frontiers | Research Topics

Physio-logging in humans: recent advances and limitations in wearable devices for biomedical applications

Edited by
Mohammad Yavarimanesh, Colin K. Drummond and
Cederick Landry

Published in
Frontiers in Physiology

ISSN 1664-8714
ISBN 978-2-8325-6659-6
DOI 10.3389/978-2-8325-6659-6

Generative AI statement
Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Physio-logging in humans: recent advances and limitations in wearable devices for biomedical applications

Topic editors
Mohammad Yavarimanesh — University of San Diego, United States
Colin K. Drummond — Case Western Reserve University, United States
Cederick Landry — Université de Sherbrooke, Canada

Citation
Yavarimanesh, M., Drummond, C. K., Landry, C., eds. (2025). *Physio-logging in humans: recent advances and limitations in wearable devices for biomedical applications*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-6659-6

Topic Editor Dr. Mohammad Yavarimanesh is a bio-algorithm data scientist at Philips and an Adjunct Professor at the University of San Diego in San Diego, CA, USA. However, his ongoing projects are related to hospital patient monitoring and are not linked to wearable/nearable technology. Topic Editor Dr. Céderick Landry has a patent on a Method and system for cueing a user of a tool using wearables. Topic Editor Prof. Colin Drummond declares no competing interests with regard to the Research Topic subject.

Table of contents

Frontiers in Physiology

TYPE Editorial
PUBLISHED 25 June 2025
DOI 10.3389/fphys.2025.1639428

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
John D. Imig,
University of Arkansas for Medical Sciences,
United States

*CORRESPONDENCE
Colin K. Drummond,
✉ colin.drummond@case.edu

RECEIVED 02 June 2025
ACCEPTED 12 June 2025
PUBLISHED 25 June 2025

CITATION
Landry C, Yavarimanesh M and Drummond CK (2025) Editorial: Physio-logging in humans: recent advances and limitations in wearable devices for biomedical applications.
Front. Physiol. 16:1639428.
doi: 10.3389/fphys.2025.1639428

Editorial: Physio-logging in humans: recent advances and limitations in wearable devices for biomedical applications

Céderick Landry[1], Mohammad Yavarimanesh[2] and Colin K. Drummond[3]*

[1]Mechanical Engineering Department, Université de Sherbrooke, Sherbrooke, QC, Canada, [2]Applied Data Science, University of San Diego, San Diego, CA, United States, [3]Biomedical Engineering, Case Western Reserve University, Cleveland, OH, United States

KEYWORDS

biomedical sensors, wearable sensors, human performance, physical sensors, chemical sensors

Editorial on the Research Topic
Physio-logging in humans: recent advances and limitations in wearable devices for biomedical applications

The recent advancements in wearables and machine learning have paved the way for unparalleled approaches to monitor physiological parameters, prevent diseases and medical conditions, and to assist, and treat patients that suffer from them. These approaches also show great potential in studying human physiology in extreme conditions. Wearable devices can provide real-time information about human health and wellbeing in extreme environments, enabling early detection of any changes or abnormalities in normal physiological function. In addition, wearables and recent advances in physio-logging can alleviate the impact of numerous diseases, and medical conditions globally. These approaches will impact our life also by reducing the cost of healthcare and increasing patients' quality of life. Noteworthy strides have already been accomplished, evoking enthusiasm among patients and researchers alike.

Based on the considerations outlined above, this Research Topic, entitled "Physio-logging in Humans: Recent Advances and Limitations in Wearable Devices for Biomedical Applications," aimed to showcase original multidisciplinary research focused on the development, validation, translational Research Topic and practical application of wearable technologies for physiological monitoring. Following rigorous peer review, eight original research papers were selected for inclusion in this Research Topic.

The work of Brady et al. aimed to explore foundational capabilities and feasibility of wearable physio-logging for remote monitoring. Recent advances in wearable technologies have expanded the scope of physio-logging in biomedical applications. Studies demonstrate the feasibility of remote health monitoring using apps like Labs Without Walls and devices such as the Apple Watch, with high adherence across diverse age groups. In a clinical trial conducted over 8 weeks, participants provided high-quality passive and active data with strong adherence and usability.

The findings support wearable-based physio-logging as a scalable, user-friendly approach for decentralized health monitoring, highlighting its potential for broad population studies and future enhancements through gamification and improved survey design.

Understanding the validity of wearable sensors to measure specific metrics play a crucial role for clinical adoption. In this line of work, Icenhower et al. conducted validation studies showing PPG-based heart rate measurements are largely unaffected by skin tone, however, accuracy declined during rapid activity changes compared to ECG readings. These results support the reliability of PPG across diverse populations, while highlighting the need for continued validation of wearable devices under dynamic conditions to ensure equitable and accurate physio-logging in biomedical applications.

With a focus on translating technology to market for low-resource settings, Mendt et al. assessed consumer-grade wearables against research-grade devices during physical activity. While the consumer grade device performed well for heart rate at low intensity, its accuracy declined with exertion. Step count, energy expenditure, and temperature readings also showed limited reliability. These findings, again, highlight both the potential and limitations of consumer wearables for physio-logging, suggesting they may be useful for long-term monitoring in low-resource settings, but are not yet suitable for precise clinical applications or, at least, should be validated according to their intended use-case.

In the spirit of interdisciplinary research combining biomedical tech and immersive media, the human factors research of Medarević et al. assessed the ability of two wearable devices—the Empatica E4 and Faros 360—to detect physiological distress in interactive virtual reality (VR) environments. Using heart rate metrics, both devices successfully identified distress, particularly during interactive VR scenes. The Faros 360 showed superior signal quality and consistency, though both devices demonstrated good agreement in heart rate measurement. These findings highlight the potential of wearable physio-logging tools in immersive settings, supporting their use in adaptive VR therapies and user experience optimization. The study also underscores the importance of device-specific performance in accurately capturing emotional and physiological states.

Practical application of wearables must also anticipate use in extreme environments Pernett et al. concluded that a custom-made chest strap equipped with strain gages, similar to consumer-grade devices, still faces limitations in accuracy under physical stress. Specifically, the study evaluated the ability of their force sensor to estimate hyperventilation risk in freedivers by predicting end-tidal CO_2 levels. Data from 21 athletes showed that chest movement amplitude and respiratory rate could explain 34% of CO_2 variability, suggesting potential for detecting unintentional hyperventilation—a key blackout risk. These findings highlight the promise of wearable physio-logging for enhancing safety in high-risk sports, while underscoring the need for further validation and improved algorithms for freediving safety.

In a dynamic field such as wearables, technical innovations and modeling abound, Ogata et al. developed a personalized method for estimating energy expenditure during heavy physical labor using wearable accelerometers and heart rate sensors. By calibrating individual models, the combined approach significantly outperformed accelerometer-only estimates. These findings underscore the limitations of single-sensor systems and highlight the value of multimodal wearables in extreme environments. These results may lead to improved health and nutrition planning for disaster relief teams and advancing physio-logging applications in real-world, high-stress scenarios.

Machine learning is increasingly featured in wearables research and Kishor Kumar Reddy et al. introduces a ResNet-LSTM deep learning model for non-invasive blood pressure estimation using ECG and PPG signals. Designed for Smart Health Monitoring in remote or underserved areas, the model achieved high accuracy despite greater computational demands. Its strong performance across datasets highlights the potential of AI-powered wearable physio-logging to enhance real-time cardiovascular monitoring and address limitations of traditional cuff-base blood pressure measurement. In a similar way, Dervieux et al. aimed to explore predictions and limitations of models for non-heated transcutaneous CO_2 sensors that have the potential to allow more accessible, real-time monitoring of arterial CO2 partial pressure in clinical and remote settings to monitor patients severe respiratory disorders. This study examined how skin temperature influences transcutaneous CO_2 diffusion, a key factor in wearable capnometer performance. In 40 adults, skin conductivity, CO_2 exhalation, and blood flow increased significantly at 35–38°C—temperatures achievable without active heating. These findings support the feasibility of non-heated $tcpCO_2$ sensors, advancing wearable physio-logging by addressing a major limitation in current monitors and enabling more practical, continuous respiratory tracking in both clinical and remote settings.

Collectively, these contributions underscore the growing reliability, versatility, and challenges of wearable physio-logging, paving the way for more inclusive, adaptive, and scalable health technologies.

Author contributions

CL: Writing – original draft, Writing – review and editing. MY: Writing – review and editing. CD: Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

We thank the authors of the papers published in this Research Topic for their valuable contributions and the referees for their rigorous review. We also thank the editorial board of the Frontiers in Physiology and the Frontiers specialists for their support.

Conflict of interest

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

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

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

frontiers | Frontiers in Physiology

TYPE Original Research
PUBLISHED 23 January 2024
DOI 10.3389/fphys.2023.1293752

Check for updates

OPEN ACCESS

EDITED BY
Xudong Ge,
University of Maryland, Baltimore County, United States

REVIEWED BY
Venkatesh Srinivasan,
University of Maryland, Baltimore County, United States
Kenta Iitani,
Tokyo Medical and Dental University, Japan

*CORRESPONDENCE
Emmanuel Dervieux,
emmanuel.dervieux@biosency.com

RECEIVED 13 September 2023
ACCEPTED 05 December 2023
PUBLISHED 23 January 2024

CITATION
Dervieux E, Guerrero F, Uhring W, Giroux-Metgès M-A and Théron M (2024), Skin temperature influence on transcutaneous carbon dioxide (CO_2) conductivity and skin blood flow in healthy human subjects at the arm and wrist.
Front. Physiol. 14:1293752.
doi: 10.3389/fphys.2023.1293752

Skin temperature influence on transcutaneous carbon dioxide (CO_2) conductivity and skin blood flow in healthy human subjects at the arm and wrist

Emmanuel Dervieux[1,2,3]*, François Guerrero[2], Wilfried Uhring[3], Marie-Agnès Giroux-Metgès[2,4] and Michaël Théron[2]

[1]Biosency, Cesson-Sévigné, France, [2]EA4324-ORPHY, Univ Brest, Brest, France, [3]ICube, University of Strasbourg and CNRS, Strasbourg, France, [4]Explorations Fonctionnelles Respiratoires, Centre Hospitalier Régional et Universitaire de Brest, Brest, France

Objective: present transcutaneous carbon dioxide (CO_2)—tcpCO_2—monitors suffer from limitations which hamper their widespread use, and call for a new tcpCO_2 measurement technique. However, the progress in this area is hindered by the lack of knowledge in transcutaneous CO_2 diffusion. To address this knowledge gap, this study focuses on investigating the influence of skin temperature on two key skin properties: CO_2 permeability and skin blood flow.

Methods: a monocentric prospective exploratory study including 40 healthy adults was undertaken. Each subject experienced a 90 min visit split into five 18 min sessions at different skin temperatures—Non-Heated (NH), 35, 38, 41, and 44°C. At each temperature, custom sensors measured transcutaneous CO_2 conductivity and exhalation rate at the arm and wrist, while Laser Doppler Flowmetry (LDF) assessed skin blood flow at the arm.

Results: the three studied metrics sharply increased with rising skin temperature. Mean values increased from the NH situation up to 44°C from 4.03 up to 8.88 and from 2.94 up to 8.11 $m{\cdot}s^{-1}$ for skin conductivity, and from 80.4 up to 177.5 and from 58.7 up to 162.3 $cm^3{\cdot}m^{-2}{\cdot}h^{-1}$ for exhalation rate at the arm and wrist, respectively. Likewise, skin blood flow increased elevenfold for the same temperature increase. Of note, all metrics already augmented significantly in the 35–38°C skin temperature range, which may be reached without active heating—*i.e.* only using a warm clothing.

Conclusion: these results are extremely encouraging for the development of next-generation tcpCO_2 sensors. Indeed, the moderate increase (× 2) in skin conductivity from NH to 44°C tends to indicate that heating the skin is not critical from a response time point of view, *i.e.* little to no skin heating would only result in a doubled sensor response time in the worst case, compared to a maximal heating at 44°C. Crucially, a skin temperature within the 35–38°C range already sharply increases the skin blood flow, suggesting that tcpCO_2 correlates well with the arterial paCO_2 even at such low skin temperatures. These two conclusions further strengthen the viability of non-heated $tcpCO_2$ sensors, thereby paving the way for the development of wearable transcutaneous capnometers.

KEYWORDS

carbon dioxide, $ptCO_2$, $tcpCO_2$, transcutaneous, exhalation rate, diffusion, blood flow, hyperaemia

1 Introduction

Due to its clinical significance, continuous monitoring of the arterial CO_2 partial pressure—$paCO_2$—is of paramount importance in medical practice, especially for patients presenting severe respiratory disorders (Wagner, 2015). The gold standard to get a single $paCO_2$ reading consists in an arterial puncture followed by a gaseous analysis of the collected blood sample. Unfortunately, this procedure—tersely referred to as the "blood gases" in clinical settings—is both painful and risky (Scheer et al., 2002), requiring trained personnel as well as expensive blood gas analysers. It also calls for a quick analysis of the blood samples following their collection, which adds stress to hospital logistics (Nanji and Whitlow, 1984). These major drawbacks led to the development of transcutaneous CO_2 monitors, which consist in a Stow-Severinghaus electrode—mainly a pH-meter bathing in a bicarbonate solution—heated in the 41–44°C range, and placed against a patient's skin (Severinghaus and Astrup, 1986; Huttmann et al., 2014). This electrode measures a transcutaneous CO_2 partial pressure—the $tcpCO_2$—which correlates well with the $paCO_2$ if the skin is heated above at least 38°C (Wimberley et al., 1985).

Yet, these monitors also suffer from several weaknesses: *i)* their important drift requires a recalibration with an appropriate gas mixture every 8 hours, at most (Bendjelid et al., 2005), *ii)* the thin membrane covering the electrode is fragile, and needs to be replaced every 2 weeks or so (Lermuzeaux et al., 2016), *iii)* the heating power required by the electrode to maintain the skin in the above-mentioned temperature range is about 100–200 mW[1], thus precluding its use in a battery-powered wearable stand-alone device and *iv)* their elevated price tag prevents their widespread use, would it be in a clinical or home-based setting. Consequently, the development of an alternative to the existing $tcpCO_2$ monitors appears mandatory, and has been an active research field in the last decades (Dervieux et al., 2022; Section 4.1).

Recently, in a review article aiming at encompassing the diversity of CO_2 measurement techniques with a focus on biomedical applications, we divided the issue of developing such an alternative $tcpCO_2$ monitor into three research areas (Dervieux et al., 2022):

1. Due to the above-mentioned drift, and high cost of the Stow-Severinghaus electrode, an alternative CO_2 *measurement technique* is needed.
2. Then, in order to dimension the sensor-to-be, it is essential to accurately know the CO_2 *exhalation rate* through the skin, as the latter directly influences the response time of the sensor.
3. Finally, it is mandatory that the $tcpCO_2$ and $paCO_2$ are in good agreement at the skin temperature considered for measurement. *Id est*, that the *correlation* between $tcpCO_2$ and $paCO_2$ is satisfactory at the latter temperature.

Regarding the first point, it appeared to us that, among the many technologies reviewed, a polymer patch embedding a CO_2-sensitive fluorophore would be particularly advisable. Interestingly, this trail has recently been followed by Cascales et al. (2022) or Tufan and Guler (2022) with some success, although no *in vivo* experiment have been conducted to date. This point will be the object of future studies and is not developed any further in this paper.

The second and third points, on the contrary, are at the very heart of the present study. Starting with the exhalation rate, the main issue with data available in the literature—see Table 3—is that the skin temperature is only mentioned once—by Eöry (1984)—and never accurately regulated when this parameter is measured (Fitzgerald, 1957). Even though Levshankov et al. (1983) crafted a heating device, they do not report the temperature setpoint that they used. Thus, the present study aims at filling this gap by measuring the influence of skin temperature on the transcutaneous CO_2 exhalation rate.

