderate from 0.50 to below 0.75, good from 0.75 to 0.90, and excellent above 0.90. These thresholds correspond to the categories proposed by Koo and Li for ICC interpretation [42] and were also applied descriptively to the CCC results to maintain a consistent terminology across agreement coefficients. Interpretation considered the confidence interval(CI) in addition to the point estimate. These qualitative categories describe the strength of agreement and were not treated as criteria of practical acceptability, which was evaluated separately using bpm-based difference measures.

Agreement across all three devices was quantified using an intraclass correlation coefficient (ICC). Whereas the CCC provides pairwise agreement, the ICC provides a single coefficient describing agreement among the three devices simultaneously. The coefficient was estimated using the two-way model described by McGraw and Wong [20], with analysis windows represented by rows and devices by columns:

$$\text{ICC} = \frac{BMS - EMS}{BMS + (k-1)EMS + \frac{k}{n}(JMS - EMS)} \tag{3.2}$$

where BMS is the between-window mean square, JMS is the between-device mean square, EMS is the residual mean square, n is the number of windows, and $k = 3$ is the number of devices. The selected form was ICC(A,1), corresponding to single-measure absolute agreement. Consequently, systematic differences between devices contributed to disagreement rather than being ignored.

The three devices represented the specific measurement systems evaluated in the study rather than a random sample from a wider population of devices. The ICC therefore characterises agreement among this specific set of devices. Its values were interpreted using the categories described above and in conjunction with their CIs.

In addition to CCC, several pairwise difference and association metrics were reported for each device pair: mean absolute error (MAE), root mean square error (RMSE), mean signed difference, Bland-Altman limits of agreement (LoA), Pearson's correlation coefficient, the percentages of windows within 5 and 10 bpm, and mean absolute percentage error (MAPE).

For a given pair, let p_i and r_i denote the heart-rate estimates obtained for window i. Persimmon was defined as the first device when included in the comparison,

while Polar was defined as the first device for the Polar-Shimmer comparison. The signed difference was therefore

$$d_i = p_i - r_i,$$

and the pairwise mean used in the Bland-Altman analysis was

$$a_i = \frac{p_i + r_i}{2}.$$

A negative mean signed difference therefore indicates that the first-named device provided lower heart-rate estimates, on average, than the second-named device.

The pairwise metrics are defined in Table 3.8. Here, n is the number of windows included in a given comparison, $SD(d_i)$ is the standard deviation of the signed differences, T is the heart-rate tolerance in bpm, and $\mathbf{1}(\cdot)$ is the indicator function, equal to one when its argument is true and zero otherwise. For Pearson's coefficient, s_{pr} denotes the covariance between the two series, and s_p and s_r their standard deviations.

Table 3.8: Pairwise difference and association metrics calculated from the heart-rate estimates of each device pair.

Metric	Formula	Interpretation
MAE	$\frac{\sum_{i=1}^{n} \lvert d_i \rvert}{n}$	Mean absolute difference in bpm
RMSE	$\sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n}}$	Root mean square difference in bpm
Bias	$\frac{\sum_{i=1}^{n} d_i}{n}$	Mean signed difference in bpm
LoA	$\frac{\sum_{i=1}^{n} d_i}{n} \pm 1.96\, SD(d_i)$	Bland-Altman limits of agreement
Pearson's r	$\frac{s_{pr}}{s_p s_r}$	Linear association between devices
$P_{\leq T}$	$100 \times \frac{\sum_{i=1}^{n} \mathbf{1}(\lvert d_i \rvert \leq T)}{n}$	Percentage of windows differing by no more than T bpm
MAPE	$100 \times \frac{\sum_{i=1}^{n} \lvert d_i \rvert / r_i}{n}$	Mean absolute percentage difference relative to the second-named device

3.3.2 Pooled analysis

All usable windows were pooled across recordings. The agreement and difference metrics described above were then calculated over the complete set of eligible windows: one CCC and one set of pairwise metrics for each device pair, together with one ICC across the three devices.

For each pair of devices, a Bland-Altman plot was generated [18]. The pairwise mean a_i was represented on the horizontal axis and the signed difference d_i on the vertical axis. The mean signed difference and limits of agreement summarised the distribution of window-level differences between devices. These conventional limits of agreement do not account for the overlap between consecutive windows or for repeated measurements obtained from the same participant [43].

Windows obtained from the same participant were not statistically independent, and consecutive 12 s windows overlapped by approximately 83%. Confidence intervals were therefore not calculated using window-level standard errors, which would treat the windows as independent and underestimate uncertainty.

CIs for CCC and ICC were instead estimated using a participant-level cluster bootstrap [44]. Participants were sampled with replacement, and all eligible windows belonging to each selected participant were retained together in each bootstrap sample. This procedure preserved the within-participant clustering while estimating uncertainty associated with the participant sample. The three study devices remained fixed throughout the resampling procedure.

A total of 1000 bootstrap samples were generated using a fixed random seed to ensure reproducibility. The 95% CIs were obtained using the percentile method. The same participant-level resampling procedure was applied to MAE, RMSE, mean signed difference, and MAPE.

3.3.3 Per-recording analysis

The same agreement and difference metrics were also calculated separately within each recording using only the windows available for that participant. This produced one recording-level value for each metric and allowed the pooled results to be interpreted relative to the variability observed across participants.

The recording-level values were also used for the participant-group comparisons described in Section 3.3.4. CIs were not calculated for individual-recording metrics because the participant-level bootstrap requires multiple participant clusters and therefore cannot be applied within a single recording.

3.3.4 Subgroup analyses

The same usable-window criterion was applied independently within each subgroup. Activity and motion level were window-level variables, whereas sex and

Fitzpatrick skin phototype were participant-level characteristics. Windows labelled `unknown` were excluded from activity-specific analyses.

Motion was quantified from the Persimmon accelerometer and was used only as a descriptor of movement; it was not used to modify the PPG signal or the heart-rate estimates. The magnitude of the three-axis acceleration vector was first calculated. A local baseline was removed using a 1 s moving median, and the absolute residual was subsequently smoothed using a 0.5 s moving mean.

The motion value associated with each heart-rate window was defined as the mean of this smoothed signal over the corresponding 12 s interval. Motion values were then scaled across the pooled cohort using the 20th and 95th percentiles and clipped to the interval 0-1. Windows were classified as low motion for values below 0.33, medium motion for values from 0.33 to below 0.66, and high motion for values greater than or equal to 0.66.

The categorised motion level was used for the subgroup agreement and difference analyses, while the continuous motion index was used only to assign each window to a motion category. Within each motion level, CCC was calculated for each of the three device pairs, together with the three-device ICC. Mean absolute difference was also calculated for Persimmon relative to Polar and Shimmer, allowing the effect of increasing movement on inter-device agreement to be examined using both concordance- and difference-based measures.

CCC, ICC, and the pairwise difference metrics were recalculated separately within each subgroup. CIs were obtained using the same participant-level cluster-bootstrap procedure applied in the pooled analysis, with resampling restricted to participants contributing data to the subgroup. Subgroups represented by fewer participants or fewer windows consequently produced greater uncertainty and wider CIs.

Formal participant-group comparisons were performed for sex and skin phototype. These variables assign each participant to a single group and therefore permit independent between-participant comparisons of recording-level metrics. Activity and motion level, in contrast, were repeated within participants because the same participants contributed windows to multiple conditions. These analyses were therefore treated descriptively using subgroup-specific estimates and their bootstrap CIs rather than by applying the independent-group test used for sex and phototype.

Sex separated the cohort directly into two groups. Four Fitzpatrick skin phototypes were represented in the sample, with relatively few participants in some individual categories. For the inferential comparison, phototypes I and II were therefore combined into one group and phototypes III and IV into another. Agreement metrics were nevertheless also reported descriptively for each of the four represented phototypes separately.

Between-group comparisons were performed using the Mann-Whitney U test

[45]. This non-parametric test compares the relative ranks of values in two independent groups and does not require an assumption of normally distributed observations. It was selected because the group sizes were small and no assumption of normality was made for the recording-level agreement and difference metrics.

The Mann-Whitney test evaluates whether values from one group tend to be larger or smaller than values from the other group. It was not interpreted specifically as a test of differences in group medians, because such an interpretation requires additional assumptions regarding the shapes of the underlying distributions. Group medians were therefore reported as descriptive summaries rather than as the direct quantity tested.

Five recording-level quantities were evaluated within each participant grouping: the Persimmon-Polar MAE, the Persimmon-Shimmer MAE, the Persimmon-Polar CCC, the Persimmon-Shimmer CCC, and the three-device ICC.

Exact p-values were calculated by enumerating all possible allocations of the observed recording-level values into groups of the observed sizes rather than using a large-sample approximation. Rank-biserial correlation, denoted by r, was reported as an effect-size measure for each comparison.

The rank-biserial correlation describes the direction and magnitude of the between-group difference in ranks. Considering all pairs formed by selecting one recording from each group, r can be interpreted as the difference between the proportion of pairs in which the second group has the higher value and the proportion in which the first group has the higher value. A value of zero indicates no systematic tendency for either group to have higher values, whereas values approaching -1 or $+1$ indicate increasingly consistent separation between groups. Because lower MAE and higher agreement coefficients both represent closer agreement, the direction of r was interpreted according to the metric being analysed.

No correction for multiple comparisons was applied. Five statistical tests were performed within each participant grouping. If these tests were independent, the probability of obtaining at least one nominally significant result at the 5% level under the global null hypothesis would be approximately 23%. Because the metrics are not independent, this value should not be interpreted as the exact family-wise error probability. The results were therefore interpreted with consideration of the number of comparisons rather than treating an isolated p-value close to 0.05 as strong evidence.