While the $tcpCO_2/paCO_2$ correlation is excellent at—or above—42°C (Conway et al., 2018), scarce are the authors who investigated lower skin temperatures, with none but Wimberley et al. (1985) experimenting with temperatures as low as 38°C. The reason for heating the skin in the first place is to trigger a local reactive hyperaemia (Roustit and Cracowski, 2012). By doing so, the subcutaneous tissues are flushed with fresh arterial blood, and their gaseous content thus gets closer to the arterial one (Koch, 1965; Rooth et al., 1987; Zavorsky et al., 2007). While temperatures in the 42–44°C range have often been used to trigger maximal hyperaemia, lower temperatures have been seldom explored (Hodges et al., 2016), and we thus took advantage of our exhalation rate measurements to measure the skin blood flow at lower temperatures simultaneously.

Then, the reader should bear in mind the importance of skin temperature for designing a new kind of $tcpCO_2$ sensor. Indeed—ideally—such a sensor should heat the skin as little as possible for two main reasons: *i)* heating the skin is uncomfortable for the patient and can wake them up in the case of night time monitoring, and *ii)* it consumes a significant amount of power, which precludes using such sensor in a wearable, as mentioned above. Yet, such an unnoticeable and wearable $tcpCO_2$ sensor would be highly desirable in a telemonitoring context for home use, reducing the need for hospital visits. Indeed, if the positive impact of $tcpCO_2$ telemonitoring is yet to be demonstrated—for the obvious reason that the corresponding wearable $tcpCO_2$ monitor does not exist at the time being—several clinical trials demonstrate the beneficial contribution of telemedicine—*a.k.a.* telehealth—on

[1] Typical value for a TCM4 $tcpCO_2$ monitor (Radiometer, Denmark), as measured by the authors.

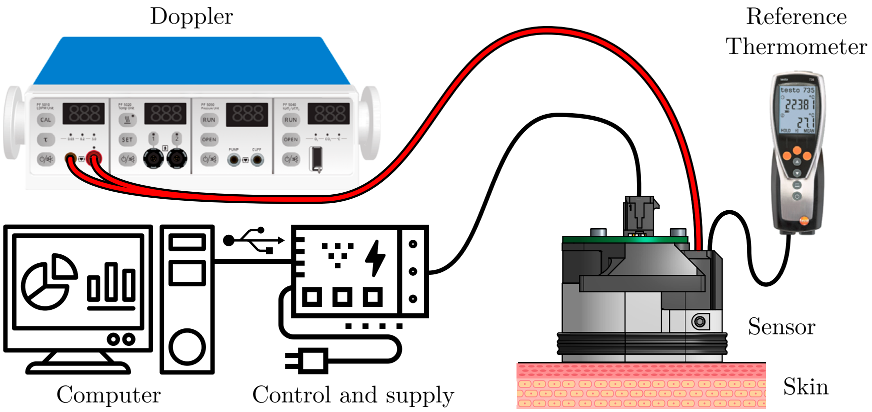


FIGURE 1
General outline of the rate sensor and its peripherals. See the text for further explanations.

both patient's outcome and costs of admission in a variety of conditions (Steventon et al., 2012; Yun et al., 2018; Kruse et al., 2019). Additionally, the outbreak of contagious pandemics—such as COVID-19 (Garfan et al., 2021)—and the rapid development of the health wearable market (Dunn et al., 2018; Yetisen et al., 2018; Chung et al., 2019; Dagher et al., 2020) may also promote the use of telemonitoring in medical practice. For these reasons, *not* heating the skin while measuring tcPCO_2 would be highly desirable.

The following study focuses on measuring the transcutaneous CO_2 exhalation rate and cutaneous micro-circulation on the full NH–44°C skin temperature range. Two measurement sites were investigated: the dorsal side of the wrist and the lateral aspect of the upper arm, while the skin blood flow was only measured at the upper arm. Additionally, a strong emphasis was placed on the transcutaneous CO_2 *conductivity*, which may be preferred to the well-known exhalation rate because of its more intrinsic nature—in particular, the latter conductivity does not depend on the ambient CO_2 level, nor on the subject's paCO_2, as opposed to the exhalation rate, which is influenced by both.

2 Materials

2.1 The transcutaneous CO_2 rate sensor

A custom transcutaneous CO_2 diffusion rate sensor—hereafter simply denoted as "the sensor"—was developed for the needs of this study. Its basic working principle is close to that evoked by Dervieux et al. (2022)—and will be further detailed in Section 3.1.1—while its design is inspired by the early works of Eletr et al. (1978) and Greenspan et al. (1981). The general outline of the sensor and its peripherals can be seen in Figure 1.

The sensor, designed to be placed against the subject's skin by mean of a double-sided adhesive, is connected to three main apparatuses: *i)* a calibrated, reference thermometer (Testo 735, Testo, Germany) equipped with a type K thermocouple (110-4482, RS Pro, United Kingdom), *ii)* a Doppler perfusion monitor (Periflux 5000, Perimed, Sweden) equipped with a 407 probe, and *iii)* a control and supply block, consisting in a thermostat, a power supply unit, and a Universal Serial Bus (USB) to universal Asynchronous Receiver Transmitter (UART) converter, embedded in a 3D-printed case. For the sake of conciseness though, the control and supply block is only detailed in the Supplementary Material S1, which also contains a thorough analysis of the safety issues that may arise when using this sensor. The sensor itself can be seen in great details in Figure 2. It consists in an aluminium (2017A) body, which serves as a support for the following elements: a CO_2 sensor, a heating resistive wire, a thermistor, a thermocouple, the Doppler probe, a poly-lactic acid (PLA) 3D-printed cover, and an interfacing Printed Circuit Board (PCB). Complete drawings of the sensor's body are provided in the Supplementary Material S1.

The CO_2 sensor is a MinIR (ExplorIR–M5%, CO_2Meter, United States), an off-the-shelf, compact, Non Dispersive Infra-Red (NDIR) CO_2 sensor, with a full range of 5% and an accuracy of 70 ppm ± 5% of reading at Standard Temperature and Pressure (STP)—see Hodgkinson and Tatam (2012) for further details on the operating principle of such sensors. Its internals—a pair of IR Light Emitting Diode (LED) and photodiode (PD) facing a spherical, gold-plated mirror—may be seen in the cut view of Figure 2. The pCO_2 inside the sensor was recorded with a sampling frequency of 2 Hz.

As the gas-tightness of the measuring chamber is a critical aspect of the sensor's operating principle—see Section 5.1.3—a two-stage sealing was implemented: *i)* a silicone-grease coated (515520, GEB, France), soft—60 Shore A hardness—silicone O-ring was placed between the aluminium body and the CO_2 sensor itself and *ii)* liquid epoxy resin (Résine Cristal, Gédéo, France) was cast in the remaining interstice between the latter two elements—illustrated in the cut view of Figure 2 in vivid red.

The thermoregulation of the sensor is performed by mean of a resistive wire for heating, coupled to a thermistor for temperature measurement and regulation. For verification purpose, an additional thermocouple was also added, as stated above. The heating wire consists in 2×15 turns of $28\ \Omega{\cdot}m^{-1}$, 0.15 mm in diameter, enamelled, resistive, constantan wire (Isotan, Thomsen), connected in parallel. The wire delivers a total heating power of 6.1 W under 12 V, and is coiled around the aluminium body, in a dedicated groove. The bottom of the groove is covered with a layer of 0.25 mm thermally conductive double sided tape (8810, 3M, United

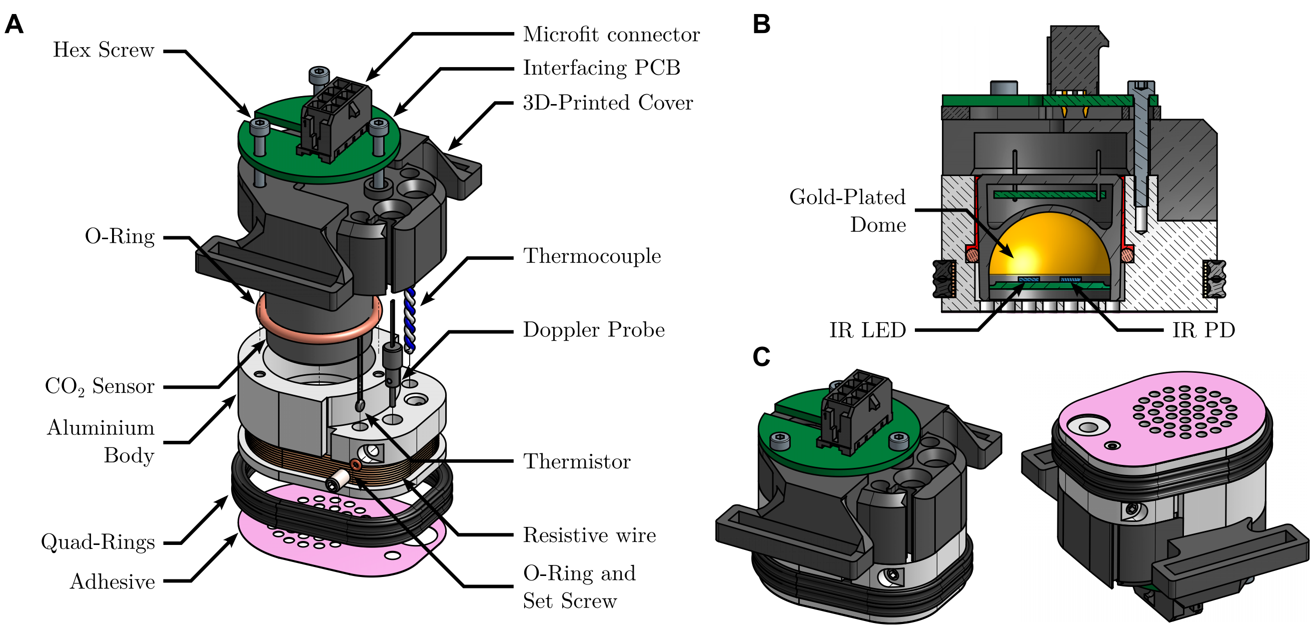


FIGURE 2
Detailed views of the sensor. **(A)**: exploded view, detailing its different parts. **(B)**: cut view, showing the inner functioning principle of the CO_2 sensor. Note the epoxy resin sealing, in red. **(C)**: isometric views from above, and below, of the fully assembled sensor, illustrating the grid-shaped sensor's sole.

States) prior to coiling the wire, and the latter is finally covered with two nitrile quad rings, as can be seen in Figure 2. This covered layout prevents burns caused by direct contact with the heating wires. The thermistor (151-237, RS Pro, United Kingdom) and thermocouple were glued in two dedicated flat-bottom mounting holes which were pre-filled with a thermally conductive, electrically non-conductive, epoxy resin (8329TFM, MG Chemicals, Canada). Care was taken that *i)* the distance between the bottom of the mounting holes and the heating wire and *ii)* that between the sole of the sensor's body and the heating wire were equal, in order to ensure that the temperature measured by the thermistor and thermocouple is close to that of the skin.

The Doppler probe is housed in a dedicated hole, and can slide vertically, in such a way that it can be adjusted to outcrop the sole of the sensor, coming in direct contact with the skin. It holds in place by mean of a cup-pointed, headless, set screw which compresses it radially via an O-ring, so as not to damage the probe. The raw Doppler perfusion signal—originally sampled at 62.5 Hz—was downsampled to 0.625 Hz and low-pass filtered using a tenth-order Butterworth filter prior to further analysis—see Section 3.1.2.

The 3D-printed cover and interfacing PCB were added for usability purposes: the 3D-printed PLA cover allows to attach a strap (HTH 833 with H83 hooks, Velcro, United Kingdom) to the sensor in order to maintain it against a subject's skin, as illustrated in Figure 3B, while the interfacing PCB gathers the four UART pins from the CO_2 sensor, the two ends of the thermistor, and those of the heating wire into a single eight-pins Microfit connector (0430450812, Molex, United States).

The adhesive itself consists in a disposable laser-cut, double-sided, clinical-grade tape (1567, 3M, United States). For ease of application, a special tooling was developed to accurately align the sensor and the adhesive together—see Supplementary Material S1.

2.2 Reference tcpCO_2 monitor

In addition to the above-detailed custom-made sensor, a clinical-grade tcpCO_2 monitor (TCM4, Radiometer, Denmark) was also used on the upper deltoid—one of the recommended sites for tcpCO_2 monitoring (SenTec, 2016)—yielding a continuous reference tcpCO_2 reading. The tcpCO_2 sensor itself (tc Sensor 54) was affixed to the skin using an appropriate attachment ring and contact gel, and it was re-membraned and re-calibrated when needed, as per the manufacturer's guidelines. All the accessories used to this end were Radiometer's (Radiometer, 2020).

2.3 Sensors positioning

The different sensors and measurement sites chosen in the study are illustrated in Figure 3A. All sensors were placed on the subject's left arm: the reference tcpCO_2 monitor was placed on the upper deltoid, as mentioned above, while two custom rate sensors were positioned as follows. The first one was equipped with the Doppler probe, and was attached on the distal side of the upper arm, immediately below the deltoid, at the junction point between the upper part of the biceps, the lower end of the deltoid, and the triceps. The second rate sensor was placed on the dorsal side of the wrist, and did not include a Doppler probe. Both sensors were affixed to the subject's skin by mean of the above-mentioned double-sided adhesive, and secured in place with a Velcro strap. Additionally, the subject's arm laid comfortably onto an arm gutter so that it remains still and relaxed for the whole duration of the experiments.

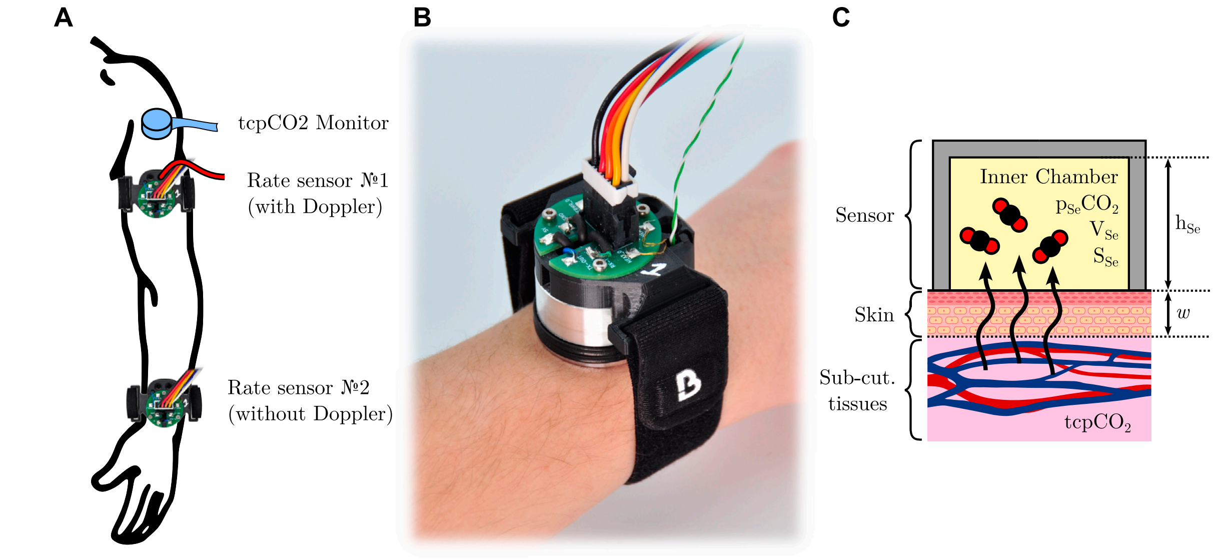


FIGURE 3
(A): outline of the sensors used in this study, and their location. **(B)**: a picture of the sensor with its connection cable assembly and strap, attached on a wrist. **(C)**: simplified model of CO_2 diffusion through the skin inside a closed sensor.

3 Methods

3.1 Measured metrics

3.1.1 Skin CO_2 conductivity and exhalation rate

The rate of diffusion—*a.k.a.* the exhalation rate—of CO_2 through the skin per unit of area—hereafter noted Q, of dimension $L^3 \cdot L^{-2} \cdot T^{-1}$—can be measured by affixing to the skin a cup-like device, which entraps the skin-exhaled CO_2. In this situation, the CO_2 diffusion through the skin can be modelled as presented in Figure 3C. Briefly, the skin is considered as a CO_2-permeable membrane of thickness w and diffusivity D towards CO_2 (unit of $m^2\ s^{-1}$). The partial CO_2 pressure inside the sub-cutaneous tissues and inner chamber are $tcpCO_2$ and $p_{Se}CO_2$, respectively, and the sensor area in contact with the skin is S_{Se}, while its equivalent height and volume are h_{Se} and V_{Se}. It can be shown under certain hypotheses (Dervieux et al., 2022) that:

$$\frac{dp_{Se}CO_2}{dt} = \frac{D}{w \cdot h_{Se}} \cdot \left(tcpCO_2 - p_{Se}CO_2\right) \tag{1}$$

Leading to a first-order behaviour for $p_{Se}CO_2$, given by:

$$p_{Se}CO_2(t) = tcpCO_2 \cdot \left(1 - e^{-\frac{t}{\tau}}\right) + p_{Se}CO_2(t=0) \cdot e^{-\frac{t}{\tau}}, \quad \text{with } \tau = \frac{h_{Se} \cdot w}{D} \tag{2}$$

And Q is then equal to:

$$Q(t) = \frac{D}{w \cdot P_0} \cdot \left(tcpCO_2 - p_{Se}CO_2(t=0)\right) \cdot e^{-\frac{t}{\tau}} \tag{3}$$

wherein P_0 is the total atmospheric pressure at measurement site. However, since Q depends on both the ambient level of CO_2—*via* $p_{Se}CO_2(t=0)$—and the subject's capnia—*via* $tcpCO_2$—we introduced the skin *conductivity*—from the thermodynamic or electrical analogy—expressed in $m \cdot s^{-1}$, and defined as:

$$\boxed{K = \frac{D}{w}} = \frac{P_0 \cdot Q(t=0)}{tcpCO_2 - p_{Se}CO_2(t=0)} \tag{4}$$

Contrary to Q, K is an intrinsic property of the skin and is independent of the $tcpCO_2/p_{Se}CO_2$ gradient. Additionally, deriving Eq. 2 and evaluating it at $t = 0$ yield:

$$K = \frac{h_{Se}}{tcpCO_2 - p_{Se}CO_2(t=0)} \cdot \left.\frac{dp_{Se}CO_2}{dt}\right|_{t=0} \tag{5}$$

In practice, K was thus measured as follows, choosing arbitrarily $t = 0$ at each temperature change:

- h_{Se} is known by dividing V_{Se} by S_{Se}. The latter is known by construction of the sensor's aluminium body, while V_{Se} was estimated by filling a clone of the sensor with a low viscosity fluid—pure ethanol—and weighting it.
- the subject's $tcpCO_2$ was measured using the above-mentioned reference medical grade monitor. The extraction of a single $tcpCO_2$ reading for each temperature set-point is detailed in Supplementary Material S1.
- $p_{Se}CO_2(t=0)$ could be measured with a simple reading of the CO_2 sensor.
- $\left.\frac{dp_{Se}CO_2}{dt}\right|_{t=0}$ was estimated by fitting a linear regression on the measured $p_{Se}CO_2$. The latter regression was performed on the $p_{Se}CO_2$ data starting 3 min after a temperature change—to allow for temperature homogenisation of the different inner parts of the sensor—and up to 18 min after, for a total of 15 min of data. This duration was chosen following a preliminary study performed on ten subjects, which yielded R^2 regression scores above 0.95 for a regression duration above 700 s (about 12 min).

In summary, the diffusion of CO_2 through the skin was quantified by the skin CO_2 conductivity K, which was measured

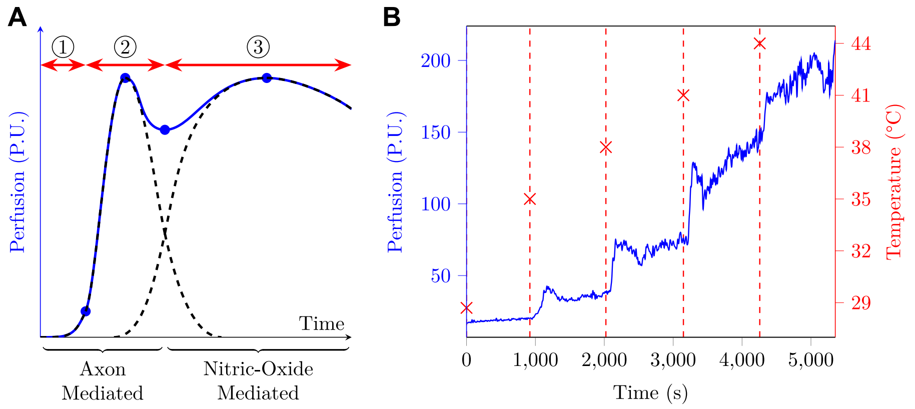


FIGURE 4
(A): typical skin perfusion response to local heating, based on data from different sources (Minson et al., 2001; Cracowski et al., 2006; Minson, 2010; Frantz et al., 2012; Roustit and Cracowski, 2012). If we suppose the heat stress to be applied at time origin, three phases are usually observed: ① an onset lag corresponding to *i)* the heating time required by the sensor to reach its set-point temperature and *ii)* the time taken by the inner skin layers to reach this temperature and trigger the axon-mediated hyperaemia. ② the axon-mediated hyperaemia which rises quickly and then fades away. ③ the nitric-oxide mediated hyperaemia, whose onset is slower and which slowly fades away if the temperature set-point is not too elevated. **(B)**: perfusion and skin temperature as a function of time as measured at the arm of a test subject.

at five different temperatures, each temperature corresponding to a 18 min measurement window for a total of 90 min of acquisition per subject, as detailed in Section 3.2, below. Additionally, Equation 4 was used to compute the corresponding equivalent $Q\ (t = 0)$ with $p_{Se}CO_2(t = 0) \approx 0$—*i.e.* the skin CO_2 exhalation rate in free air as commonly referred to in the literature. Of note, this $Q\ (t = 0)$ was **not** observed in practice, since the sensor was left in place—and thus $p_{Se}CO_2(t = 0) \neq 0$ for most temperatures. For the sake of conciseness, in the remainder of this article, the letter Q alone or the mention of "skin CO_2 exhalation rate" without further indications always designate the above-mentioned $Q\ (t = 0)$. Finally, it should be noted that these Q values were derived mainly as a mean to compare with the existing literature, and that the actual statistical analyses were performed on K—see Section 3.3.

3.1.2 Skin blood flow

The skin blood flow—*a.k.a.* (sub)cutaneous micro-circulation or perfusion—was measured using LDF, and expressed in Perfusion Units (P.U.), a dimensionless arbitrary unit that reflects both the amount and the speed of moving elements—mainly erythrocytes—seen by the Doppler probe (Bonner and Nossal, 1990). When the skin temperature rises, perfusion increases, a phenomenon known as heat-triggered—or thermal—reactive hyperaemia (Minson, 2010), whose dynamics is illustrated in Figure 4A. The respective durations of phases ①–③ were not specified in abscissa since they may vary markedly depending on the heating rate and temperature (Magerl and Treede, 1996; Del Pozzi et al., 2016). To give the reader an order of magnitude, phase ① usually lasts a few minutes, phase ② from 5 up to 10 min, and phase ③ from 30 up to 60 min (Minson et al., 2001; Cracowski et al., 2006; Minson, 2010; Frantz et al., 2012; Roustit and Cracowski, 2012).

Such a behaviour calls for some kind of data processing to yield a single representative perfusion metric for the initial bump, after-bump nadir, and final plateau. In this paper, $\text{SkBF}_{90}(\text{T})$ was defined as the 90th percentile of the measured skin blood flow—SkBF—on an 18 min window at temperature T. This choice was made following preliminary measurements at the arm, an example of which is plotted in Figure 4B. The latter clearly exhibits five perfusion plateaux corresponding to the five temperature set points, and one can also distinguish a small initial bump at the onset of a new temperature, which is especially visible at 35, 38 and 41°C. Due to *i)* the high variability exhibited by the measured perfusion, especially at high temperature, *ii)* the fact that the nitric-oxide phase can take up to 30–40 min to establish (Barcroft and Edholm, 1943; Taylor et al., 1984; Minson et al., 2001; Frantz et al., 2012; Del Pozzi et al., 2016), and *iii)* the fact that each temperature window lasts 18 min, it seemed a good strategy to choose a metric which was more robust than the mean—*e.g.* the *n*th percentile—and focused on the very end of the observation window. In this regard, $\text{SkBF}_{90}(\text{T})$ seemed to meet the latter requirements and was therefore chosen as perfusion metric. It was additionally normed by the maximal perfusion measured on a given subject—*i.e.* the SkBF_{90} value measured at 44°C. Finally, the LDF metric used throughout this study is thus:

$$\boxed{\text{nSkBF}_{90}(\text{T}) = \frac{\text{SkBF}_{90}(\text{T})}{\text{SkBF}_{90}(44°\text{C})}} \tag{6}$$

3.2 Protocol design

The clinical study was interventional, monocentric, and involved 40 healthy human subjects. Inclusion criteria were an age between 18 and 80, and having given a free and informed consent. Exclusion criteria were the presence of cutaneous lesions at the measurement sites or skin conditions such as dermatitis or psoriasis, and taking vasodilator therapy. The research was approved by the local ethics committee (Comité de protection des personnes Sud-Méditerrannée II, IDRCB ref.: 2020-A02185-38), registered on Clinical Trials (NCT05637138), and it was carried out in accordance with the declaration of Helsinki.

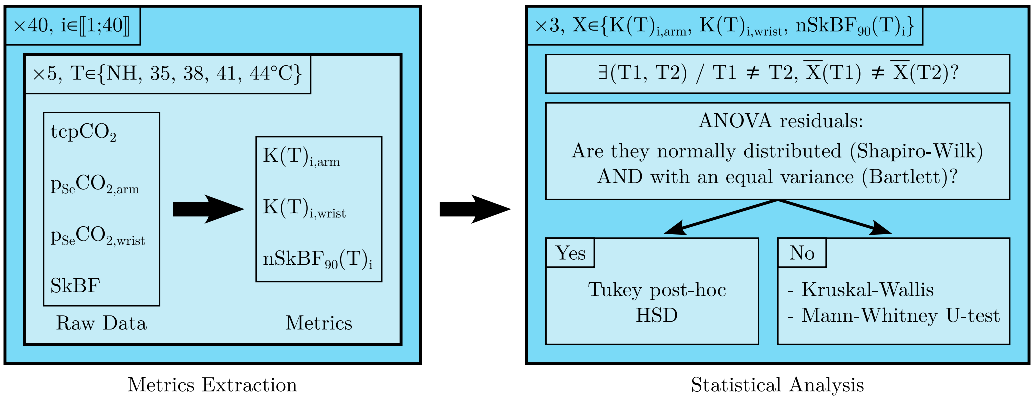


FIGURE 5
Graphical representation of the data analysis workflow, see the text for further details.

After a preliminary visit during which the subjects were informed of the study modalities and gave consent, all measurements were performed during a single visit, during which the subjects were seated. This visit began with a visual inspection of the measurement sites for detection of cutaneous lesions. The sites were then shaved if needed for a good adhesion of the sensors, using an electric trimmer in order to avoid skin inflammation. The skin was then degreased and cleaned using isopropyl alcohol, and the three above-mentioned sensors were attached to their respective measurement sites. These preliminary steps also allowed for subject acclimation and lasted about 5 min. The measurement itself then began, consisting of five 18 min periods, corresponding to five temperatures for the two rate sensors: NH, 35, 38, 41, and 44°C. At the end of the 90 min measurement period, the sensors were gently peeled off, and the skin was cleaned again. All $tcpCO_2$ and $p_{Se}CO_2$ data were recorded on computers for future analysis, and the room temperature was also recorded using a calibrated thermometer (Testo 735, Testo, Germany).

3.3 Data analysis

The data analysis workflow is summarised in Figure 5. Raw data were collected for all 40 subjects at five different temperatures and three metrics were extracted for each ith subject/temperature pair: K at the arm and wrist, and $nSkBF_{90}$ at the arm only. For each of those metrics, an ANOVA was performed across all subjects to determine whether their mean values differ significantly between two temperatures. If the ANOVA residuals did not significantly differ from a normal distribution—according to Shapiro-Wilk testing—and if the hypothesis of variance equality between the temperature groups was also verified—according to Bartlett testing—a Tuckey *post hoc* HSD test was then performed. Otherwise, a Kruskal-Wallis test followed by a series of Mann-Whitney U-tests were performed. Additionally, Pearson and Spearman correlation tests were also carried out to study the influence of temperature on the three afore-mentioned metrics (not represented in Figure 5). When applicable, all tests were two-sided and a 5% alpha risk was chosen as significance threshold.

4 Results

4.1 Demographics and temperatures

The fourty subjects consisted in 24 men and 16 women, aged between 20 and 61 years (mean/median/Standard Deviation (SD): 40/39/13 years). The laboratory temperature was in the 20.1–22.7°C range for the whole duration of the experiments (mean/median/SD: 21.3/21.2/0.7°C). The skin temperature during the non-heated 18 min phase was measured twice, at 10 and 18 min after sensor application, and these two measurements were averaged to yield a single temperature value per subject. The latter was in the 27.1–31.8°C range (mean/median/SD: 29.3/29.2/1.2°C) at the arm, and in the 24.7–33.1°C range (mean/median/SD: 28.6/28.5/1.7°C) at the wrist.

4.2 Skin CO_2 conductivity

The linear regressions leading to K values—see Section 3.1.1—yielded excellent regression coefficients, with average R^2 values of 0.98 and 0.96 at the arm and wrist, respectively. The resulting skin conductivities are summarised in Figure 6, and show a sharp tendency to increase with an increasing skin temperature. Interestingly, the five upper outlying values at the arm—at all temperatures—and the four upper and below outlying values at the wrist—in the 35–44°C range—belonged each time to a single subject, who exhibited an especially high, or low skin conductivity. However, these two latter subjects were not one and the same person at the arm and at the wrist. The dispersion of skin conductivity values is also glaring with max/min ratios for a given skin temperature/location pair in the 2.9–9.0 range.