Chapter 4

Results

4.1 Data availability

The results are based on the twenty-two recordings included in the study, subject to the sensor-specific exclusions described in Section 3.2.7. Table 4.1 summarises the availability of heart-rate estimates for each sensor after applying these exclusions and the usable-window criterion. For each agreement analysis, a window was included only when all devices required for that comparison provided a usable heart-rate estimate. The corresponding recording and window counts for each comparison are reported in Table 4.3.

After sensor-specific exclusions, each of the three PPG devices contributed data from twenty-one of the twenty-two recordings. A heart-rate estimate was available for nearly all eligible windows. Persimmon and Shimmer provided estimates for all 3125 windows, whereas Polar provided estimates for 3123 of 3125 windows. The two windows without a Polar heart-rate estimate occurred in the same recording during the `drinking water` activity. In both cases, only two valid inter-beat intervals remained after physiological range filtering, below the minimum of three required to generate a heart-rate estimate.

Table 4.1: Availability of heart-rate estimates for each sensor after applying the sensor-specific exclusions and the usable-window criterion.

Sensor	Recordings	Usable windows	Usable (%)
Persimmon	21	3125	100.0
Polar	21	3123	99.9
Shimmer	21	3125	100.0

4.2 Choice of the bandpass filter

The two candidate bandpass filters described in Section 3.2.4 were evaluated over the complete dataset while all other processing stages were kept unchanged. Table 4.2 summarises the resulting agreement and difference metrics.

The linear-phase FIR filter produced higher CCC values for all three device pairs and a higher three-device ICC than the Butterworth filter. The improvement was also observed for the Polar-Shimmer comparison, which does not involve Persimmon, indicating that the effect was not limited to comparisons involving the prototype. In addition, the FIR filter produced lower Persimmon-Polar and Persimmon-Shimmer mean absolute differences.

Persimmon provided a heart-rate estimate for every eligible window with both filter designs. The improved agreement obtained with the FIR filter was therefore not attributable to a reduction in the number of analysed Persimmon windows. Based on the overall improvement in agreement and pairwise differences, the linear-phase FIR filter was selected for the final processing pipeline. All subsequent results reported in this thesis were obtained using this filter.

Table 4.2: Agreement and pairwise difference metrics obtained using the two candidate bandpass filters, with all other processing stages kept unchanged. CCC and ICC are dimensionless; MAE is reported in bpm. Usable windows are the percentage of Persimmon windows returning an estimate.

Quantity	Butterworth	Linear-phase FIR
CCC Persimmon-Polar	0.581	0.744
CCC Persimmon-Shimmer	0.494	0.715
CCC Polar-Shimmer	0.719	0.868
ICC, three sensors	0.608	0.787
MAE Persimmon-Polar	7.83	5.25
MAE Persimmon-Shimmer	8.62	5.36
MAE Polar-Shimmer	5.38	3.00
Usable windows	100.0%	100.0%

4.3 Overall agreement

The pooled agreement results differed across the three device pairs (Table 4.3). The Polar-Shimmer comparison showed the highest CCC, at 0.868, whereas the Persimmon-Polar and Persimmon-Shimmer comparisons yielded CCC values of 0.744 and 0.715, respectively.

According to the interpretative categories described in Section 3.3.1, the Polar-Shimmer point estimate was within the good-agreement range, while the two

Persimmon comparisons were within the moderate-agreement range. The corresponding CI extended across adjacent categories: the Polar-Shimmer interval ranged from 0.753 to 0.934, spanning the good and excellent ranges, whereas the Persimmon-Polar and Persimmon-Shimmer intervals extended from the moderate into the good range. The three-device ICC was 0.787, corresponding to good agreement, with a 95% CI of 0.715-0.832 extending into the moderate range.

Overall, Polar and Shimmer showed greater pooled concordance with each other than either device showed with Persimmon. However, the CIs of the three pairwise CCC estimates overlapped and were therefore interpreted descriptively rather than as evidence of statistically significant differences between device pairs.

Table 4.3: Pooled concordance for each device pair and for the three devices jointly. CIs were obtained using participant-level cluster bootstrap resampling.

Analysis	Statistic	Estimate	95% CI	Windows	Recordings
Persimmon-Polar	CCC	0.744	0.648-0.804	2978	20
Persimmon-Shimmer	CCC	0.715	0.629-0.774	2980	20
Polar-Shimmer	CCC	0.868	0.753-0.934	2978	20
All three devices	ICC	0.787	0.715-0.832	2833	19

The pairwise difference metrics showed the same ordering (Table 4.4). MAE was similar for the two comparisons involving Persimmon, at 5.3 bpm for Persimmon-Polar and 5.4 bpm for Persimmon-Shimmer. The Polar-Shimmer comparison showed a lower MAE of 3.0 bpm.

The mean signed difference for Persimmon-Polar was -0.5 bpm (95% CI: -2.0 to $+1.2$ bpm), with a CI that included zero. For Persimmon-Shimmer, the mean signed difference was -2.2 bpm (95% CI: -4.1 to -0.3 bpm), indicating lower Persimmon estimates on average. The Polar-Shimmer comparison showed a mean signed difference of -1.8 bpm (95% CI: -2.8 to -0.8 bpm), indicating lower Polar estimates on average.

For all three comparisons, the Bland-Altman limits of agreement were wide relative to the corresponding mean signed differences, indicating that window-level variability contributed substantially more to the observed differences than a constant mean offset alone. RMSE was also substantially higher than MAE for each comparison, indicating a greater influence of windows with relatively large pairwise differences.

The association and tolerance summaries showed a similar pattern. Pearson's correlation coefficient was 0.75 for Persimmon-Polar, 0.73 for Persimmon-Shimmer, and 0.88 for Polar-Shimmer. For the two Persimmon comparisons, approximately 74% of windows differed by no more than 5 bpm and approximately 85% by no more than 10 bpm. The corresponding percentages for Polar-Shimmer were 86.9% and 94.1%, respectively. MAPE was 6.8% for Persimmon-Polar, 6.4% for

Persimmon-Shimmer, and 3.4% for Polar-Shimmer.

Table 4.4: Pairwise difference and association metrics for each device pair, pooled across all usable windows. Differences were calculated as the first-named device minus the second-named device. MAE, RMSE, bias, and limits of agreement are reported in beats per minute; MAPE and tolerance values are reported as percentages. Parentheses indicate 95% participant-level cluster-bootstrap CIs.

Comparison	N	MAE	RMSE	Bias
Persimmon-Polar	2978	5.3 (4.0-6.4)	10.0 (8.2-11.6)	−0.5 (−2.0 to +1.2)
Persimmon-Shimmer	2980	5.4 (3.9-6.8)	10.7 (8.4-12.9)	−2.2 (−4.1 to −0.3)
Polar-Shimmer	2978	3.0 (2.1-4.0)	7.4 (4.7-9.6)	−1.8 (−2.8 to −0.8)

Comparison	r	≤5 bpm (%)	≤10 bpm (%)	MAPE (%)	95% LoA (bpm)
Persimmon-Polar	0.75	73.5	84.8	6.8 (5.1-8.9)	−20.0 to +19.1
Persimmon-Shimmer	0.73	74.3	85.8	6.4 (4.8-8.3)	−22.8 to +18.4
Polar-Shimmer	0.88	86.9	94.1	3.4 (2.6-4.2)	−15.9 to +12.3

Figure 4.1 shows the Bland-Altman plots for the three device pairs. The 95% limits of agreement were −20.0 to +19.1 bpm for Persimmon-Polar and −22.8 to +18.4 bpm for Persimmon-Shimmer. The Polar-Shimmer comparison showed narrower limits, from −15.9 to +12.3 bpm. In all three comparisons, substantial dispersion remained around the mean signed difference across the observed heart-rate range.

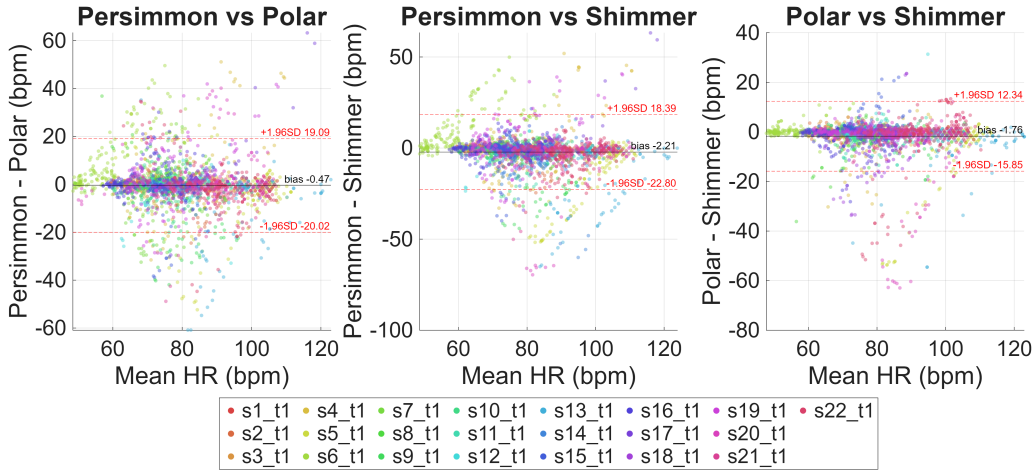


Figure 4.1. Bland-Altman plots for each device pair, with one point per usable window and colours indicating individual recordings. Horizontal lines represent the mean signed difference and the 95% limits of agreement.

Figure 4.2 shows the paired heart-rate estimates for the three device comparisons. The Persimmon-Polar and Persimmon-Shimmer points showed greater dispersion around the identity line than the Polar-Shimmer points, consistent with the corresponding CCC and pairwise difference metrics. The proportion of windows in which the first-named device produced the lower estimate was 51% for Persimmon-Polar, 58% for Persimmon-Shimmer, and 54% for Polar-Shimmer. The corresponding proportions in which the first-named device produced the higher estimate were 41%, 35%, and 33%, respectively; the remaining windows contained identical estimates from the two devices.