A normalisation of the variable K—using the change of variable $K \mapsto \sqrt{K}$—was carried out prior to the ANOVA, resulting in the Shapiro-Wilk and Bartlett tests to be passed (all p-values above 0.05). The ANOVA was significant (p-value below 10^{-15}) and the results of the following Tukey *post hoc* HSD test are detailed in Table 1. The apparent increase in K with an increasing skin temperature seen

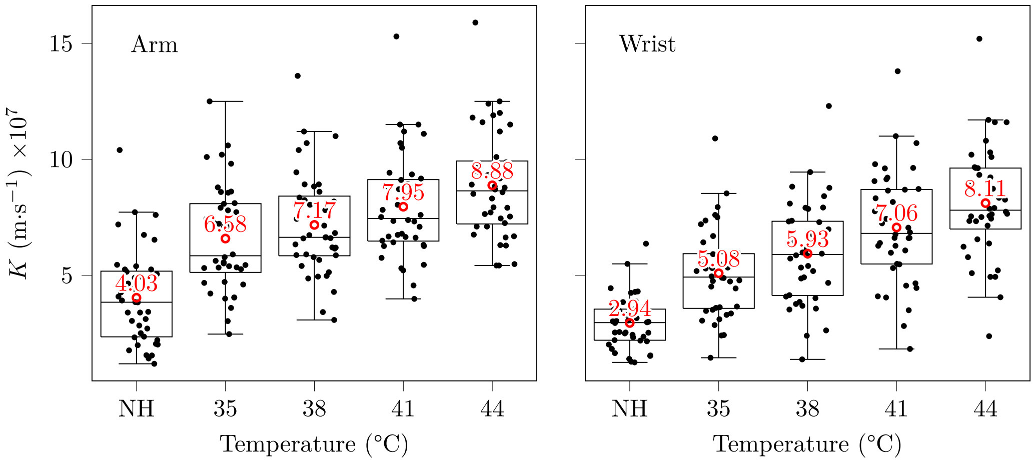


FIGURE 6
Skin conductivities at the arm and wrist. Each black mark corresponds to a subject/temperature pair, the red circles and texts indicate mean values, and some horizontal jitter was added to the black marks for legibility. The whiskers extend at most to 1.5 times the interquartile range, and descriptive statistics—range, and SD—are provided in Supplementary Material S1—two properties shared by the three box plots of the present paper.

TABLE 1 p-values for the Tuckey HSD *post hoc* test for differences of the mean *K* at different temperatures.

Arm	NH	35°C	38°C	41°C
35°C	$< 10^{-6}$†	-	-	-
38°C	$< 10^{-6}$†	0.72	-	-
41°C	$< 10^{-6}$†	0.05†	0.54	-
44°C	$< 10^{-6}$†	$< 10^{-3}$†	0.01†	0.44

Wrist	NH	35°C	38°C	41°C
35°C	$< 10^{-6}$†	-	-	-
38°C	$< 10^{-6}$†	0.35	-	-
41°C	$< 10^{-6}$†	$< 10^{-3}$†	0.14	-
44°C	$< 10^{-6}$†	$< 10^{-6}$†	$< 10^{-3}$†	0.73

Values with a † are considered significant with a risk $\alpha = 0.05$. (Only the lower-left part of the tables are filled-in for the sake of clarity).

in Figure 6 is hereby confirmed, with significantly different mean K values for most temperature differences except the closest ones—*i.e.* for the 35/38, 41/38 and 44/41°C pairs.

Pearson and Spearman correlations were both significant (p-values below 10^{-15}) with correlation coefficients of 0.60 and 0.59 at the arm, respectively, and 0.66 and 0.67 at the wrist. These coefficients indicate a moderate positive influence of the skin temperature on its diffusivity towards CO_2.

4.3 Skin CO_2 exhalation rate

In order to provide a more accessible parameter than K for the reader, as well as to allow direct comparison with existing literature, equivalent initial exhalation rates $Q\,(t = 0)$ were also computed using Eq. 4. The resulting values are presented in Table 2.

Although our results tend to indicate a higher CO_2 exhalation rate at the upper arm than at the wrist—a MANOVA was performed considering the measurement site as independent variable and the five Q_T as dependent variables, and yielded a p-value of 0.01 using Pillai's trace—the size of this effect is moderate, especially in view of the wide Q dispersion.

4.4 Laser Doppler Flowmetry

nSkBF$_{90}$ values were computed as described in Section 3.1.2, and are presented in Figure 7. Skin perfusion exhibits a strong increase with temperature, with a tenfold multiplication between the NH basal state and the maximum vasodilation 44°C stage. In particular, a mild heating to a temperature of 38°C already entails a fourfold increase in perfusion. As was the case for skin conductivity, the dispersion of the nSkBF$_{90}$ values is also considerable, with max/min ratios for a given temperature in the 2.2–9.7 range.

Regarding statistical analyses, the normality and variance homogeneity hypotheses could not be verified regardless of the changes of variable performed. A Kruskal-Wallis test was thus performed, followed by a series of Mann-Whitney U-tests, all of which proved significant (all p-values below 10^{-10}).

Pearson and Spearman correlations were both significant (p-values below 10^{-15}) with correlation coefficients of 0.90 and 0.96, respectively. These coefficients indicate a strong positive influence of the skin temperature on its perfusion. The fact that Pearson's correlation is below Spearman's is not surprising since the relationship between skin temperature and nSkBF$_{90}$ is strongly non-linear, as emphasised by the sigmoid fit performed in Figure 7.

TABLE 2 *Q* (*t* = 0) values, as computed using Eq. 4, expressed in $cm^3 \cdot m^{-2} \cdot h^{-1}$.

Arm	Mean	Range	SD
NH	80.4	25.1–191.7	40.3
35°C	130.3	52.4–227.4	43.5
38°C	142.7	63.0–248.1	41.4
41°C	159.9	81.9–280.2	42.8
44°C	177.5	104.6–288.4	42.5

Wrist	Mean	Range	SD
NH	58.7	24.2–141.1	22.7
35°C	100.9	27.7–220.6	38.3
38°C	118.2	27.1–248.0	42.8
41°C	141.8	36.0–282.1	47.2
44°C	162.3	46.5–300.0	45.9

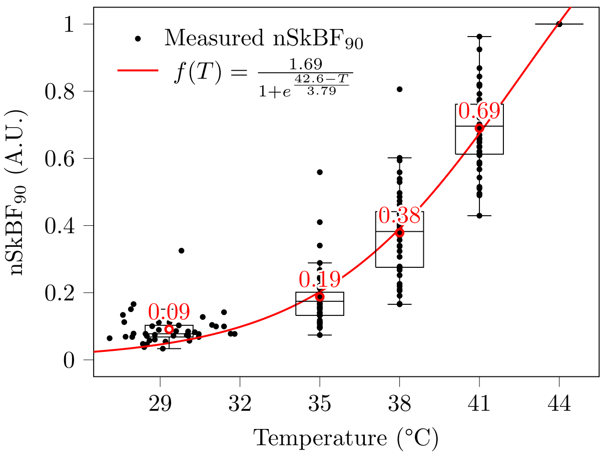


FIGURE 7
Measured nSkBF$_{90}$ at the arm. Note that since nSkBF$_{90}$ is normalised with respect to SkBF$_{90}$ at 44°C, all subjects merge into a single unitary value at this latter temperature. Each black mark corresponds to a subject/temperature pair, the red circles and texts indicate mean values, and the red curve is a least square sigmoid fit. Contrary to Figure 6, no horizontal jitter was added to the data, and the dispersion observed in the 27–32°C range corresponds to the inter-subject variability in non-heated skin temperatures.

5 Discussion

The main objectives of this research were to ascertain the influence of skin temperature on *i)* its permeability towards CO_2—through the study of the skin CO_2 conductivity *K*, and exhalation rate *Q*—and *ii)* the skin blood flow—through nSkBF$_{90}$.

5.1 Sensor design

5.1.1 Skin contact surface

Although the sensor's aluminium body was precisely machined following the drawing given in Supplementary Material S1, the exact surface area in contact with the skin that participates in gaseous exchange may slightly vary from one subject to another. Indeed, at each hole of the sensor's sole, the skin forms a small dome, whose convexity is essentially function of the mechanical properties of the skin. Yet, those mechanical properties are sex-, moisture-, age-, and temperature-dependant (Salter et al., 1993; Held et al., 2018), thereby introducing small intra- or inter-subject variations in skin contact surface area. Since this area is used to calculate *K* through S_{Se}—see Section 3.1.1—the latter is in turn influenced by these small variations.

While this would likely not change the conclusions of the present study given the order of magnitude of the above-described phenomenon reported in the literature, future research could look into replacing the grid-shaped sole that we used by a thin metal mesh, or a metallic foam. These two latter techniques were for instance implemented by Eletr et al. (1978), McIlroy et al. (1978), and Hansen et al. (1980). However, it must be emphasised that the shape of the sole of a thermally-regulated transcutaneous exhalation rate sensor is essentially a compromise between: *i)* the degree of perforation or porosity of the surface, which should be as high as possible to ensure a large diffusion surface, and *ii)* heat transfer considerations, which call for a plain, dense, surface, with a minimum number of holes, to ensure temperature homogeneity of the skin. Additionally, while the use of a wire mesh, or metallic foam reduces the above-mentioned "dome effect", it also makes the surface estimation more tedious. Therefore, this aspect of the sensor design should be further investigated to find a satisfactory technical solution which addresses the above concerns.

5.1.2 CO_2 sensor choice

The choice of the selected NDIR CO_2 sensor was mainly motivated by its compact form factor and ease of implementation. Additionally, the 5% range was especially adapted for CO_2 diffusion rate measurement. Indeed, tcpCO_2 in healthy subjects is typically in the 35–45 mmHg range (Rithalia et al., 1984), corresponding to 4.6–5.2% of CO_2. Since the CO_2 diffusion rate measurements taking place in the present study were only limited to the first moments of CO_2 diffusion from the skin into the sensor—see Section 3.1.1—measured CO_2 fraction values stayed below 1–2%. The 5% range was thus adapted to our need.

5.1.3 Gas tightness

As mentioned in Section 2.1, the gas tightness of the sensor's chamber with respect to ambient air was of paramount importance for the success of the study. Indeed, any leak of inner-chamber CO_2 towards the outer air would subtract from the measured rate of exhalation of CO_2 through the skin, and thus impair the resulting *K* values. During the sensor's design, gas tightness was assessed by sticking the sensor onto a glass plate using the same adhesive as for the human-testing part of the study. The so-obtained glass plate/sensor pair was then put inside a chamber which was successively filled with a 2.5% CO_2/di-nitrogen (N_2) mixture and fresh air. The resulting measurements are presented in Figure 8B, and clearly demonstrate that the greased O-ring alone was not gas tight, while the epoxy sealing was. The Figure 8A also illustrates in

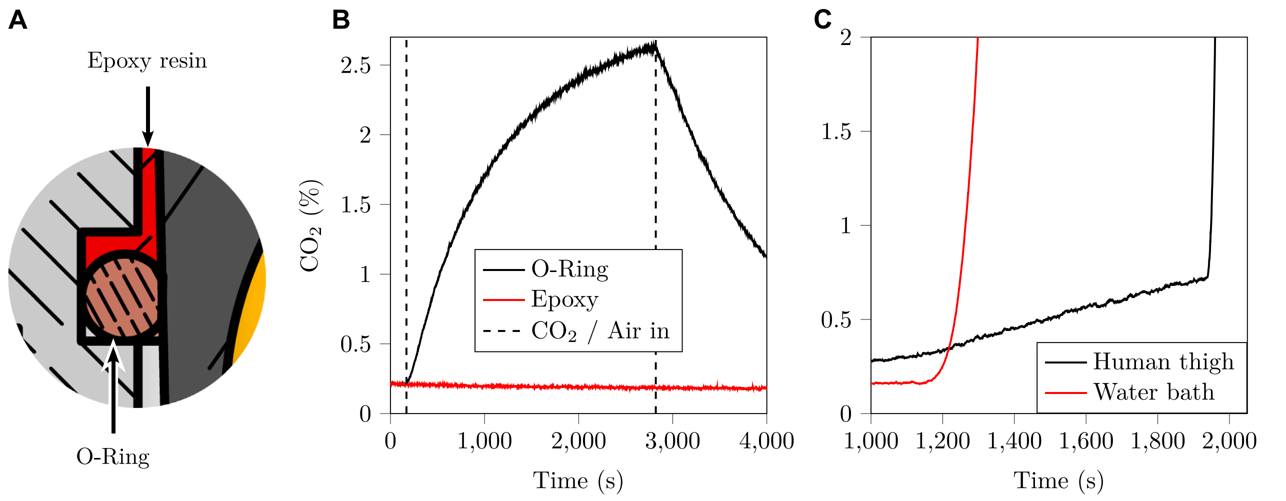


FIGURE 8
(A): close up of the section view of Figure 2, showing the two key elements of the gas tight seal: a grease-coated O-ring, and cast epoxy resin. **(B)**: a sealing test comparing the greased O-ring alone, with the greased O-ring and the cast epoxy resin. This test clearly indicates that the greased O-ring alone does not accomplish gas tightness, whereas the cast epoxy resin does. **(C)**: sensor sensitivity towards humidity, showing the onset of condensation onto the gold-plated dome. This condensation drastically reduces the quantity of light reaching the detector, effectively blinding it, which is interpreted as an exceedingly elevated CO_2 concentration.

a cut view how the epoxy seal complements the O-ring. In practice, the grease-coated O-ring acts as a resin-proof sealing that prevents the resin from flowing inside the sensor's chamber during its casting process.

This gas tightness allows CO_2 to accumulate into the sensor chamber until an equilibrium is reached between the subcutaneous tissue and the sensor's chamber—*i.e.* until $p_{Se}CO_2 = tcpCO_2$. While this equilibrium—although it would take several hours given the respective order of magnitudes of Q and h_{Se}—and the associated CO_2 diffusion process are at the very heart of this study, another undesirable chemical species will also accumulate into the sensor's chamber due to the combined action of diffusion and sweating: water vapour.

While the influence of water vapour on NDIR CO_2 measurements due to the infrared absorbance of water vapour is expected to be negligible given the large gap between CO_2 and water vapour infrared absorption bands (Mranvick, 2023), the onset of condensation onto the reflective part of the sensor—namely the gold-coated reflective dome—can still be an issue. Indeed, the formation of micro-droplets of condensing water onto the latter dome would drastically reduce its reflectance, fooling the sensor into believing that a large amount of CO_2 is present inside the sensor's chamber—a well-known issue in NDIR sensing (Fietzek et al., 2014; Wang et al., 2018). In order to study the influence of condensing humidity levels onto the sensor used in the present research, two experiments were carried out whose outcomes are presented in Figure 8C. The first experiment consisted in placing the sensor on a human thigh at increasing temperatures and waiting for condensation to occur, which happened after 30 min at 44°C. The second experiment consisted in bubbling ambient air (20°C) through pumice stone inside a hot water bath, yielding water-saturated hot air (40°C), which was then flowed onto the un-heated sensor. Even in these unfavourable conditions—*i.e.* a cold sensor and water-saturated hot air—it took about 20 min to detect the onset of condensation on CO_2 measurements. Given that the latter onset was particularly sudden and visible on $p_{Se}CO_2$ in both experiments, the influence of water vapour condensation on this study was deemed negligible. Indeed, it would be easily detected—were it to happen while measuring a given subject—and the corresponding measurement would be discarded, something which did not happen in practice.

Finally, the reader should bear in mind that the gas tightness of the sensor and the accumulation of humidity underneath it both create a condition called *skin occlusion*. This occlusion, while out of the scope of this paper, has been studied by several authors (Frame et al., 1972; King et al., 1978; Faergemann et al., 1983), who reported much higher CO_2 exhalation rates for long-term—*i.e.* days—occluded skin, as compared to its basal state. This phenomenon was not investigated in the present study due to the long time scale that it involves, but further research on this topic would be welcome.

5.1.4 Sensors positioning

To our knowledge, only three studies compared the influence of the measurement site on the transcutaneous CO_2 diffusion rate in humans: that of Schulze (1943) on twelve subjects (Table 16 *op. cit.*), that of Adamczyk et al. (1966) on one subject and that of Levshankov et al. (1983) on an unspecified number of subjects. The results of the latter two authors are summarised in Figure 9—Schulze indications were difficult to interpret and were thus not illustrated.

The high variability in Adamczyk *et al.* data—probably caused by the inclusion of only one subject—is glaring, especially when studying left-body/right-body differences. Interesting are the extremely important values reported for the axilla. Those values may be measurement artefacts, or they may be caused by a peculiar behaviour towards CO_2 diffusion of the apocrine glands, which are mainly located in the axilla—see Baker (2019). However, we found no evidence in the literature for or against this hypothesis.

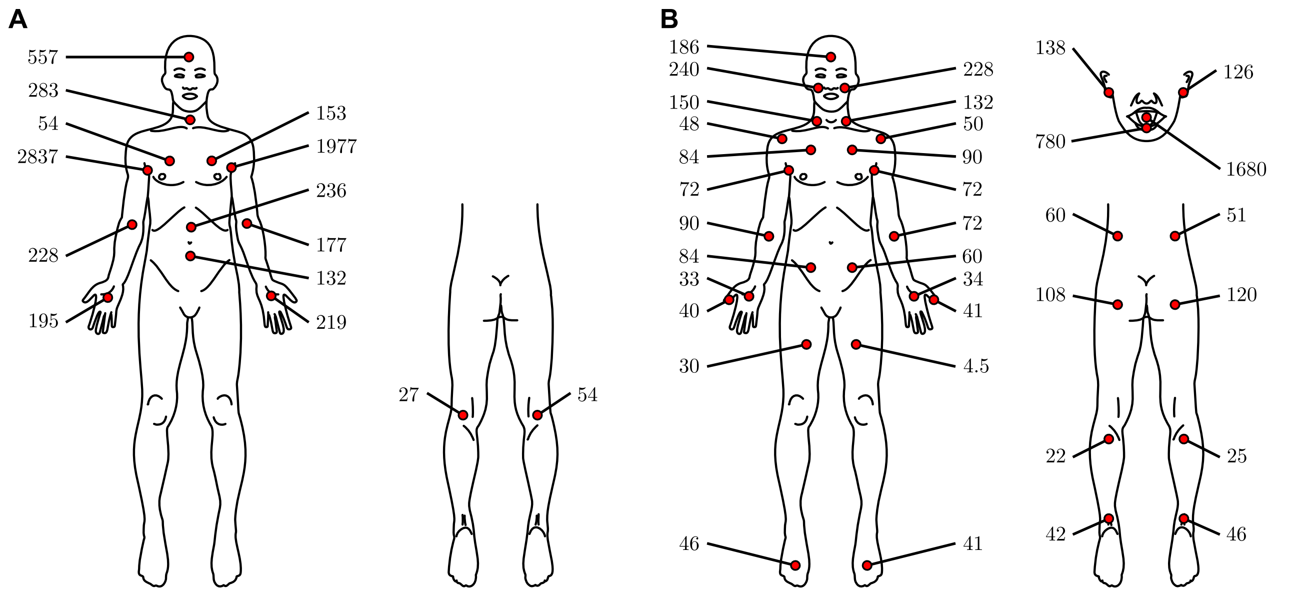


FIGURE 9
CO_2 diffusion rates through human skin at various sites, expressed in $cm^3 \cdot m^{-2} \cdot h^{-1}$, based on data from Adamczyk et al. (1966) **(A)** and Levshankov et al. (1983) **(B)**.

Alternatively, those elevated values may be caused by the skin temperature, which is much higher at the axilla on resting subjects than at the extremities (Niu et al., 2001; Sund-Levander et al., 2002), since the skin was not heated in their study. At the opposite, the results of Levshenkov *et al.* are more homogenous concerning left and right body measurements. All in all, and apart from the extreme axilla values, the reported CO_2 diffusion rates exhibit no extreme variations and are of the same order of magnitude, regardless of the measurement site. In this aspect, it thus seems from the limited information at our disposal that no measurement site is far better than the other from the CO_2 diffusion rate perspective.

Consequently, we chose our measurement sites mainly for their ease of access and acceptability, with a view to using these sites for a future wearable tcpCO_2 sensor. In this respect, the dorsal side of the wrist and the upper arm were found to be particularly interesting, as evidenced by the rapidly expanding and widespread use of health-related wristband and armband in the recent years (Al-Eidan et al., 2018; Cosoli et al., 2020; Soon et al., 2020).

5.2 Skin CO_2 conductivity

5.2.1 Metric choice

It must be emphasised here that in the simplified skin diffusion model introduced in our previous publication (Dervieux et al., 2022) and detailed in Figure 3, the membrane called "skin" does *not* correspond to an actual physiological membrane. Consequently, its thickness *w* and diffusivity towards CO_2 does *not* correspond to any physical property that might have been measured on a specific part of the dermis or epidermis. Rather, this "skin" membrane corresponds to a physical modelling of gas transport between the subcutaneous tissue and the outer air. As such, the latter membrane models both the diffusion of CO_2 through the stratum corneum *and* the circulation of blood and diffusion of CO_2 in the dermis and subcutaneous tissues.

Moreover, this model also integrates the difference in tcpCO_2 between that measured by the reference Radiometer tcpCO_2 monitor, and that measured at the sensor's location. Indeed, since the reference tcpCO_2 monitor was set to 41°C, it is likely that the tcpCO_2 that we injected in Equation 4 is slightly over-estimated—as per the dilution principle presented in Figure 11—at temperatures below 41°C. Consequently, reported *K*—or *Q*—values below 41°C are likely to be slightly over-estimated. The amplitude of this over-estimation should be in the same order of magnitude as the arterio-venous pCO_2 gradient in resting, healthy subjects—*i.e.* about 5–15% in the NH–38°C skin temperature range (Kowalchuk et al., 1988; Schneider et al., 2013). However, this state of fact was inevitable since, to the best of our knowledge, no clinical tcpCO_2 monitor working at a temperature below 37°C exists at the time being, and manufacturers recommend using 41–42°C—an injunction that we followed. Future research aiming at extending our work may consider the design of a tcpCO_2 sensor working at low temperature in order to establish the appropriate corrections to the obtained *K* values.

5.2.2 Impact on the response time of a future tcpCO_2 sensor

Contrary to perfusion—which increases over elevenfold with skin heating—*K* only doubles from NH to 44°C, and its increase is even smaller between 35–38 and 44°C values. This latter fact is all the more interesting when having in mind the design of a future energy-efficient tcpCO_2 sensor. Indeed, internal studies measuring skin temperature under a wearable device positioned at the upper arm (Bora Band, Biosency, France) on ten healthy subjects revealed that a mean skin temperature of 33.9°C could easily be achieved at the upper arm without additional heating, and that covering the arm with an additional layer of isolation—*i.e.* shirt or jumper sleeves—makes it rise even higher to reach 35.1°C.

With such skin temperatures, there is no strong incentive—from a response time point of view—to heat the skin actively any further—*i.e.* by mean of an external electrical heating system. Indeed, the measured increase of 35% in K at the arm from 35 to 44°C—see Figure 6—would result in a decrease in response time of the same magnitude for a given tcpCO$_2$ sensor, according to the response time model presented in our previous paper (Dervieux et al., 2022). While having a slower sensor may seem like a burning issue for critical care applications, it is not the case for telemonitoring for which long-term tendencies are to be observed over several months (Jang et al., 2021).

Additionally, it should be noted that since the Q values measured in the present study—80–178 cm^3·m^{-2}·h^{-1} on average—are in line with that used in our previous publication (Dervieux et al., 2022) for response time calculations—100 cm^3·m^{-2}·h^{-1}—the afore-proposed sensor thickness of 100 μm for a response time below 10 min remains credible. As a reminder, it was shown in the latter publication that a linear relationship exists between the response time of an equilibrium-based tcpCO$_2$ sensor, and the volume to surface ratio—*i.e.* equivalent thickness—of its equilibration medium. Thus, there is essentially a compromise to be made between this thickness, which cannot be infinitely small for technological reasons, and the response time of a so-designed sensor (Dervieux et al., 2022).

Of note, and to the best of our knowledge, there is a lack of clinical guidelines specifying the required response time for tcpCO$_2$ monitors. Nevertheless, there exists a considerable amount of literature focusing on transcutaneous monitor testing in clinical environments, from which it can be inferred that a typical *in vitro* 90% response time of about 1 min (Bendjelid et al., 2005; Eberhard, 2007) is achievable with current tcpCO$_2$ monitors. *In vivo* performance reports, on their, part mention an approximatively 10 min initial equilibration time before a first tcpCO$_2$ reading can be taken (Carter and Banham, 2000; Domingo et al., 2010; Restrepo et al., 2012). Regarding the response time of tcpCO$_2$ monitors following a sudden change in paCO$_2$, a lag has been reported in the literature between end-tidal pCO$_2$—petCO$_2$—paCO$_2$, and tcpCO$_2$, inducing a higher *in vivo* response time than in the ideal *in vitro* case. Reported values for this latter lag fall within the 1–5 min range (Kesten et al., 1991; Carter and Banham, 2000; Cuvelier et al., 2005; Rafl et al., 2018). An overall response time requirement of approximatively 5 min can thus *de facto* be assumed for a tcpCO$_2$ monitor to meet field expectations. Still, this latter assumption mainly holds for the intensive care of critically ill patients (Mari et al., 2019) and no information exists concerning long-term tcpCO$_2$ (tele-)monitoring for the obvious reason that the corresponding monitors do not exist yet.

5.3 Exhalation Rate

5.3.1 An Imperfect Metric

Considering Equation 1, it readily appears that the exhalation rate Q is not constant, and logically depends on the initial $p_{Se}CO_2$, and of the passing of time. This issue has however been largely ignored by the literature on the topic—see Table 3—and Q has been considered by most authors as if it had a single constant value. The latter, which has been reported as *the* CO$_2$ diffusion rate through the skin, is actually the *initial one in free air*—*i.e.* Q (t=0) with $p_{Se}CO_2(t = 0) \approx 0$—and corresponds to the slopes of the tangents to the $p_{Se}CO_2$ curves at $t = 0$ in Figure 10A.

Unfortunately, if measuring Q as illustrated in the latter figure is theoretically feasible at different temperatures, it would also require to remove the sensor at each temperature change, in order to renew the gas inside the inner chamber of the sensor with fresh air. This would in turn require to peel off the sensor from the subject's skin at each temperature change, which would distort the Q measurement, as the skin—and more specifically the *stratum corneum*, its outermost layer—would become thinner and thinner at each sensor replacement—actually, stripping the skin with multiple tape applications is a well-known technique to drastically increase Q (Scheuplein, 1976; Eletr et al., 1978; Greenspan et al., 1981).

Thus, the sensor was left in place in this study, while the temperature was successively changed from NH up to 35, 38, 41, and finally 44°C. This led to measured $p_{Se}CO_2$ alike that represented in Figure 10B. In that case, using Q as a metric would be unpractical, since the $p_{Se}CO_2$ value at t_H is not null, and Q values would no longer represent *initial* CO$_2$ diffusion rates as measured in free air. In practice, the obtained Q values at different temperatures would then not be comparable with each other, each one being measured with a slightly different $p_{Se}CO_2$ initial value.

5.3.2 Comparison with existing literature

Thanks to Equation 4, we can however obtain equivalent Q values in free air from our K measurements, and compare them with those of the literature, given in Table 3—at least for the NH case. The values reported in Tables 2 and 3 are of the same magnitude, and the wide amplitudes that we report here—*e.g.* 25–192 and 24–141 cm^3·m^{-2}·h^{-1} at the NH arm and wrist, respectively—are on par with those reported in previous research.

5.4 Laser doppler flowmetry

5.4.1 Choosing nSkBF$_{90}$ as a metric

Both inter-subject and inter-site LDF variabilities have often been reported in the literature (Johnson et al., 1984; Cracowski et al., 2006; Minson, 2010; Roustit and Cracowski, 2012; Cracowski and Roustit, 2016; Hodges et al., 2016), and appears to be inherent to this modality of skin blood flow measurement as well as to human physiology in general. Nonetheless, certain guidelines may be followed to obtain the most reproducible results (Cracowski et al., 2006). In particular, when it comes to derive a single explicit LDF metric from a given measurement period—such as a skin-site/sensor-temperature pair, for instance—several techniques have been developed to obtain meaningful results from raw LDF data.

At first, some authors—*e.g.* Hodges et al. (2016)—prefer to express the skin blood flow as Cutaneous Vascular Conductance (CVC), which is given by the LDF in P.U. or V, divided by the Mean Arterial Pressure (MAP). The CVC is said to be more "physiological" (Cracowski et al., 2006), since an increase in skin blood flow could be caused by an increase in MAP but also by an increase in vascular compliance, for instance. By dividing the LDF-acquired blood flow

TABLE 3 CO_2 exhalation rates through the skin in the literature compared with the present studies.

Exhalation rate *Q* [$cm^3 \cdot m^{-2} \cdot h^{-1}$]	Temp. [°C]	Num. Of Subjects	Reference
25–120	22–36 (air)	2	Shaw and Messer (1930a)
10–160	25–37 (air)	1	Shaw and Messer (1930b)
58–169	**26–31 (air)**	**38**	**Ernstene and Volk (1932), Table 1**
12–143	23–37 (air)	1	Whitehouse et al. (1932)
32–69	**25–28 (air)**	**13**	**Schulze (1943), Table 12**
180–2500†	–	1	Adamczyk et al. (1966)
25–87	–	5	Thiele and Van Kempen (1972)
11–28	25–35 (air)	3	Frame et al. (1972)
50	**–**	**27**	**Levshankov et al. (1983)**
140–221	**36 (skin)**	**14**	**Eöry (1984)**
25–192	27–32 (skin)	40	This work, NH arm
24–141	25–33 (skin)	40	This work, NH wrist
25–288	27–44 (skin)	40	This work, all temp. arm
24–300	25–44 (skin)	40	This work, all temp. wrist

Past studies with more than ten participants are indicated **in bold** and were used for sample size determination.

†Axilla measured value, possibly erroneous.

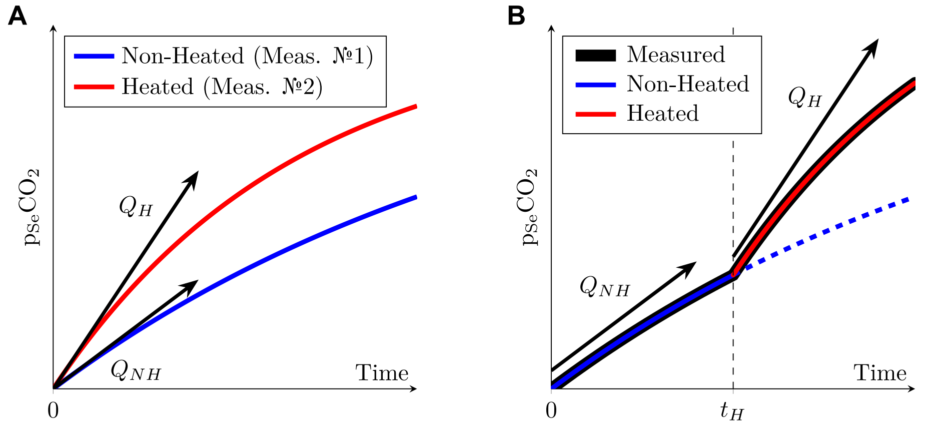


FIGURE 10
(A): schematic evolution of the pCO_2 inside the sensor's chamber when the skin is heated or not. **(B)**: same as Left, but with a non-heated sensor placed onto the skin for a duration t_H before being heated. Note the difference between the two sensing schemes: the left one requires two successive measurements—one heated, the other one non-heated—while the right one consists in a single measurement during which the skin is successively non-heated and then heated. The dashed line on the right represents what would have happened without heating during the whole acquisition, which is equivalent to the solid blue line on the left.

by the MAP, the obtained CVC value is thus in theory more representative of the arteriovenous compliance, a theory supported by several works in haemodynamics (Johnson, 1986; Lautt, 1989; Herring and Paterson, 2018). At the same time, skin blood flow alone—often abbreviated as SkBF, and either expressed in P.U. or Volts—has been used for several decades (Johnson et al., 1984; Frantz et al., 2012) and remains a good alternative to CVC when MAP is not available.