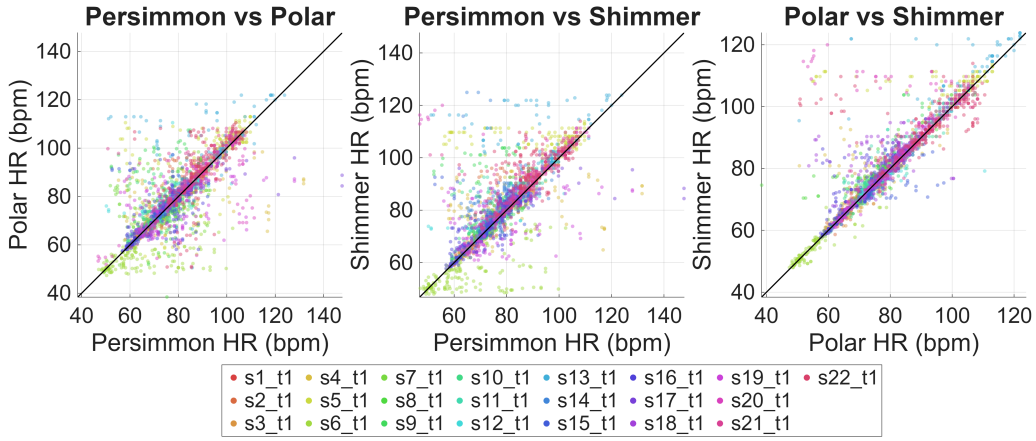


Figure 4.2. Paired heart-rate estimates for each device pair, with one point per usable window and colours indicating individual recordings. The diagonal represents identical estimates from the two devices.

4.4 Agreement per recording

Agreement varied substantially across individual recordings (Figure 4.3). For the Persimmon-Polar comparison, MAE ranged from 1.2 bpm in `s1` to 10.1 bpm in `s13`, while CCC ranged from 0.984 in `s1` to 0.089 in `s15`.

Among the nineteen recordings for which all three pairwise comparisons were available, the Polar-Shimmer MAE was lower than both Persimmon pairwise MAEs in fifteen recordings. The opposite pattern was observed in `s8_t1`, `s11_t1`, and `s20_t1`, where the Polar-Shimmer MAE exceeded both Persimmon pairwise MAEs. In `s20_t1`, the Polar-Shimmer CCC was negative.

A low MAE did not necessarily correspond to a high CCC. For example, in `s8_t1`, the Persimmon-Polar MAE was 3.2 bpm, among the lower values observed across recordings, whereas the corresponding CCC was 0.436. This difference reflects the fact that MAE quantifies the magnitude of pairwise differences, whereas

CCC also depends on the concordance of the values across the range observed within the recording. Consequently, recordings were not ranked identically by MAE and CCC.

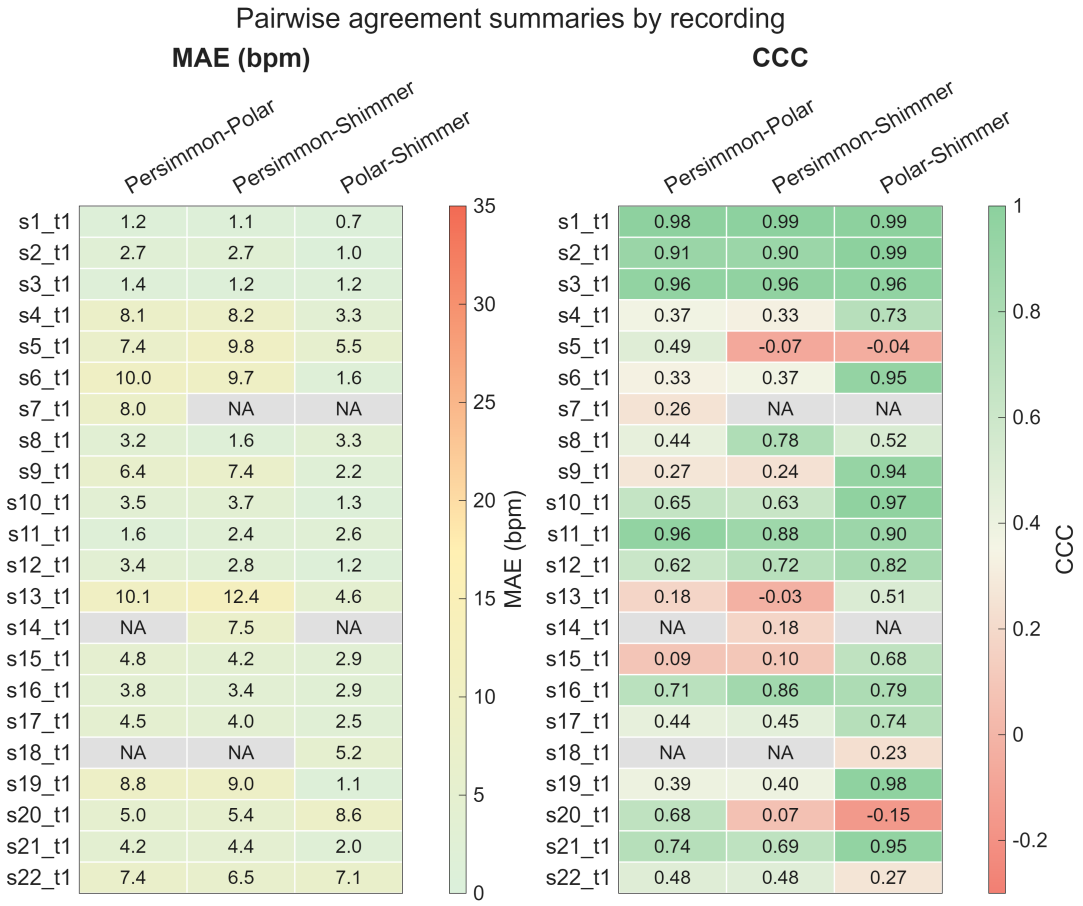


Figure 4.3. Pairwise MAE, in bpm, and CCC calculated separately for each recording and device pair. Grey cells indicate comparisons for which one of the two devices was excluded from that recording.

Agreement across all three devices also varied substantially between recordings (Figure 4.4). Among the nineteen recordings with usable data from all three devices, the ICC ranged from 0.17 to 0.99. Seven recordings had an ICC below 0.50, whereas four exceeded 0.90.

These recording-level results show that the pooled agreement estimates reported in Section 4.3 summarise substantial between-recording variability rather than uniform agreement across the study sample.

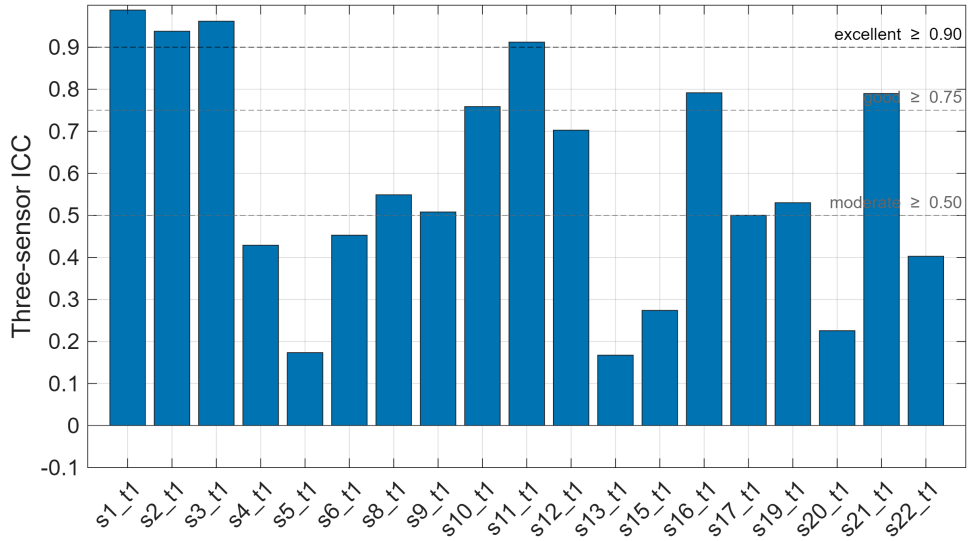


Figure 4.4. Three-device ICC calculated separately for each recording with usable data from all three devices.

4.5 Agreement by activity

Agreement varied substantially across the activities of the protocol (Table 4.5; Figure 4.5). Overall, the activity ranking obtained from MAE was consistent with that indicated by the concordance coefficients. Seated rest showed the best agreement, followed by the seated tasks and standing rest, whereas walking and sitting down showed the largest pairwise differences and the lowest concordance. The three-device ICC followed the same general pattern.

The three seated tasks showed intermediate performance, with using the computer generally producing higher agreement than using the phone or drinking water. Standing rest also fell within this intermediate range despite being a resting condition. Compared with seated rest, however, both Persimmon comparisons showed substantially larger differences and lower concordance during standing. This indicates that resting conditions were not equivalent and that posture may contribute to the observed variability.

Across activities, Polar and Shimmer consistently showed closer agreement with each other than either device showed with Persimmon. This pattern was observed in all sustained activities and in both postural transitions. The relative performance of the two Persimmon comparisons varied across the intermediate conditions, whereas during walking and sitting down Persimmon showed closer agreement with Polar than with Shimmer.

Walking represented the poorest sustained activity and showed a marked reduction in agreement relative to seated rest. This deterioration was expected because walking introduces repeated arm movement and greater susceptibility to motion-related interference in wearable PPG. Importantly, agreement between Polar and Shimmer also decreased during walking, indicating that the deterioration under this condition was not limited to Persimmon.

The two postural transitions showed a less expected pattern. Standing up produced intermediate agreement, whereas sitting down showed the poorest Persimmon agreement of all activities. Polar and Shimmer, in contrast, remained in close agreement during both transitions. Several factors may contribute to this asymmetry. Rapid changes in sensor-skin coupling and contact pressure can affect the PPG waveform during movement [12], while posture and arm position can also influence signal quality [13]. In addition, the transition estimates were based on substantially fewer windows and participants than the sustained activities, resulting in wide CIs. The difference between standing up and sitting down should therefore be interpreted cautiously.

Although the point estimates provide a clear overall ranking of measurement conditions, CIs overlapped for several of the intermediate activities. The activity-specific results are therefore interpreted primarily in terms of the broader pattern: highest agreement during seated rest, intermediate agreement during the seated tasks and standing conditions, and markedly lower agreement during walking and sitting down.

MAE and CCC did not always rank individual intermediate conditions identically. This difference is expected because MAE quantifies the magnitude of pairwise differences, whereas CCC also depends on the variation of heart-rate estimates within a condition. The two measures should therefore be interpreted as complementary descriptions of inter-device agreement.

Table 4.5: Agreement by activity for the three device pairs. For each activity, the first two rows report Persimmon-Polar and Persimmon-Shimmer comparisons, respectively, and the third row reports Polar-Shimmer. N denotes the number of usable windows. MAE and its 95% participant-level cluster-bootstrap CI are reported in bpm; MAPE and tolerance values are percentages. ICC is reported once per activity because it quantifies agreement across all three devices jointly.

Activity	Comparison	N	MAE (bpm)	95% CI (bpm)	$\leq$5 (%)	$\leq$10 (%)	MAPE (%)	CCC	ICC
still (seated)	Polar	1173	2.7	1.8-3.7	88.3	93.7	3.7	0.886	0.930
	Shimmer	1161	2.7	1.8-3.5	87.2	94.7	3.5	0.899	
	Polar-Shimmer	1165	1.2	1.0-1.4	96.9	99.2	1.5	0.982	
using the computer	Polar	339	4.8	3.1-6.9	72.0	88.2	6.3	0.793	0.850
	Shimmer	340	4.9	3.0-7.3	73.8	86.5	6.3	0.793	
	Polar-Shimmer	340	2.8	1.8-3.9	85.9	95.0	3.5	0.928	
using the phone	Polar	377	5.5	3.3-8.3	71.9	84.4	7.9	0.638	0.778
	Shimmer	375	5.1	2.8-7.7	74.1	85.1	7.2	0.679	
	Polar-Shimmer	376	2.6	1.8-3.7	87.5	96.0	3.3	0.924	
still (standing)	Polar	398	6.5	4.0-9.8	69.6	81.9	7.9	0.573	0.705
	Shimmer	409	5.6	3.2-8.6	71.6	86.8	6.7	0.646	
	Polar-Shimmer	407	2.6	1.5-4.7	88.5	95.6	2.8	0.891	
drinking water	Polar	238	5.2	3.6-7.2	68.5	86.1	6.9	0.712	0.783
	Shimmer	242	5.4	2.9-8.8	73.1	88.0	6.1	0.629	
	Polar-Shimmer	238	4.5	3.0-6.2	72.3	87.8	5.5	0.824	
walking	Polar	246	12.9	9.2-16.5	33.7	51.2	15.8	0.373	0.162
	Shimmer	246	16.2	10.1-22.9	37.0	51.2	16.8	-0.084	
	Polar-Shimmer	245	9.1	4.0-16.0	63.3	76.7	9.0	0.266	
stand up	Polar	26	6.4	3.3-11.3	69.2	76.9	7.3	0.630	0.702
	Shimmer	26	6.6	3.7-11.7	65.4	80.8	6.9	0.584	
	Polar-Shimmer	26	2.5	0.9-5.5	88.5	96.2	2.8	0.876	
sit down	Polar	19	20.9	6.1-34.4	26.3	42.1	20.1	-0.074	0.101
	Shimmer	19	22.7	7.0-37.6	21.1	36.8	21.6	-0.128	
	Polar-Shimmer	19	2.9	1.2-5.6	84.2	94.7	3.2	0.943	

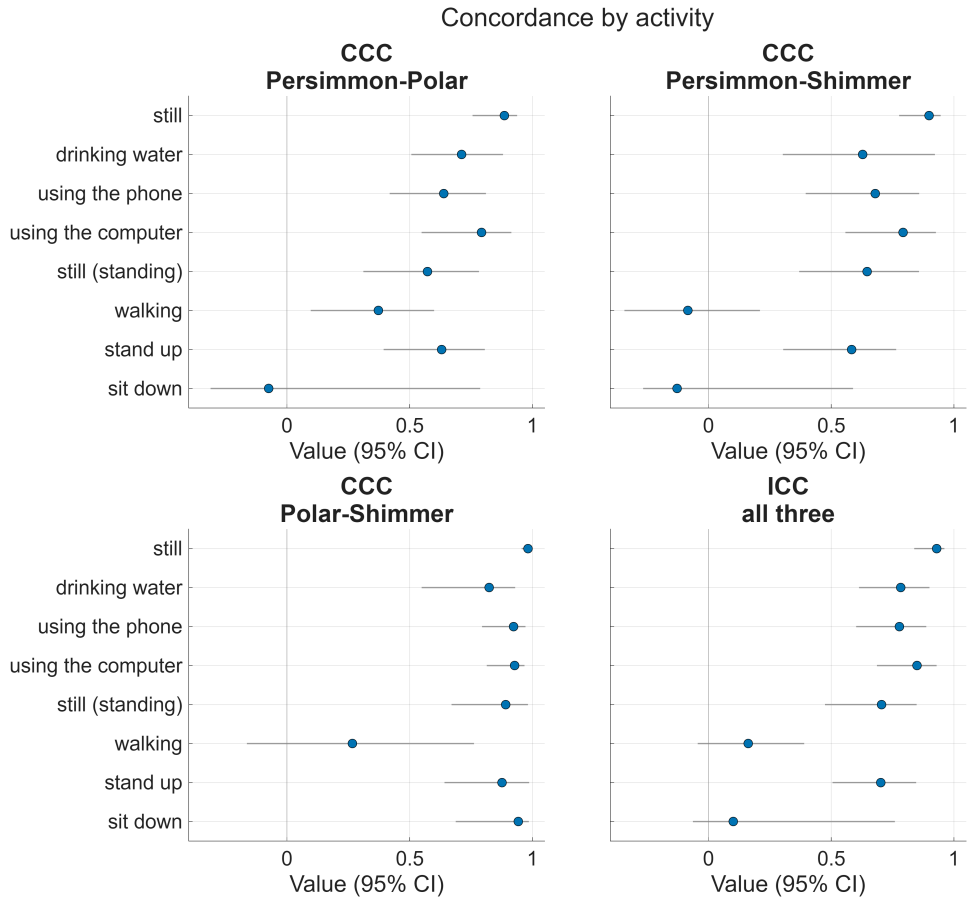


Figure 4.5. Activity-specific pairwise CCC and three-device ICC estimates. Points indicate the estimated coefficient and bars indicate the corresponding 95% participant-level cluster-bootstrap CIs.

4.6 Agreement by motion level

The motion index provided a complementary description of the experimental conditions by grouping windows according to measured arm movement rather than according to the instructed activity. Agreement decreased progressively from low to medium and high motion (Figures 4.7 and 4.8). This trend was observed for all three device pairs and for the three-device ICC.

At low motion, all three device pairs showed relatively high agreement, with Polar-Shimmer showing the highest concordance and the smallest pairwise differences. Agreement decreased at medium motion and deteriorated further at high motion. The same overall progression was observed in the three-device ICC, which decreased across successive motion levels.

The relative ordering of the device pairs also changed with increasing motion. At low and medium motion, Polar-Shimmer showed higher concordance than either comparison involving Persimmon. At high motion, however, Persimmon-Polar showed the highest CCC, whereas Persimmon-Shimmer and Polar-Shimmer showed lower concordance. In terms of pairwise differences, the separation between the three comparisons also became less pronounced at high motion. Thus, the clear Polar-Shimmer advantage observed at low motion was no longer present under high-motion conditions.

The heart-rate estimates showed a related change across motion levels. Median heart rate increased from low to high motion for all three devices, but the increase was larger for Shimmer than for Persimmon and Polar. As a result, Persimmon and Polar remained closer to each other at high motion, whereas their differences from Shimmer increased. This pattern is considered further in the Discussion.

The reduction in agreement at higher motion levels was not accompanied by a reduction in heart-rate availability. At high motion, each device continued to provide a heart-rate estimate for all eligible windows that it contributed. Therefore, the lower agreement observed at high motion was not the result of selective loss of high-motion windows.

Figure 4.6 provides an illustrative example of the optical signals at the three motion levels. At low motion, the filtered waveforms were more regular and the detected peak sequences were similar across the three devices. At higher motion levels, waveform regularity decreased and differences between the detected peak sequences became more apparent. Because the figure shows a single recording, it is included as a qualitative illustration rather than as a quantitative summary of the complete dataset.