Then, once the type of measurement—SkBF or CVC—is chosen, the question that comes next is that of the extraction of a single perfusion metric from a long-lasting acquisition. Indeed, due to the peculiar dynamics of thermal hyperaemia—see Figure 4, above—a simple time-averaging on the whole acquisition duration would make little sense.

To circumvent this issue, several research teams used temporal averaging on manually-set periods of interest in the raw LDF data. The averaging duration that they used depended on the studied phenomenon, with durations of 1–3 min for transient phenomena—*i.e.* initial bump and after-bump nadir—up to 5–10 min for long-lasting ones—*i.e.* baseline or maximum perfusion plateau (Minson et al., 2001; Frantz et al., 2012). Other authors, for their parts, chose to average the two to three last minutes of a 10–25 min measurement window at the maximal perfusion value, which obtention is detailed below (Kellogg et al., 2008; Hodges et al., 2016). However, Barcroft and Edholm (1943) and Taylor et al. (1984) mentioned even longer durations for thermal hyperaemia to fully settle following a change in skin temperature—up to 40–60 min. Such a lengthy onset period would result in a total acquisition duration in the 3–5 h range for the five different temperatures involved in the present study. At the opposite, a total experiment duration—including informing the subjects and obtaining their consent—of about 2 h seemed to us to be an acceptable maximum for easily recruiting volunteers. This 2 h duration in turn entails that each temperature window of the present study only lasted 18 min, which may not be enough for the establishment of the nitric-oxide mediated hyperaemia detailed in Figure 4. Thankfully, this 18 min duration is by far long enough for the axon mediated response to take place, and the latter often yields perfusion levels comparable to that reached at the end of the nitric-oxide mediated phase (Kellogg et al., 2008; Minson, 2010; Frantz et al., 2012). Thus, by taking the 90-th percentile of SkBF—see Section 3.1.2—the obtained SkBF$_{90}$ values are likely to be representative of the SkBF plateau values which would have been observed by increasing the duration of the temperature windows. The latter hypothesis is further confirmed by the similarity between our results and that of the literature, as discussed in the next section.

Finally, it is also common practice to normalise the measured skin blood flow—whether expressed as SkBF or CVC—by its maximum value, often taken after a prolonged (≥15 min) period at an elevated (≥44°C) temperature (Taylor et al., 1984; Vionnet et al., 2014; Hodges et al., 2016) or by direct injection of sodium nitroprusside (Kellogg et al., 2008). Although it has been seldom proposed to normalise the measured values by the baseline blood flow value instead of the maximum one (Magerl and Treede, 1996; Mayrovitz and Leedham, 2001), this is considered bad practice because intra-subject baseline variations can be important even in a temperature controlled room (Bircher et al., 1994; Cracowski et al., 2006).

In this study, CVC was not considered due to the invasiveness of a continuous MAP measurement and SkBF was thus chosen as raw perfusion metric. Then, we proposed to take the 90-th percentile of a given temperature window instead of time averaging. Finally, normalisation by the maximum perfusion value—*i.e.* the one reached at the end of the 44°C window—was performed, as per literature guidelines.

5.4.2 Comparison with the literature

The LDF measurements that were gathered in the present study are consistent with existing literature on the topic. In particular, the sigmoid behaviour observed in Figure 7—revealing a strong onset of hyperaemia in the 35–41°C range—is on par with the observations of Magerl and Treede (1996), Stephens et al. (2001), and Hodges et al. (2016).

5.4.3 Impact on the accuracy of a future tcpCO$_2$ sensor

The fact that the perfusion is doubled at 35°C and quadrupled at 38°C compared to baseline—see Figure 7—is especially encouraging for the development of a future energy-efficient tcpCO$_2$ sensor, since these temperatures can be easily achieved without—or with minimal—heating, as already discussed in Section 5.2.2 (reaching over 35°C at the upper arm under jumper sleeves). Indeed, since arterialised capillary blood—either obtained by local heating or application of a vasoactive cream—is gaseously close to arterial blood (Zavorsky et al., 2007), it is to be expected that partially arterialised capillary blood obtained by a mild heating—*i.e.* below 44°C—lies somewhere between venous and arterial blood, from a gaseous content point of view. More specifically, Rooth et al. (1987) hypothesised that the subcutaneous capillary pCO$_2$—*i.e.* tcpCO$_2$—would be a barycentre between venous and arterial pCO$_2$, as illustrated in Figure 11.

This figure emphasises the fact that—especially for a resting subject—even a mild heating of the skin in the 35–38°C range could be enough to yield a tcpCO$_2$ only a few mmHg away from the paCO$_2$. The latter error may be acceptable depending on the clinical application targeted. For example, the Food and Drug Administration (FDA) requires tcpCO$_2$ monitors to be accurate within 5 mmHg, with an allowed drift of up to 10% of the initial reading over a 1-h period (Food and Drug Administration, 2002).

5.5 Sample size

The main objective of the present study was to estimate the mean K value as a function of temperature. The latter mean can be estimated at each temperature T by:

$$\widehat{K}_T = \frac{1}{N} \cdot \sum_{i=1}^{N} K_{S,i} \tag{7}$$

wherein the index i stands for the ith subject of the study and N stands for its sample size. Contrary to hypothesis testing, for which a sample size may be derived straightforwardly from targeted alpha or beta risks, and some prior knowledge of the data (Ambrosius, 2007, Chap. 19; Chow et al., 2017), sample size determination in the case of an exploratory—or pilot—study is more challenging, with its share of arbitrary decision (Ko and Lim, 2021). Indeed, while a 95% confidence interval can be computed for K_T as:

$$C.I._{K_T}^{95\%} = \left[\widehat{K}_T - \varepsilon, \widehat{K}_T + \varepsilon\right], \quad \text{and} \quad \varepsilon = -t_{N-1}\left(\frac{0.05}{2}\right) \cdot \frac{s}{\sqrt{N}} \tag{8}$$

wherein t_{N-1} is the percentile score of a Student distribution with $N-1$ degrees of freedom, and s is the SD of the sample—*i.e.* the *estimated* standard deviation of the population—the value of the latter SD is vastly unknown. In order to estimate an adequate

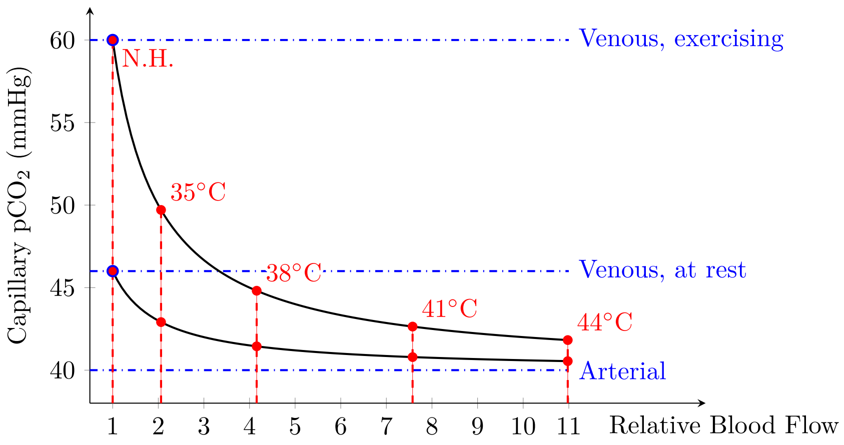


FIGURE 11
Capillary pCO$_2$ as a function of relative blood flow considering two venous pCO$_2$ levels: at rest, and while exercising. Relative blood flow values measured in the present study were also added in red with their respective temperature labels. A normal paCO$_2$ of 40 mmHg (Schneider et al., 2013) was taken for arterial blood, while venous blood levels were set to 46 and 60 mmHg at rest and while exercising, respectively. Of note, while 46 mmHg at rest is generally accepted in the literature (Byrne et al., 2014), the 60 mmHg exercising value was mainly chosen for legibility reasons. Indeed, exercising values may exceed 100 mmHg during heavy exercise, or in case of septic shock (Kowalchuk et al., 1988; Diaztagle Fernández et al., 2017). Modified from Rooth et al. (1987).

sample size for the study at hand—based on an *acceptable* margin of error on $\widehat{K}_T$—a prior estimation of s is thus needed, and is the object of the upcoming section. Importantly, since Q—that is $Q\,(t=0)$ using the above-presented notation—was studied by earlier authors instead of K, the following reasoning will be made using the former. This can be done safely since the two values are linked by a proportionality constant—see Equation 4. Of note, Equation 8 holds only if the K_T follow a normal distribution, which had neither been confirmed nor denied in the literature, to the best of our knowledge, but which was verified in the current study—see Section 5.5.3.

5.5.1 Literature review

Among the literature studies on the topic of skin CO$_2$ exhalation rate measurements on human subjects detailed in Table 3, only four of them have been performed on more than ten subjects, and are highlighted **in bold** in the aforementioned table. Unfortunately, the latter studies are sometimes unclear about Q measuring conditions—measurement site and skin temperature, in particular. We did our best not to distort or misinterpret the works of their authors, but what follows is essentially our best *interpretation* of their writings. These four studies reported—or made it possible to derive from raw data—a $\widehat{Q}$ and s value for each measurement site investigated. These data can be used to compute 95% confidence intervals on Q estimation—as shown in Equation 8—or on s, yielding (Ambrosius, 2007, Chap. 4):

$$C.I._{\sigma}^{95\%} = \left[\frac{s^2 \cdot (N-1)}{\chi^2_{N-1}\left(\frac{0.05}{2}\right)} \; ; \; \frac{s^2 \cdot (N-1)}{\chi^2_{N-1}\left(1-\frac{0.05}{2}\right)} \right] \tag{9}$$

wherein χ^2_{N-1} is the percentile score of a χ^2 distribution with $N-1$ degrees of freedom. The resulting confidence intervals are reported in Table 4 and tend to indicate a relative uncertainty on Q estimation in the order of 5–30% for relatively small sample sizes—*i.e.* 13–38 subjects.

5.5.2 Chosen sample size

The reported s value, as well as the upper and lower bounds of s 95% confidence interval were then used to compute the relative uncertainty on Q as a function of the number of subjects using Equation 8. While this relative uncertainty decreases when the sample size increases, its reducing rate—as well as the associated uncertainty values—varies wildly depending on the considered data source. Indeed, while a sample size of 20–30 subjects should lead to a relative uncertainty on Q in the 5–10% range using Levshankov et al. (1983) data, much larger sample sizes—*i.e.* 100-150 subjects—would be needed to reach the same level of accuracy using Schulze (1943) or Ernstene and Volk (1932). measurements. In the end, since the works of the latter two authors were much older—1932 and 1943, respectively—than that of Eöry (1984), Levshankov et al. (1983) and , respectively—it was decided to put them aside. The sample size determination was thus grounded only on the works of Eöry (1984), Levshankov et al. (1983), and a sample size of 40 subjects was deemed acceptable, since it should have resulted into a relative uncertainty on Q estimation below 10%.

5.5.3 Results

Unfortunately, this initial estimation of a 40 subjects cohort proved to be rather optimistic in practice. Indeed, the relative uncertainty on measured Q values can be computed using Equation 8, and falls in the 15–32% range, depending on the skin temperature and measurement site. In this aspect, our results are close to those presented by Ernstene and Volk (1932), Schulze (1943), and Eöry (1984) who reported relative uncertainties in the 11–30% range. Yet, the latter authors used less than 40 subjects, and our uncertainty range was thus expected to be narrower than theirs.

TABLE 4 Confidence intervals at the 95% level for σ and Q, computed from s and $\widehat{Q}$ values reported in the literature. Aside from the relative uncertainties—defined as $2 \cdot \varepsilon / X$ wherein X is σ or Q—all values are given in $cm^3 \cdot m^{-2} \cdot h^{-1}$.

	Lower bound	Reported	Upper bound	Relative uncertainty (%)	Measurement site	Reference
σ	18.7	22.9	29.6	48	Whole arm	Ernstene and Volk (1932)
	8.5	11.9	19.6	93	Abdomen	Schulze (1943)
	2.65	3.36	4.60	58	Left forearm	Levshankov et al. (1983)
	2.41	3.06	4.19		Right forearm	Levshankov et al. (1983)
	15	21	34	89	Acupuncture site	Eöry (1984)
	13	18	28		“adjacent skin area”	Eöry (1984)
Q	113	120	128	13	Whole arm	Ernstene and Volk (1932)
	41	48	55	30	Abdomen	Schulze (1943)
	48.0	49.3	50.7	5.39	Left forearm	Levshankov et al. (1983)
	48.4	49.6	50.8	4.88	Right forearm	Levshankov et al. (1983)
	209	221	233	11	Acupuncture site	Eöry (1984)
	130	140	150	14	“adjacent skin area”	Eöry (1984)

Moreover, the present study also exhibits a higher variability than that of Levshankov et al. (1983)—whose results indicate a relative uncertainty about 5%, see Table 4.

The origin of these discrepancies between literature-driven expectations and the above-presented results is not fully understood at the moment. One possible explanation could be differing measurement sites between the above-mentioned studies and the ones that we chose. Indeed, the data reported by Eöry (1984), Schulze (1943) (Table 16 *op. cit.*) tend to indicate some degree of variability in the relative uncertainty expected at different sites—in particular when comparing the abdomen and hand in Schulze's data (30 vs. 45%), or the two sites used by Eöry (10 vs. 14%). Thus, it seems plausible that Q variability at the upper arm and wrist is above that reported in earlier studies for differing sites.

Ultimately, we cannot but recommend using larger sample sizes in future studies of a similar nature, considering the significant variability we observed. As a side note related to sample size determination, the normality of the Q distribution was ascertained using a series of Bonferroni-corrected Shapiro-Wilk tests which were non-significant, further justifying the approach presented in Section 5.5. To the best of our knowledge, this is the first report of normality for transcutaneous CO_2 exhalation rates.

6 Conclusion

As stated in introduction, the aim of this study was twofold: measuring the influence of skin temperature on the transcutaneous diffusion of CO_2, and on the skin blood flow. To this end, a custom sensor was designed and used on 40 healthy human subjects at two measurements sites: the upper arm, and the wrist.

Our results indicate comparable behaviours at both sites, with an increasing relationship between temperature on the one hand, and CO_2 exhalation rate, CO_2 conductivity and perfusion on the other hand. These results are encouraging for the development of a future energy-efficient tcpCO_2 sensor for the following reasons:

- Skin conductivity towards CO_2 increases only moderately with an increase in skin temperature, at most doubling from NH to 44°C. Thus, if the response time of the sensor-to-be is not critical—*i.e.* if a 35% slower response is acceptable compared to the one reachable at maximum skin heating—the latter may not require additional heating. This is especially encouraging in the perspective of building a wearable, battery-operated device.
- Perfusion, for its part, increases strongly with an increase in skin temperature, already doubling from NH to 35°C, and quadrupling from NH to 38°C. This phenomenon is especially interesting since—according to Rooth et al. (1987)—this should bring tcpCO_2 close to paCO_2 even for skin temperatures as low as 35–38°C, which are reachable at the arm without additional heating, given that the latter is covered by warm clothings. However, this latter hypothesis—*i.e.* the existence of a clinically-satisfying tcpCO_2/paCO_2 correlation in the 35–38°C skin temperature range—is yet to be demonstrated experimentally *in vivo*, which will be the subject of future research.