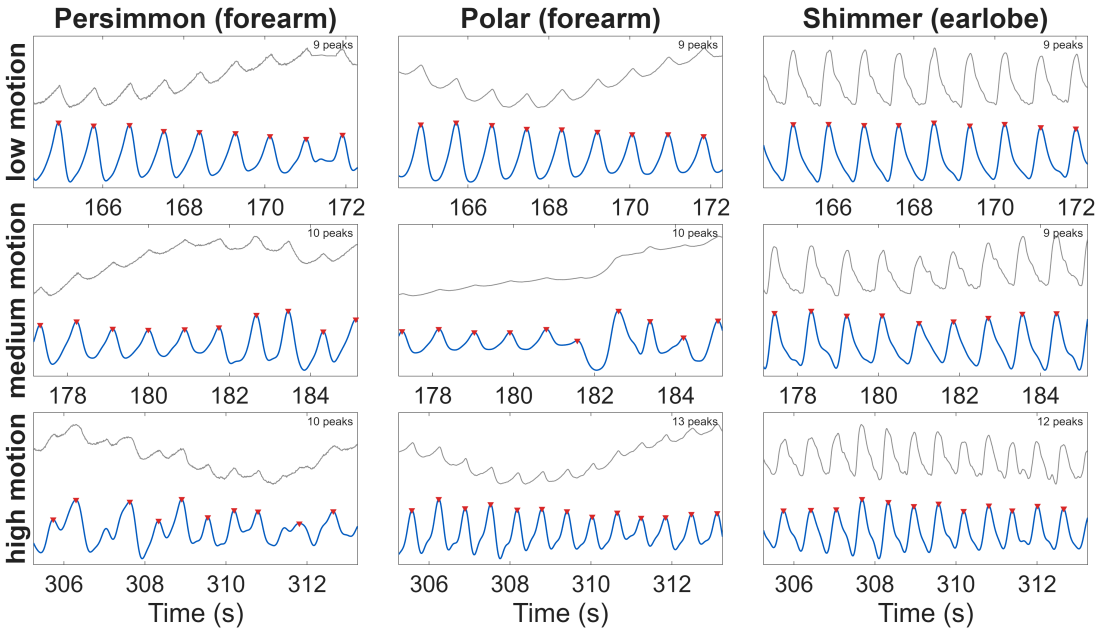


Figure 4.6. Illustrative optical signals from the three devices at low, medium, and high motion within one recording. In each panel, the recorded and filtered signals are shown together with the accepted systolic peaks. Each trace is scaled independently for visualisation.

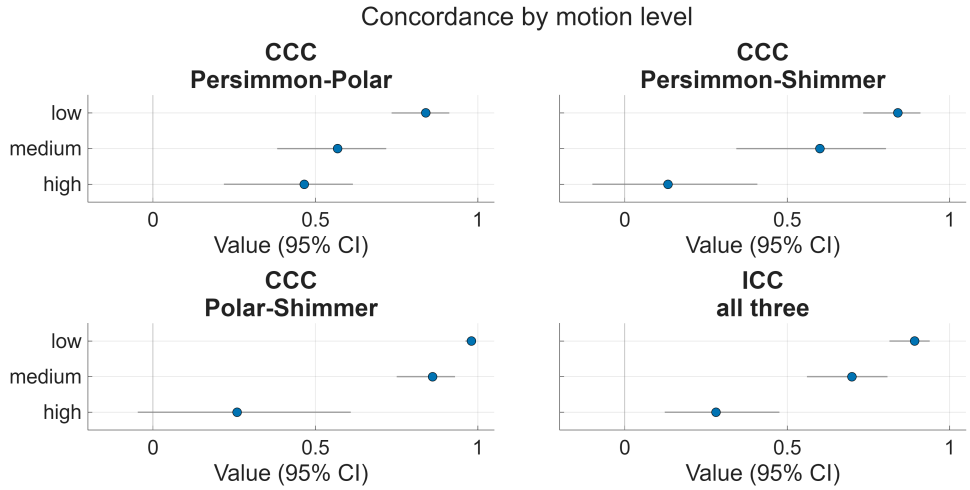


Figure 4.7. Pairwise CCC and three-device ICC estimates across low, medium, and high motion levels. Points indicate the estimated coefficient and bars indicate the corresponding 95% participant-level cluster-bootstrap CIs.

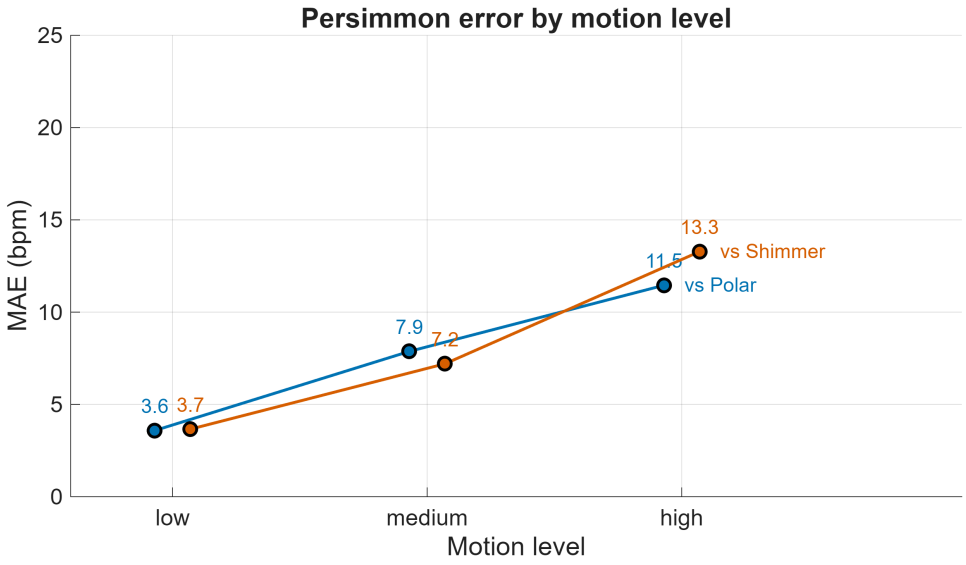


Figure 4.8. Mean absolute difference between Persimmon and each comparison device across low, medium, and high motion levels.

4.7 Agreement by sex

The pooled estimates showed higher agreement in men than in women for both comparisons involving Persimmon. CCC was higher in men for Persimmon-Polar and Persimmon-Shimmer, and the three-device ICC showed the same ordering. The corresponding MAE values were also lower in men. In contrast, the Polar-Shimmer comparison showed slightly lower MAE and higher CCC in women. Thus, the descriptive difference between sexes was not consistent across all three device pairs.

The 95% CIs for the male and female groups overlapped substantially for both MAE and concordance. The observed differences between the pooled subgroup estimates should therefore be interpreted cautiously.

Formal between-group comparisons were performed using the recording-level metrics described in Section 3.3.4. The exact Mann-Whitney test was applied to Persimmon-Polar MAE, Persimmon-Shimmer MAE, Persimmon-Polar CCC, Persimmon-Shimmer CCC, and the three-device ICC (Table 4.6).

None of the five comparisons reached statistical significance. The lowest p-value was obtained for Persimmon-Polar CCC ($p = 0.157$). For this metric, the median recording-level CCC was 0.414 in women and 0.665 in men. The effect-size magnitudes across the five comparisons ranged from 0.11 to 0.40, with the recording-level estimates consistently indicating greater agreement in men according to the direction of each metric.

These analyses were based on relatively small groups. The Persimmon pairwise comparisons included 12 women and 8 men, whereas the Polar-Shimmer comparison included 11 women and 9 men because the sensor-specific exclusions differed between device pairs. The absence of statistically significant differences should therefore not be interpreted as evidence that sex has no effect on inter-device agreement in the broader population.

4.8 Agreement by skin phototype

Agreement was first examined descriptively across the four Fitzpatrick skin phototypes represented in the cohort (Figure 4.9). The pairwise CCC and three-device ICC estimates varied across types I-IV without a consistent trend. Polar-Shimmer showed higher concordance than either comparison involving Persimmon across all four phototypes, while its pairwise differences also remained relatively stable across the groups.

Because the individual phototype groups contained only two to seven participants, the inferential analysis combined types I and II into a lighter phototype group and types III and IV into a darker phototype group, as described in Section 3.3.4. The recording-level agreement metrics were then compared between the two groups using the exact Mann-Whitney test (Table 4.6).

None of the five comparisons reached statistical significance. The smallest p-value was observed for Persimmon-Polar MAE ($p = 0.080$), for which the median recording-level MAE was 7.36 bpm in the types I-II group and 3.83 bpm in the types III-IV group. The corresponding rank-biserial effect size was $r = -0.47$. The remaining comparisons showed the same descriptive direction, with lower MAE and higher concordance estimates in the types III-IV group, but none reached the 5% significance level.

Given the small number of participants within the phototype groups and the lack of statistically significant between-group differences, these results do not provide evidence of a systematic association between Fitzpatrick phototype and inter-device agreement in this cohort.

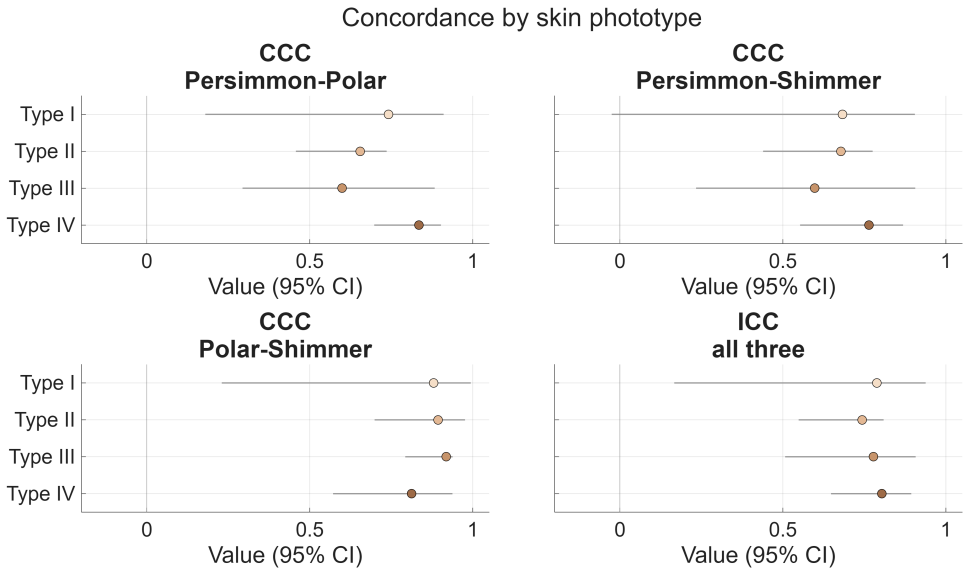


Figure 4.9. Pairwise CCC and three-device ICC estimates across Fitzpatrick skin phototypes I-IV. Points indicate the estimates and bars indicate the corresponding 95% CIs. Types V and VI were not represented in the cohort.

Table 4.6: Between-participant comparisons of recording-level agreement metrics by sex and aggregated Fitzpatrick skin phototype. MAE is reported in bpm; CCC and ICC are dimensionless. p denotes the exact two-sided Mann-Whitney value and r the rank-biserial correlation. Positive r values indicate higher values in the second group and negative values indicate higher values in the first group.

Quantity	Median, women	Median, men	p	r
MAE Persimmon-Polar	5.61	3.69	0.343	−0.27
MAE Persimmon-Shimmer	5.45	3.54	0.427	−0.23
CCC Persimmon-Polar	0.414	0.665	0.157	0.40
CCC Persimmon-Shimmer	0.427	0.704	0.624	0.15
ICC, three devices	0.508	0.654	0.717	0.11
Quantity	Median, types I-II	Median, types III-IV	p	r
MAE Persimmon-Polar	7.36	3.83	0.080	−0.47
MAE Persimmon-Shimmer	6.49	3.69	0.112	−0.43
CCC Persimmon-Polar	0.391	0.645	0.175	0.37
CCC Persimmon-Shimmer	0.400	0.628	0.552	0.17
ICC, three devices	0.453	0.654	0.278	0.31

4.9 Agreement against the practical benchmarks

The PERSIMMON project specifies heart-rate accuracy targets of ±5 bpm at rest and ±10 bpm during movement [21]. Because no ECG reference was available in the present study, these thresholds cannot be used to establish absolute accuracy. They are therefore used here as practical benchmarks for interpreting the differences between Persimmon and the two comparison devices.

Across the complete protocol, Persimmon showed an MAE of 5.3 bpm relative to Polar and 5.4 bpm relative to Shimmer. Approximately 74% of paired windows differed by no more than 5 bpm and approximately 85% by no more than 10 bpm. These pooled values, however, combine conditions with substantially different levels of agreement.

Seated rest showed the closest agreement with both comparison devices. The Persimmon MAE was 2.7 bpm relative to both Polar and Shimmer, and approximately 88% of windows were within 5 bpm. Low-motion windows showed a similar pattern, with pairwise differences remaining substantially smaller than at higher motion levels. In contrast, standing rest showed larger Persimmon differences, with MAE values slightly above 5 bpm for both comparison devices.

Most seated activities and the standing-up transition remained below a mean absolute difference of 10 bpm. Walking and sitting down were the main exceptions. During walking, Persimmon MAE increased to 12.9 bpm relative to Polar and 16.2 bpm relative to Shimmer, while sitting down showed even larger estimates. The corresponding CIs were wide, particularly for sitting down, which was based on a small number of windows. These values should therefore be interpreted as evidence of substantially poorer agreement under these conditions rather than as precise estimates of the amount by which a performance target was exceeded.

The walking result also requires additional caution because Polar-Shimmer agreement deteriorated markedly under the same condition. Consequently, the larger Persimmon differences observed during walking cannot be attributed to the prototype alone.

Overall, the practical benchmark comparison indicates that Persimmon agreement was closest to the project targets under seated-rest and low-motion conditions, whereas substantially larger inter-device differences were observed during walking and sitting down. These comparisons describe agreement with the two comparison devices and should not be interpreted as direct validation of the PERSIMMON accuracy requirements.

Chapter 5

Discussion

Across the complete protocol, Persimmon showed an MAE of 5.3 bpm relative to Polar. However, this pooled value combines conditions with markedly different levels of agreement. MAE ranged from 2.7 bpm during seated rest to 12.9 bpm during walking and 20.9 bpm during the transition into sitting. The pooled result therefore does not represent a uniform level of agreement across the protocol.

This dependence on measurement condition was also evident in the motion-based analysis. When windows were grouped according to the accelerometer-derived motion index, pairwise differences increased progressively from low to medium and high motion. These results indicate a clear association between movement intensity and reduced inter-device agreement.

5.1 Interpretation of the overall agreement

Across the complete protocol, Polar and Shimmer showed higher concordance with each other than either device showed with Persimmon. The two Persimmon comparisons were nevertheless similar in magnitude, suggesting that the overall result was not driven by disagreement with only one comparison device.

The mean signed differences were small relative to the corresponding Bland-Altman limits of agreement. In particular, the Persimmon-Polar mean difference was close to zero, while the Persimmon-Shimmer comparison showed a small negative mean difference. The substantially wider limits of agreement indicate that the main source of disagreement was variation between individual windows rather than a uniform offset between devices. A simple calibration correction would therefore address only a limited component of the observed differences.

The use of a common heart-rate processing pipeline is important in interpreting these results. The same filtering, peak-detection, and heart-rate estimation procedures were applied to all three optical signals, reducing differences that would

otherwise arise from proprietary heart-rate algorithms. However, the devices still differed in sensing hardware, optical configuration, attachment, sampling characteristics, measurement site, and any analogue or device-level conditioning applied before the recorded signal became available. The common digital pipeline therefore reduces, but does not eliminate, device-related sources of variability.

The activity-dependent pattern observed here is consistent with previous evidence showing that wearable PPG agreement is generally higher under sedentary conditions and deteriorates during movement [15]. The present study extends this observation by comparing three simultaneously recorded optical signals processed using the same heart-rate estimation pipeline.

5.2 Effects of activity and movement

Activity-specific and motion-specific analyses showed the same overall pattern: agreement was highest under low-movement conditions and decreased as movement increased. The motion analysis is complementary to the activity analysis because the activity labels describe the instructed task, whereas the motion index quantifies the amount of arm movement measured during each heart-rate window.

Seated rest showed the highest agreement of the protocol. The three seated daily tasks showed intermediate performance, while walking produced the poorest agreement among the sustained activities. This progression suggests that the difficulty of the measurement condition was more closely related to movement than to the physiological level of heart rate alone.

Standing rest is particularly informative because it performed substantially worse than seated rest despite both conditions being classified as rest. Previous work has shown that posture, arm support, and sensor position can influence wrist PPG signal quality [13]. In the present protocol, seated rest was performed with different arm support and posture from standing rest, while the standing condition also showed somewhat greater measured movement. These factors were not manipulated independently, so their individual contributions cannot be determined. Nevertheless, the result indicates that "rest" should not be treated as a single homogeneous measurement condition.

Walking produced a marked reduction in agreement across all three device pairs. Motion artefact is a well-established limitation of wearable PPG, arising in part from changes in sensor-skin coupling and contact pressure during movement [12]. During quasi-periodic activities such as walking, motion-related components can also overlap with the frequency range of the cardiac signal, which limits the effectiveness of conventional linear filtering [33]. These mechanisms are consistent with the increasing irregularity observed in the optical signals at higher motion levels.

Importantly, Polar-Shimmer agreement also deteriorated substantially during walking and high-motion windows. The reduced Persimmon agreement under these conditions therefore cannot be interpreted as evidence of a failure specific to the prototype. Without an electrical reference, the present study cannot determine which of the three devices was closest to the underlying cardiac rate when their estimates diverged.

The relative behaviour of the three device pairs at high motion also raises the possibility of an effect of measurement site. Persimmon and Polar recorded PPG from the forearm, whereas the Shimmer optical probe was positioned on the earlobe. At high motion, Persimmon and Polar remained more concordant with each other than either was with Shimmer, and the increase in median heart rate from low to high motion was larger for Shimmer than for the two forearm devices.

This pattern is compatible with a contribution of anatomical location because the forearm undergoes greater displacement during walking than the head and earlobe. Previous work has also reported favourable PPG signal quality at the ear among several anatomical sites [9]. However, measurement site cannot be isolated from device design, attachment, optical configuration, or upstream signal conditioning in the present experiment. Furthermore, movement was measured at the arm rather than at the Shimmer optical probe. The observed pattern should therefore be regarded as a hypothesis for future investigation rather than evidence that anatomical location caused the difference.

The two postural transitions showed an additional, less expected pattern. Standing up produced intermediate agreement, whereas sitting down produced some of the largest Persimmon differences observed in the study. Polar and Shimmer, in contrast, remained comparatively close to each other during both transitions. One possible explanation is that the two movements produce different transient changes in sensor-skin coupling, contact pressure, arm position, or strap loading. These factors could affect Persimmon differently even if the two transitions appear mechanically similar at a broader level.

The transition results require caution, however, because they were based on far fewer windows and participants than the sustained activities, resulting in wide CIs. A small number of recordings can therefore have a large influence on these estimates. Repeated transition trials, with arm position and support controlled independently, would be required to determine whether the apparent asymmetry between standing up and sitting down is reproducible.

5.3 Effects of participant characteristics

The analyses by sex showed descriptive differences but no statistically significant between-group effect. Both comparisons involving Persimmon showed higher concordance and lower MAE in men, whereas Polar-Shimmer showed the opposite descriptive ordering. The absence of a consistent direction across all three device pairs makes it difficult to interpret the pattern as a general effect of sex on optical heart-rate measurement.

One possible interpretation is that the observed difference involves the interaction between participant characteristics and the specific device or its fit rather than a participant characteristic acting equally on all three measurements. However, the small subgroup sizes prevent a reliable distinction between these explanations. The present results therefore provide no evidence of a statistically detectable sex-related difference in inter-device agreement, but they do not exclude smaller effects.