Additionally, our results highlight the significant variability of transcutaneous CO_2 exhalation rate and conductivity measurements in human subjects. Hence, we strongly advise future research on the topic to consider large sample sizes—*i.e.* more than 40 subjects—in order to ensure accurate estimates of the latter metrics. The present study also focuses only on two measurement sites—the upper arm and the wrist—and further investigations at other sites would be welcome. In particular, the remarkably high axilla values reported by some authors is intriguing, and could benefit from a special attention. Of note, the study data—demographics, K, Q, and nSkBF_{90} values—are provided in Supplementary Material S2.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by the Comité de protection des personnes Sud-Méditerrannée II, IDRCB ref.: 2020-A02185-38. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

ED: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Visualization, Writing–original draft, Writing–review and editing. FG: Supervision, Validation, Writing–review and editing. WU: Supervision, Validation, Writing–review and editing. M-AG-M: Funding acquisition, Investigation, Supervision, Validation, Writing–review and editing, Project administration. MT: Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Validation, Writing–review and editing, Project administration.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by Biosency and the Centre Hospitalier Régional et Universitaire de Brest.

Acknowledgments

The authors wish to warmly thank Hervé Abitbol and the Tecmatel Engineering team for their design of the adhesive mounting jig, Julien Dolou for his help in the administrative formalities involved in conducting the clinical study, and Alexandre Desvaux, Mikaël Jamet (Cimtech) and Lucien Lecurieux Lafayette (Troisel) for their help in the design and manufacturing of the aluminium sensor body. The authors also acknowledge the following designers/entities for their drawings: Whyakup (computer) and Rflor (power supply) from the Noun Project, Mobius (USB logo) and Mariana Ruiz Villarreal (venous system) from Wikipedia, Pixabay (male anatomy) and the NASA (left arm drawing extracted from Pioneer plaque).

Conflict of interest

ED works for Biosency, a company which develops wearable cardiopulmonary assessment devices. Although this paper concerns fundamental research in physiology and even if the presented sensor is mainly a research tool—and not the prototype of a commercial sensor—this disclosure may help to avoid the appearance of a conflict of interest.

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2023.1293752/full#supplementary-material

References

Adamczyk, B., Boerboom, A. J., and Kistemaker, J. (1966). A mass spectrometer for continuous analysis of gaseous compounds excreted by human skin. *J. Appl. Physiology* 21, 1903–1906. doi:10.1152/jappl.1966.21.6.1903

Al-Eidan, R. M., Al-Khalifa, H., and Al-Salman, A. M. (2018). A review of wrist-worn wearable: sensors, models, and challenges. *J. Sensors* 2018, 1–20. doi:10.1155/2018/5853917

Ambrosius, W. T. (2007). *Topics in biostatistics. methods in molecular biology*. Humana Totowa, NJ: Humana Press. Available at: https://link.springer.com/book/10.1007/978-1-59745-530-5.

Baker, L. B. (2019). Physiology of sweat gland function: the roles of sweating and sweat composition in human health. *Temperature* 6, 211–259. doi:10.1080/23328940.2019.1632145

Barcroft, H., and Edholm, O. G. (1943). The effect of temperature on blood flow and deep temperature in the human forearm. *J. Physiology* 102, 5–20. doi:10.1113/jphysiol.1943.sp004009

Bendjelid, K., Schütz, N., Stotz, M., Gerard, I., Suter, P., and Romand, J.-A. (2005). Transcutaneous pCO_2 monitoring in critically ill adults: clinical evaluation of a new sensor. *Crit. care Med.* 33, 2203–2206. doi:10.1097/01.CCM.0000181734.26070.26

Bircher, A., de Boer, E. M., Agner, T., Wahlberg, J. E., and Serup, J. (1994). Guidelines for measurement of cutaneous blood flow by laser Doppler flowmetry. A report from the Standardization Group of the European Society of Contact Dermatitis. *Contact Dermat.* 30, 65–72. doi:10.1111/j.1600-0536.1994.tb00565.x

Bonner, R. F., and Nossal, R. (1990). *Principles of laser-Doppler Flowmetry*. Boston, MA: Springer US, 17–45. doi:10.1007/978-1-4757-2083-9_2

Byrne, A. L., Bennett, M., Chatterji, R., Symons, R., Pace, N. L., and Thomas, P. S. (2014). Peripheral venous and arterial blood gas analysis in adults: are they comparable? a systematic review and meta-analysis. *Respirology* 19, 168–175. doi:10.1111/resp.12225

Carter, R., and Banham, S. (2000). Use of transcutaneous oxygen and carbon dioxide tensions for assessing indices of gas exchange during exercise testing. *Respir. Med.* 94, 350–355. doi:10.1053/rmed.1999.0714

Cascales, J. P., Li, X., Roussakis, E., and Evans, C. L. (2022). A patient-ready wearable transcutaneous CO_2 sensor. *Biosensors* 12, 333. doi:10.3390/bios12050333

Chow, S., Shao, J., Wang, H., and Lokhnygina, Y. (2017). *Sample size calculations in clinical research*. London, United Kingdom: Chapman and Hall/CRC Biostatistics Series (Taylor and Francis).

Chung, M., Fortunato, G., and Radacsi, N. (2019). Wearable flexible sweat sensors for healthcare monitoring: a review. *J. R. Soc. Interface* 16, 20190217. doi:10.1098/rsif.2019.0217

Conway, A., Tipton, E., Liu, W.-H., Conway, Z., Soalheira, K., Sutherland, J., et al. (2018). Accuracy and precision of transcutaneous carbon dioxide monitoring: a systematic review and meta-analysis. *Thorax* 74, 157–163. doi:10.1136/thoraxjnl-2017-211466

Cosoli, G., Spinsante, S., and Scalise, L. (2020). Wrist-worn and chest-strap wearable devices: systematic review on accuracy and metrological characteristics. *Measurement* 159, 107789. doi:10.1016/j.measurement.2020.107789

Cracowski, J.-L., Minson, C. T., Salvat-Melis, M., and Halliwill, J. R. (2006). Methodological issues in the assessment of skin microvascular endothelial function in humans. *Trends Pharmacol. Sci.* 27, 503–508. doi:10.1016/j.tips.2006.07.008

Cracowski, J.-L., and Roustit, M. (2016). Current methods to assess human cutaneous blood flow: an updated focus on laser-based-techniques. *Microcirculation* 23, 337–344. doi:10.1111/micc.12257

Cuvelier, A., Grigoriu, B., Molano, L. C., and Muir, J.-F. (2005). Limitations of transcutaneous carbon dioxide measurements for assessing long-term mechanical ventilation. *Chest* 127, 1744–1748. doi:10.1378/chest.127.5.1744

Dagher, L., Shi, H., Zhao, Y., and Marrouche, N. F. (2020). Wearables in cardiology: here to stay. *Heart rhythm.* 17, 889–895. doi:10.1016/j.hrthm.2020.02.023

Del Pozzi, A. T., Miller, J. T., and Hodges, G. J. (2016). The effect of heating rate on the cutaneous vasomotion responses of forearm and leg skin in humans. *Microvasc. Res.* 105, 77–84. doi:10.1016/j.mvr.2016.01.004

Dervieux, E., Théron, M., and Uhring, W. (2022). Carbon dioxide sensing—biomedical applications to human subjects. *Sensors* 22, 188. doi:10.3390/s22010188

Diaztagle Fernández, J. J., Rodríguez Murcia, J. C., and Sprockel Díaz, J. J. (2017). Venous-to-arterial carbon dioxide difference in the resuscitation of patients with severe sepsis and septic shock: a systematic review. *Med. Intensiva* 41, 401–410. doi:10.1016/j.medin.2017.03.008

Domingo, C., Canturri, E., Moreno, A., Espuelas, H., Vigil, L., and Luján, M. (2010). Optimal clinical time for reliable measurement of transcutaneous co2 with ear probes: counterbalancing overshoot and the vasodilatation effect. *Sensors* 10, 491–500. doi:10.3390/s100100491

Dunn, J., Runge, R., and Snyder, M. (2018). Wearables and the medical revolution. *Pers. Med.* 15, 429–448. doi:10.2217/pme-2018-0044

Eberhard, P. (2007). The design, use, and results of transcutaneous carbon dioxide analysis: current and future directions. *Anesth. Analgesia* 105, S48–S52. doi:10.1213/01.ane.0000278642.16117.f8

Eletr, S., Jimison, H., Ream, A. K., Dolan, W. M., and Rosenthal, M. H. (1978). Cutaneous monitoring of systemic pCO_2 on patients in the respiratory intensive care unit being weaned from the ventilator. *Acta Anaesthesiol. Scand.* 22, 123–127. doi:10.1111/j.1399-6576.1978.tb01406.x

Eöry, A. (1984). *In-vivo* skin respiration (CO_2) measurements in the acupuncture loci. *Acupunct. Electro-Therapeutics Res.* 9, 217–223. doi:10.3727/036012984816714668

Ernstene, A. C., and Volk, M. C. (1932). Cutaneous respiration in man: iv. the rate of carbon dioxide elimination and oxygen absorption in normal subjects. *J. Clin. Investigation* 11, 363–376. doi:10.1172/JCI100418

Faergemann, J., Aly, R., Wilson, D. R., and Maibach, H. I. (1983). Skin occlusion: effect on pityrosporum orbiculare, skin pCO_2, pH, transepidermal water loss, and water content. *Archives Dermatological Res.* 275, 383–387. doi:10.1007/BF00417338

Fietzek, P., Fiedler, B., Steinhoff, T., and Körtzinger, A. (2014). *In situ* quality assessment of a novel underwater pCO_2 sensor based on membrane equilibration and ndir spectrometry. *J. Atmos. Ocean. Technol.* 31, 181–196. doi:10.1175/JTECH-D-13-00083.1

Fitzgerald, L. R. (1957). Cutaneous respiration in man. *Physiol. Rev.* 37, 325–336. doi:10.1152/physrev.1957.37.3.325

Food and Drug Administration (2002). *Cutaneous carbon dioxide (PcCO2) and oxygen (PcO2) monitors - class II special controls guidance document for industry and FDA*. U.S. Department Of Healsth and Human Services. Tech. rep.

Frame, G. W., Strauss, W. G., and Maibach, H. I. (1972). Carbon dioxide emission of the human arm and hand. *J. Investigative Dermatology* 59, 155–159. doi:10.1111/1523-1747.ep12625939

Frantz, J., Engelberger, R. P., Liaudet, L., Mazzolai, L., Waeber, B., and François, F. (2012). Desensitization of thermal hyperemia in the skin is reproducible. *Microcirculation* 19, 78–85. doi:10.1111/j.1549-8719.2011.00124.x

Garfan, S., Alamoodi, A. H., Zaidan, B. B., Al-Zobbi, M., Hamid, R. A., Alwan, J. K., et al. (2021). Telehealth utilization during the Covid-19 pandemic: a systematic review. *Comput. Biol. Med.* 138, 104878. doi:10.1016/j.compbiomed.2021.104878

Greenspan, G., Block, A., Haldeman, L., Lindsey, S., and Martin, C. (1981). Transcutaneous noninvasive monitoring of carbon dioxide tension. *Chest* 80, 442–446. doi:10.1378/chest.80.4.442

Hansen, T. N., Sonoda, Y., and McIlroy, M. B. (1980). Transfer of oxygen, nitrogen, and carbon dioxide through normal adult human skin. *J. Appl. Physiology* 49, 438–443. doi:10.1152/jappl.1980.49.3.438

Held, M., Tweer, S., Medved, F., Rothenberger, J., Daigeler, A., and Petersen, W. (2018). Changes in the biomechanical properties of human skin in hyperthermic and hypothermic ranges. *Wounds* 30, 257–262.

Herring, N., and Paterson, D. J. (2018). *Haemodynamics: flow, pressure and resistance*. CRC Press. chap. 8. 121–147. doi:10.1201/9781351107754

Hodges, G. J., McGarr, G. W., Mallette, M. M., Del Pozzi, A. T., and Cheung, S. S. (2016). The contribution of sensory nerves to the onset threshold for cutaneous vasodilatation during gradual local skin heating of the forearm and leg. *Microvasc. Res.* 105, 1–6. doi:10.1016/j.mvr.2015.12.004

Hodgkinson, J., and Tatam, R. P. (2012). Optical gas sensing: a review. *Meas. Sci. Technol.* 24, 012004. doi:10.1088/0957-0233/24/1/012004

Huttmann, S. E., Windisch, W., and Storre, J. H. (2014). Techniques for the measurement and monitoring of carbon dioxide in the blood. *Ann. Am. Thorac. Soc.* 11, 645–652. doi:10.1513/AnnalsATS.201311-387FR

Jang, S., Kim, Y., and Cho, W.-K. (2021). A systematic review and meta-analysis of telemonitoring interventions on severe COPD exacerbations. *Int. J. Environ. Res.* 18 (13), 6757. doi:10.3390/ijerph18136757

Johnson, J. M., Taylor, W. F., Shepherd, A. P., and Park, M. K. (1984). Laser-Doppler measurement of skin blood flow: comparison with plethysmography. *J. Appl. Physiology* 56, 798–803. doi:10.1152/jappl.1984.56.3.798

Johnson, P. C. (1986). Autoregulation of blood flow. *Circulation Res.* 59, 483–495. doi:10.1161/01.RES.59.5.483

Kellogg, D. L., Zhao, J. L., and Wu, Y. (2008). Endothelial nitric oxide synthase control mechanisms in the cutaneous vasculature of humans *in vivo*. *Am. J. Physiology-Heart Circulatory Physiology* 295, H123–H129. doi:10.1152/ajpheart.00082.2008

Kesten, S., Chapman, K. R., and Rebuck, A. S. (1991). Response characteristics of a dual transcutaneous oxygen/carbon dioxide monitoring system. *Chest* 99, 1211–1215. doi:10.1378/chest.99.5.1211

King, R., Rl, C., Maibach, H. I., Jh, G., Ml, W., and Jc, J. (1978). The effect of occlusion on carbon dioxide emission from human skin. *Acta dermato-venereologica* 58 (2), 135–138. doi:10.2340/0001555558135138

Ko, M. J., and Lim, C.-Y. (2021). General considerations for sample size estimation in animal study. *Korean J. Anesthesiol.* 74, 23–29. doi:10.4097/kja.20662

Koch, G. (1965). Comparison of carbon dioxide tension, pH and standard bicarbonate in capillary blood and in arterial blood with special respect to relations in patients with impaired cardiovascular and pulmonary function and during exercise. *Scand. J. Clin. Laboratory Investigation* 17, 223–229. doi:10.1080/00365516509075339

Kowalchuk, J. M., Heigenhauser, G. J., Lindinger, M. I., Sutton, J. R., and Jones, N. L. (1988). Factors influencing hydrogen ion concentration in muscle after intense exercise. *J. Appl. Physiology* 65, 2080–2089. doi:10.1152/jappl.1988.65.5.2080

Kruse, C., Pesek, B., Anderson, M., Brennan, K., and Comfort, H. (2019). Telemonitoring to manage chronic obstructive pulmonary disease: systematic literature review. *JMIR Med. Inf.* 7, e11496. doi:10.2196/11496

Lautt, W. (1989). Resistance or conductance for expression of arterial vascular tone. *Microvasc. Res.* 37, 230–236. doi:10.1016/0026-2862(89)90040-x

Lermuzeaux, M., Meric, H., Sauneuf, B., Girard, S., Normand, H., Lofaso, F., et al. (2016). Superiority of transcutaneous CO_2 over end-tidal CO_2 measurement for monitoring respiratory failure in nonintubated patients: a pilot study. *J. Crit. Care* 31, 150–156. doi:10.1016/j.jcrc.2015.09.014

Levshankov, A. I., Pushkina, M. A., Slutskaia, M. E., and Uvarov, B. S. (1983). Determination of local gas exchange on the body surface by the method of mass spectrometry. *Meditsinskaia tekhnika* 1, 21–26. doi:10.1007/BF00560505

Magerl, W., and Treede, R. D. (1996). Heat-evoked vasodilatation in human hairy skin: axon reflexes due to low-level activity of nociceptive afferents. *J. Physiology* 497, 837–848. doi:10.1113/jphysiol.1996.sp021814

Mari, A., Nougue, H., Mateo, J., Vallet, B., and Vallée, F. (2019). Transcutaneous pCO_2 monitoring in critically ill patients: update and perspectives. *J. Thorac. Dis.* 11, S1558–S1567. doi:10.21037/jtd.2019.04.64

Mayrovitz, H. N., and Leedham, J. A. (2001). Laser–Doppler imaging of forearm skin: perfusion features and dependence of the biological zero on heat-induced hyperemia. *Microvasc. Res.* 62, 74–78. doi:10.1006/mvre.2001.2314

McIlroy, M. B., Simbruner, G., and Sonoda, Y. (1978). Transcutaneous blood gas measurements using a mass spectrometer. *Acta Anaesthesiol. Scand.* 22, 128–130. doi:10.1111/j.1399-6576.1978.tb01407.x

Minson, C. T. (2010). Thermal provocation to evaluate microvascular reactivity in human skin. *J. Appl. Physiology* 109, 1239–1246. doi:10.1152/japplphysiol.00414.2010

Minson, C. T., Berry, L. T., and Joyner, M. J. (2001). Nitric oxide and neurally mediated regulation of skin blood flow during local heating. *J. Appl. Physiology* 91, 1619–1626. doi:10.1152/jappl.2001.91.4.1619

Mranvick, (2023). What is the influence of humidity on NDIR CO_2 measurements? Physics Stack Exchange. Available at: https://physics.stackexchange.com/q/788799 (version: 2023-11-22.