Fitzpatrick skin phototype showed a similar result. Agreement estimates varied across types I-IV without showing a consistent increase or decrease across phototypes. After aggregation into types I-II and III-IV, none of the five recording-level comparisons reached statistical significance. Descriptively, the darker phototype group showed lower MAE and higher concordance, which does not suggest progressive deterioration with increasing phototype in this sample.

The literature on skin pigmentation and wearable optical heart-rate measurement is not entirely uniform. Some individual-device studies have reported performance differences associated with skin tone [26], whereas a more recent synthesis found no clinically significant pigmentation bias in wearable pulse-rate estimation [14]. The present results are compatible with the absence of a large effect in the studied sample, but the limited phototype distribution and subgroup size prevent stronger conclusions.

5.4 Practical implications

The PERSIMMON project specifies heart-rate accuracy targets of ± 5 bpm at rest and ± 10 bpm during movement [21]. Because no ECG reference was included in this study, these thresholds can only be used as practical benchmarks for inter-device differences and not as direct tests of absolute accuracy.

Persimmon agreement was closest to the project targets during seated rest and low-motion periods. Agreement remained intermediate during the seated daily tasks and standing conditions, whereas walking and sitting down showed substantially larger differences. The pooled MAE therefore conceals a strong dependence on measurement condition.

For continuous ambulatory monitoring, the deterioration observed during walking is particularly relevant because locomotion is a routine component of daily life. At the same time, the reduction in Polar-Shimmer agreement during walking shows that this condition is challenging for the comparison devices as well. Consequently, the present data identify motion robustness as an important area for improvement but do not establish the absolute magnitude of Persimmon error during locomotion.

The transition into sitting may represent a more device-specific issue because Polar and Shimmer remained comparatively close while Persimmon diverged from both. However, the small number of transition windows means that this finding requires confirmation in a protocol designed specifically to repeat and characterise postural transitions.

5.5 Methodological strengths and considerations

A major strength of the experimental design is the simultaneous acquisition of the three optical signals from the same participant. Each device therefore recorded the same physiological period and was exposed to the same activity sequence. Applying the same heart-rate estimation pipeline to all three signals further reduced algorithm-related differences and allowed the comparison to focus more directly on the acquired optical signals and device-specific characteristics.

The use of two comparison devices also improved interpretation. When Persimmon disagreed with one comparison device, the third device provided information about whether that disagreement was specific to Persimmon or reflected a condition in which the optical measurements generally diverged. This was particularly useful during walking, where Polar and Shimmer also lost agreement with each other.

Synchronisation was both performed and checked. The initial synchronisation gesture aligned the independently recorded device signals, while the final gesture was used to quantify residual drift. This reduced the likelihood that the observed disagreement arose from systematic temporal misalignment between devices.

Participant-level cluster bootstrap resampling was used to estimate uncertainty. This was necessary because overlapping windows from the same participant are not independent observations. Resampling participants rather than individual windows therefore provided CIs that better reflect between-recording variability.

The processing pipeline was designed with online implementation in mind. The FIR filter operates causally and introduces a fixed group delay that can be compensated in the signal timestamps. The peak detector, however, uses centred moving averages and therefore requires a limited amount of future data before a peak can be finalised. The complete pipeline should consequently be considered compatible

with delayed online processing rather than strictly causal at each sample instant.

The bandpass filter and heart-rate estimation strategy were selected after comparison of candidate approaches on the available recordings. The selected FIR filter produced higher agreement than the Butterworth alternative, including for the Polar-Shimmer comparison, while alternative heart-rate estimation rules did not improve the balance between agreement and data availability. These comparisons demonstrate that processing choices materially affect the resulting agreement estimates and should therefore be reported transparently.

5.6 Limitations

The principal limitation is the absence of a simultaneous ECG reference. All reported quantities describe agreement among optical measurements rather than accuracy relative to cardiac electrical activity. When two devices disagree, the present data cannot establish which estimate is closer to the underlying heart rate. This limitation is particularly relevant during walking and high-motion conditions, where the two comparison devices also diverged.

The cohort also limits the generalisability of the findings. Twenty-two participants completed one recording session each, and the sample was concentrated in younger adults. Fitzpatrick phototypes V and VI were not represented, leaving the darker end of the phototype scale unexamined. The relatively small subgroup sizes also limited the ability to detect differences associated with sex or skin phototype.

Sex and skin phototype were unevenly distributed within the cohort. The lighter phototype group contained a greater proportion of women, whereas the darker phototype group contained a greater proportion of men. Consequently, any observed between-group difference could not be attributed uniquely to sex or skin phototype.

Motion was quantified using acceleration measured at the arm. Although this provided an objective measure of movement intensity, it did not quantify the movement experienced directly at every optical measurement site. In particular, the Shimmer PPG probe was positioned at the earlobe while its accelerometer was worn on the wrist. The contribution of anatomical measurement site to the high-motion results therefore could not be isolated from differences in device design, attachment, and local movement.

No dedicated motion-artifact suppression algorithm was applied. Acceleration was used for synchronisation and motion characterisation but not for adaptive correction of the PPG signals. The present results therefore do not establish how much of the reduction in agreement at higher motion levels could be recovered through motion-aware processing.

The bandpass filter and heart-rate estimation strategy were selected after comparison of candidate approaches on the same dataset used for the final agreement analyses. Although each candidate was applied consistently across the devices, this data-driven selection may favour the processing configuration that performed best in this particular cohort. The reported performance should therefore be confirmed using independent data.

Activity- and motion-specific comparisons were descriptive rather than formal between-condition statistical tests. Repeated windows from the same participants contributed to multiple conditions, and the reported CIs quantify uncertainty within each condition rather than directly testing differences between activities or motion levels.

Finally, each participant completed only one recording session. The observed between-recording variability therefore combines between-participant differences with session-specific factors such as device placement, strap tightness, and attachment. Test-retest repeatability and the relative contribution of these sources of variability could not be assessed.

Chapter 6

Conclusion and future work

This study evaluated inter-device heart-rate agreement between the Persimmon prototype, the Polar Verity Sense, and the Shimmer3 GSR+ during activities of daily living. Raw optical signals from the three devices were processed using the same heart-rate estimation pipeline, reducing the influence of differences between proprietary heart-rate algorithms.

The results showed that inter-device agreement depended strongly on the measurement condition. Agreement was highest during seated rest and low-motion periods and decreased as movement increased. The difference between seated and standing rest further showed that conditions classified broadly as rest should not necessarily be treated as equivalent when evaluating wearable PPG performance. The pooled results therefore provide only a partial description of performance when substantially different measurement conditions are combined.

The study provides several methodological contributions relative to the wearable PPG validation approaches reviewed in Chapter 2. First, the same filtering, peak-detection, and heart-rate estimation procedures were applied to all three optical signals. This approach does not isolate the sensing hardware, because measurement site, attachment, optical configuration, and upstream acquisition remain device-specific. However, it reduces variability associated with different heart-rate algorithms and enables a more direct comparison of the acquired optical signals.

Second, two comparison devices were recorded simultaneously, and their agreement with each other was evaluated alongside the comparisons involving Persimmon. This provided information that would not be available from a design using a single comparison device. During walking and high-motion conditions, Polar-Shimmer agreement also deteriorated, showing that large differences between Persimmon and a comparison device under these conditions could not automatically be attributed to the prototype. This result highlights the value of examining agreement among comparison devices when more than one is available.

Third, activity labels and accelerometer-derived motion levels provided complementary descriptions of the measurement conditions. Activity labels identified the task being performed, whereas the motion index quantified the movement occurring within individual heart-rate windows. Their combined use showed that protocol-defined activities alone may not fully characterise the conditions affecting wearable PPG measurements.

The study also demonstrates the importance of reporting condition-specific results rather than relying exclusively on a pooled performance value. Substantial differences between resting, daily-life, and higher-motion conditions would have been obscured by the overall estimate. Combining pooled, activity-specific, and motion-specific analyses therefore provides a more informative description of wearable PPG performance under heterogeneous daily-life conditions.

If shared, the dataset could provide an additional resource for the development and comparison of wearable PPG processing methods. It contains simultaneous raw optical recordings from three devices, PPG measurements from forearm and earlobe sites, concurrent accelerometry, synchronised video-based activity labels, and recordings of activities of daily living from twenty-two participants. Because the raw signals are available on a common temporal basis, alternative filtering, peak-detection, heart-rate estimation, and motion-artifact suppression methods could be evaluated on the same underlying recordings. The absence of a synchronous ECG reference limits its use for absolute accuracy validation but does not prevent its use for studies of inter-device agreement and motion-related signal behaviour.

The primary next step is validation against a synchronous ECG reference. This would allow Persimmon to be evaluated in terms of absolute heart-rate accuracy, clarify which optical estimate is closest to the underlying cardiac rate when the devices diverge, and permit direct evaluation of the PERSIMMON accuracy requirements.

Improving robustness to movement is another priority. The inertial signals recorded by Persimmon could be incorporated into adaptive or motion-informed processing methods rather than being used only to characterise movement. Because the recordings also include synchronised activity labels, the effect of these methods could be evaluated not only across motion levels but also within specific activities. This would make it possible to determine whether motion-artifact suppression performs consistently across different daily-life conditions and to quantify how much of the motion-related degradation can be recovered.

The possible contribution of anatomical measurement site should also be investigated under controlled conditions. At high motion, the two forearm measurements remained more concordant with each other than with the earlobe measurement, but measurement site was confounded with device design and attachment. Recording comparable optical and inertial signals directly at multiple anatomical sites would

allow these factors to be examined separately.

Posture and postural transitions also require more controlled investigation. Seated and standing rest could be compared while varying arm support independently of posture. Repeated standing-up and sitting-down trials would provide more precise estimates of transition-specific agreement and determine whether the difference observed between the two transitions is reproducible.

Finally, validation should be extended to a larger and more diverse cohort, including Fitzpatrick phototypes V and VI and a more balanced distribution of sex, phototype, and age. Repeated recording sessions would also allow test-retest repeatability to be quantified and would help separate between-participant variability from session-specific effects such as sensor placement and attachment.

Together, these extensions would allow the present inter-device agreement study to progress toward direct validation of Persimmon accuracy and toward the development of more robust heart-rate monitoring under ambulatory conditions.

Bibliography

[1] Richard Gordan, Judith K Gwathmey, and Lai-Hua Xie. Autonomic and endocrine control of cardiovascular function. *World Journal of Cardiology*, 7(4):204–214, 2015.

[2] Kim Fox, Jeffrey S Borer, A John Camm, Nicolas Danchin, Roberto Ferrari, Jose L Lopez Sendon, Philippe Gabriel Steg, Jean-Claude Tardif, Luigi Tavazzi, and Michal Tendera. Resting heart rate in cardiovascular disease. *Journal of the American College of Cardiology*, 50(9):823–830, 2007.

[3] Hirofumi Tanaka, Kevin D Monahan, and Douglas R Seals. Age-predicted maximal heart rate revisited. *Journal of the American College of Cardiology*, 37(1):153–156, 2001.

[4] Jiapu Pan and Willis J Tompkins. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, 1985.

[5] Anthony Atkielski and Kychot. ECG-RRinterval.svg. https://commons.wikimedia.org/wiki/File:ECG-RRinterval.svg, 2013. Wikimedia Commons; copyrighted free use, accessed 2026.

[6] John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3):R1–R39, 2007.

[7] Peter H Charlton. Photoplethysmogram signal components. https://commons.wikimedia.org/wiki/File:Photoplethysmogram_signal_components.svg, 2021. Wikimedia Commons; licensed CC BY 4.0, accessed 2026.

[8] Peter H Charlton, John Allen, Raquel Bailón, Stephanie Baker, Joachim A Behar, Fei Chen, Gari D Clifford, David A Clifton, Harry J Davies, Cheng Ding, et al. The 2023 wearable photoplethysmography roadmap. *Physiological Measurement*, 44(11):111001, 2023.

[9] V. Hartmann, H. Liu, F. Chen, Q. Qiu, S. Hughes, and D. Zheng. Quantitative comparison of photoplethysmographic waveform characteristics: effect of measurement site. *Frontiers in Physiology*, 10:198, 2019.

[10] Masoud Moghaddam, James P Collins, Caroline E Gardner, and Michael C Rabel. Impact of anatomical placement on the accuracy of wearable heart rate monitors during rest and various exercise intensities. *Sensors*, 26(1):176, 2025.

[11] David Pollreisz and Nima TaheriNejad. Detection and removal of motion artifacts in PPG signals. *Mobile Networks and Applications*, 27(2):728–738, 2022.

[12] Daniel Ray, Tim Collins, Sandra Woolley, and Prasad Ponnapalli. A review of wearable multi-wavelength photoplethysmography. *IEEE Reviews in Biomedical Engineering*, 16:136–151, 2023.

[13] Peter H Charlton, Vaidotas Marozas, Elisa Mejía-Mejía, Panicos A Kyriacou, and Jonathan Mant. Determinants of photoplethysmography signal quality at the wrist. *PLOS Digital Health*, 4(6):e0000585, 2025.

[14] Sanidhya Singh, Miles Romney Bennett, Chen Chen, Sooyoon Shin, Hamid Ghanbari, and Benjamin W Nelson. Impact of skin pigmentation on pulse oximetry blood oxygenation and wearable pulse rate accuracy: systematic review and meta-analysis. *Journal of Medical Internet Research*, 26:e62769, 2024.

[15] Oonagh M Giggins, Julie Doyle, Suzanne Smith, Daniel R Crabtree, and Matthew Fraser. Measurement of heart rate using the Withings ScanWatch device during free-living activities: validation study. *JMIR Formative Research*, 6(9):e34280, 2022.

[16] Cailbhe Doherty, Maximus Baldwin, Alison Keogh, Brian Caulfield, and Rob Argent. Keeping pace with wearables: a living umbrella review of systematic reviews evaluating the accuracy of consumer wearable technologies in health measurement. *Sports Medicine*, 54(11):2907–2926, 2024.

[17] Joel D Reece, Jennifer A Bunn, Minsoo Choi, and James W Navalta. Assessing heart rate using consumer technology association standards. *Technologies*, 9(3):46, 2021.

[18] J Martin Bland and Douglas G Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 1(8476):307–310, 1986.

[19] Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, 1989.

[20] Kenneth O. McGraw and S. P. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46, 1996.

[21] PERSIMMON Consortium. First tests in clinic and sports. Deliverable 1.8, PERSIMMON project, Horizon Europe, 2025. Accuracy targets originally specified in Deliverable 1.1.

[22] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types I through VI. *Archives of Dermatology*, 124(6):869–871, 1988.

[23] Theresa Schweizer and Rahel Gilgen-Ammann. Wrist-worn and arm-worn wearables for monitoring heart rate during sedentary and light-to-vigorous physical activities: Device validation study. *JMIR Cardio*, 9:e67110, 2025.

[24] Fernanda Laís Loro, Riane Martins, Janaína Barcellos Ferreira, et al. Validation of a wearable sensor prototype for measuring heart rate to prescribe physical activity: cross-sectional exploratory study. *JMIR Biomedical Engineering*, 9:e57373, 2024.

[25] Brinnae Bent, Benjamin A Goldstein, Warren A Kibbe, and Jessilyn P Dunn. Investigating sources of inaccuracy in wearable optical heart rate sensors. *npj Digital Medicine*, 3:18, 2020.

[26] Anna Shcherbina, C Mikael Mattsson, Daryl Waggott, Heidi Salisbury, Jeffrey W Christle, Trevor Hastie, Matthew T Wheeler, and Euan A Ashley. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of Personalized Medicine*, 7(2):3, 2017.

[27] Jiayi Shao, Jiaying Ye, Shengyao Liu, Zachary Englhardt, Girish Narayanswamy, Vikram Iyer, and Qiuyue Shirley Xue. Multi-site ppg: An in-the-wild physiological dataset from emerging multi-site wearables, 2026.

[28] Vincenzo Ronca, Ana C Martinez-Levy, Alessia Vozzi, Andrea Giorgi, Pietro Aricò, Rossella Capotorto, Gianluca Borghini, Fabio Babiloni, and Gianluca Di Flumeri. Wearable technologies for electrodermal and cardiac activity measurements: a comparison between Fitbit Sense, Empatica E4 and Shimmer GSR3+. *Sensors*, 23(13):5847, 2023.

[29] Junyung Park, Hyeon Seok Seok, Sang-Su Kim, and Hangsik Shin. Photoplethysmogram analysis and applications: an integrative review. *Frontiers in Physiology*, 12:808451, 2022.

[30] Yongbo Liang, Mohamed Elgendi, Zhencheng Chen, and Rabab Ward. An optimal filter for short photoplethysmogram signals. *Scientific Data*, 5:180076, 2018.

[31] Alan V Oppenheim and Ronald W Schafer. *Discrete-Time Signal Processing*. Prentice Hall, 3rd edition, 2009.

[32] Mohamed Elgendi, Ian Norton, Matt Brearley, Derek Abbott, and Dale Schuurmans. Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions. *PLoS ONE*, 8(10):e76585, 2013.

[33] Zhilin Zhang, Zhouyue Pi, and Benyuan Liu. TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. *IEEE Transactions on Biomedical Engineering*, 62(2):522–531, 2015.

[34] Francesco Sartor, Gabriele Papini, Lieke Gertruda Elisabeth Cox, and John Cleland. Methodological shortcomings of wrist-worn heart rate monitors validations. *Journal of Medical Internet Research*, 20(7):e10108, 2018.

[35] Jukka Happonen. Polar sensor logger. https://play.google.com/store/apps/details?id=com.j_ware.polarsensorlogger, 2024. Third-party Android application, not an official Polar product; accessed 2026.

[36] Polar Electro Oy. Polar BLE SDK. https://github.com/polarofficial/polar-ble-sdk, 2025. Official software development kit and device documentation, accessed 2026.

[37] Polar Electro Oy. Polar verity sense. https://www.polar.com/us-en/products/accessories/polar-verity-sense/, 2024. Manufacturer product page, accessed 2026.

[38] Shimmer Research Ltd. Optical pulse ear-clip – shimmer wearable sensor technology. https://www.shimmersensing.com/product/optical-pulse-ear-clip/, 2024. Manufacturer product documentation, accessed 2026.

[39] Nordic Semiconductor. Nordic thingy:52 iot sensor kit. https://www.nordicsemi.com/Products/Development-hardware/Nordic-Thingy-52, 2024. Manufacturer product page, accessed 2026.

[40] Florenc Demrozi, Cristian Turetta, Philipp H Kindt, Fabio Chiarani, Ruggero Angelo Bacchin, Nicola Valè, Francesco Pascucci, Paola Cesari, Nicola Smania, Stefano Tamburin, and Graziano Pravadelli. A low-cost wireless body

area network for human activity recognition in healthy life and medical applications. *IEEE Transactions on Emerging Topics in Computing*, 11(4):839–850, 2023.

[41] Jiawei Guo, Shiyuan Chen, Ting Lan, Ruochen Li, Lichao Wang, Yunchong Wu, Jun Zhong, and Wei Zhu. Research on heart rate estimation algorithm based on dynamic PPG. *Frontiers in Signal Processing*, 6, 2026.

[42] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016.

[43] J Martin Bland and Douglas G Altman. Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics*, 17(4):571–582, 2007.

[44] C. A. Field and A. H. Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):369–390, 2007.

[45] H B Mann and D R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.