Nanji, A. A., and Whitlow, K. J. (1984). Is it necessary to transport arterial blood samples on ice for pH and gas analysis? *Can. Anaesth. Soc. J.* 31, 568–571. doi:10.1007/BF03009545

Niu, H. H., Lui, P. W., Hu, J. S., Ting, C. K., Yin, Y. C., Lo, Y. L., et al. (2001). Thermal symmetry of skin temperature: normative data of normal subjects in Taiwan. *Chin. Med. J.* 64, 459–468.

Radiometer (2020). TCM CombiM, continuous blood gas monitoring. Available at: https://pdfhost.io/v/DaeEjvgEq_radiometer_combim.

Rafl, J., Kulhanek, F., Kudrna, P., Ort, V., and Roubik, K. (2018). Response time of indirectly accessed gas exchange depends on measurement method. *Biomed. Eng.* 63, 647–655. doi:10.1515/bmt-2017-0070

Restrepo, R. D., Hirst, K. R., Wittnebel, L., and Wettstein, R. (2012). AARC clinical practice guideline: transcutaneous monitoring of carbon dioxide and oxygen: 2012. *Respir. Care* 57, 1955–1962. doi:10.4187/respcare.02011

Rithalia, S. V. S., Clutton-Brock, T. H., and Tinker, J. (1984). Characteristics of transcutaneous carbon dioxide tension monitors in normal adults and critically ill patients. *Intensive Care Med.* 10, 149–153. doi:10.1007/BF00265805

Rooth, G., Ewald, U. W., and Caligara, F. (1987). Transcutaneous pO_2 and pCO_2 monitoring at 37°C. cutaneous pO_2 and pCO_2. *Adv. Exp. Med. Biol.* 220, 23–32.

Roustit, M., and Cracowski, J.-L. (2012). Non-invasive assessment of skin microvascular function in humans: an insight into methods. *Microcirculation* 19, 47–64. doi:10.1111/j.1549-8719.2011.00129.x

Salter, D., McArthur, H. C., Crosse, J., and Dickens, A. (1993). Skin mechanics measured *in vivo* using torsion: a new and accurate model more sensitive to age, sex and moisturizing treatment. *Int. J. Cosmet. Sci.* 15, 200–218. doi:10.1111/j.1467-2494.1993.tb00075.x

Scheer, B. V., Perel, A., and Pfeiffer, U. J. (2002). Clinical review: complications and risk factors of peripheral arterial catheters used for haemodynamic monitoring in anaesthesia and intensive care medicine. *Crit. Care* 6, 199–204. doi:10.1186/cc1489

Scheuplein, R. J. (1976). Permeability of the skin: a review of major concepts and some new developments. *J. Investigative Dermatology* 67, 672–676. doi:10.1111/1523-1747.ep12544513

Schneider, A. G., Eastwood, G. M., Bellomo, R., Bailey, M., Lipcsey, M., Pilcher, D., et al. (2013). Arterial carbon dioxide tension and outcome in patients admitted to the intensive care unit after cardiac arrest. *Resuscitation* 84, 927–934. doi:10.1016/j.resuscitation.2013.02.014

Schulze, W. (1943). Untersuchungen über die Alkaliempfindlichkeit, das Alkalineutralisationsvermögen und die Kohlensäureabgabe der Haut. *Arch. für Dermatol. Syph.* 185, 93–161. doi:10.1007/BF02714173

SenTec (2016). *Sentec v-sign brochure, rf-007857-bm*.

Severinghaus, J. W., and Astrup, P. B. (1986). History of blood gas analysis. III. Carbon dioxide tension. *J. Clin. Monit.* 2, 60–73. doi:10.1007/BF01619178

Shaw, L. A., and Messer, A. C. (1930b). Cutaneous respiration in man: II. The effect of temperature and of relative humidity upon the rate of carbon dioxide elimination and oxygen absorption. *Am. J. Physiology-Legacy Content* 95, 13–19. doi:10.1152/ajplegacy.1930.95.1.13

Shaw, L. A., Messer, A. C., and Weiss, S. (1930a). Cutaneous respiration in man: I. factors affecting the rate of carbon dioxide elimination and oxygen absorption. *Am. J. Physiology-Legacy Content* 95, 107–118. doi:10.1152/ajplegacy.1929.90.1.107

Soon, S., Svavarsdottir, H., Downey, C., and Jayne, D. G. (2020). Wearable devices for remote vital signs monitoring in the outpatient setting: an overview of the field. *BMJ Innov.* 6, 55–71. doi:10.1136/bmjinnov-2019-000354

Stephens, D. P., Charkoudian, N., Benevento, J. M., Johnson, J. M., and Saumet, J. L. (2001). The influence of topical capsaicin on the local thermal control of skin blood flow in humans. *Am. J. Physiology-Regulatory, Integr. Comp. Physiology* 281, R894–R901. doi:10.1152/ajpregu.2001.281.3.R894

Steventon, A., Bardsley, M., Billings, J., Dixon, J., Doll, H., Hirani, S., et al. (2012). Effect of telehealth on use of secondary care and mortality: findings from the whole system demonstrator cluster randomised trial. *BMJ* 344, e3874. doi:10.1136/bmj.e3874

Sund-Levander, M., Forsberg, C., and Wahren, L. K. (2002). Normal oral, rectal, tympanic and axillary body temperature in adult men and women: a systematic literature review. *Scand. J. Caring Sci.* 16, 122–128. doi:10.1046/j.1471-6712.2002.00069.x

Taylor, W. F., Johnson, J. M., O'Leary, D., and Park, M. K. (1984). Effect of high local temperature on reflex cutaneous vasodilation. *J. Appl. Physiology* 57, 191–196. doi:10.1152/jappl.1984.57.1.191

Thiele, F. A. J., and Van Kempen, L. H. J. (1972). A micro method for measuring the carbon dioxide release by small skin areas. *Br. J. Dermatology* 86, 463–471. doi:10.1111/j.1365-2133.1972.tb16098.x

Tufan, T. B., and Guler, U. (2022). "A miniaturized transcutaneous carbon dioxide monitor based on dual lifetime referencing," in 2022 IEEE Biomedical Circuits and Systems Conference (BioCAS), 144–148. doi:10.1109/BioCAS54905.2022.9948600

Vionnet, J., Calero-Romero, I., Heim, A., Rotaru, C., Engelberger, R. P., Dischl, B., et al. (2014). No major impact of skin aging on the response of skin blood flow to a submaximal local thermal stimulus. *Microcirculation* 21, 730–737. doi:10.1111/micc.12154

Wagner, P. D. (2015). The physiological basis of pulmonary gas exchange: implications for clinical interpretation of arterial blood gases. *Eur. Respir. J.* 45, 227–243. doi:10.1183/09031936.00039214

Wang, J. N., Xue, Q. S., Lin, G. Y., and Ma, Q. J. (2018). Mid-infrared carbon dioxide sensor with wireless and anti-condensation capability for use in greenhouses. *Spectrosc. Lett.* 51, 266–273. doi:10.1080/00387010.2018.1468785

Whitehouse, A. G. R., Hancock, W., and Haldane, J. S. (1932). The osmotic passage of water and gases through the human skin. *Proc. R. Soc. Lond. Ser. B, Contain. Pap. A Biol. Character* 111, 412–429. doi:10.1098/rspb.1932.0065

Wimberley, P. D., Grønlund Pedersen, K., Olsson, J., and Siggaard-Andersen, O. (1985). Transcutaneous carbon dioxide and oxygen tension measured at different temperatures in healthy adults. *Clin. Chem.* 31, 1611–1615. doi:10.1093/clinchem/31.10.1611

Yetisen, A. K., Martinez-Hurtado, J. L., Ünal, B., Khademhosseini, A., and Butt, H. (2018). Wearables in medicine. *Adv. Mater.* 30, 1706910. doi:10.1002/adma.201706910

Yun, J. E., Park, J.-E., Park, H.-Y., Lee, H.-Y., and Park, D.-A. (2018). Comparative effectiveness of telemonitoring versus usual care for heart failure: a systematic review and meta-analysis. *J. Cardiac Fail.* 24, 19–28. doi:10.1016/j.cardfail.2017.09.006

Zavorsky, G. S., Cao, J., Mayo, N. E., Gabbay, R., and Murias, J. M. (2007). Arterial versus capillary blood gases: a meta-analysis. *Respir. Physiology Neurobiol.* 155, 268–279. doi:10.1016/j.resp.2006.07.002

Frontiers in Physiology

TYPE Original Research
PUBLISHED 15 February 2024
DOI 10.3389/fphys.2024.1322881

OPEN ACCESS

EDITED BY
Emiliano Cè,
University of Milan, Italy

REVIEWED BY
Clayton L. Camic,
Northern Illinois University, United States
Takafumi Ando,
National Institute of Advanced Industrial Science and Technology (AIST), Japan

*CORRESPONDENCE
Naomi Omi,
✉ omi.naomi.gn@u.tsukuba.ac.jp

RECEIVED 17 October 2023
ACCEPTED 23 January 2024
PUBLISHED 15 February 2024

CITATION
Ogata H, Negishi Y, Koizumi N, Nagayama H, Kaneko M, Kiyono K and Omi N (2024), Individually optimized estimation of energy expenditure in rescue workers using a tri-axial accelerometer and heart rate monitor.
Front. Physiol. 15:1322881.
doi: 10.3389/fphys.2024.1322881

Individually optimized estimation of energy expenditure in rescue workers using a tri-axial accelerometer and heart rate monitor

Hitomi Ogata[1], Yutaro Negishi[2], Nao Koizumi[2], Hisashi Nagayama[2], Miki Kaneko[3], Ken Kiyono[3] and Naomi Omi[2]*

[1]Graduate School of Humanities and Social Sciences, Hiroshima University, Hiroshima, Japan, [2]Faculty of Health and Sport Sciences, University of Tsukuba, Tsukuba, Japan, [3]Graduate School of Engineering Science, Osaka University, Toyonaka, Japan

Objectives: This study aimed to provide an improved energy expenditure estimation for heavy-load physical labor using accelerometer data and heart rate (HR) measured by wearables and to support food preparation and supply management for disaster relief and rescue operations as an expedition team.

Methods: To achieve an individually optimized estimation for energy expenditure, a model equation parameter was determined based on the measurements of physical activity and HR during simulated rescue operations. The metabolic equivalent of task (MET), which was measured by using a tri-axial accelerometer and individual HR, was used, where two (minimum and maximum) or three (minimum, intermediate, and maximum) representative reference points were selected for each individual model fitting. In demonstrating the applicability of our approach in a realistic situation, accelerometer-based METs and HR of 30 males were measured using the tri-axial accelerometer and wearable HR during simulated rescue operations over 2 days.

Results: Data sets of 27 rescue operations (age:34.2 ± 7.5 years; body mass index (BMI):22.9 ± 1.5 kg/m^2) were used for the energy expenditure estimation after excluding three rescue workers due to their activity type and insufficient HR measurement. Using the combined approach with a tri-axial accelerometer and HR, the total energy expenditure increased by 143% for two points and 133% for three points, compared with the estimated total energy expenditure using only the accelerometer-based method.

Conclusion: The use of wearables provided a reasonable estimation of energy expenditure for physical workers with heavy equipment. The application of our approach to disaster relief and rescue operations can provide important insights into nutrition and healthcare management.

KEYWORDS

rescue operations, metabolic equivalent of task, disaster simulation training, estimation equation, wearable

1 Introduction

An accurate estimation of energy expenditure can provide important information and guidelines for nutrition and healthcare management of physical workers. In particular, disaster relief and rescue teams must prepare and transport their food supplies in advance. The higher the degree of energy deficiency, the lower the energy expenditure in activity, especially during high-intensity physical activity, and the more significant the decrease in energy expenditure (Martin et al., 2011). Based on the studies conducted by the U.S. military, the lack of energy promotes muscle damage and muscle soreness, decreases the performance of physical activities (Margolis et al., 2014), causes weight loss and a greater percentage of muscle mass loss than fat mass (Berryman et al., 2017), decreased immune function (T lymphocyte response) (Kramer et al., 1997), and decreased cardiac function (altered left ventricular diastolic function) (Planer et al., 2012). Therefore, the total amount of food supplied must be sufficient to meet the nutritional needs of the body and to maintain the energy required for high-intensity physical activity. However, energy expenditure estimation methods for such workers have not been well established, and the energy expenditure during such operations remains unclear.

With recent advancements in sensor technologies, portable devices are becoming smaller, capable of longer data collection (because of their storage and battery lives), multidimensional, and more sensitive (Chen et al., 2012). Many movement sensors can be used to measure human physical activities, including electromechanical switches (for heel strike detection), mercury switches, pedometers, inclinometers, gyroscopes, goniometers (for angles or postures), accelerometers, and even global positioning systems (GPS) (Chen et al., 2012). Among these devices, accelerometers are currently the most widely used sensors for human physical activity monitoring in clinical and free-living settings (Chen et al., 2012) because of their small size, noninvasiveness, and relatively low cost (Plasqui and Westerterp, 2007). Moreover, accelerometers can easily estimate the amount of energy expenditure during high-intensity physical activities, except for heavy equipment and static load activities (Bouten et al., 1997). However, discriminating movements during daily low-intensity physical activities and estimating the appropriate energy expenditure remain a challenge (Ndahimana and Kim, 2017). Some tri-axial accelerometers have shown potential application in estimating the amount of energy expenditure during low-intensity physical activities, such as sitting, standing, housework, and walking, which are important for estimating total energy expenditure (Midorikawa et al., 2007; Yamada et al., 2009; Ohkawara et al., 2011). When measuring the total energy expenditure in field validation studies (double-labeled water (DLW), which is the gold standard for measuring total energy expenditure in the field), high correlations with the total energy expenditure estimate were found in most activity monitors (Van Remoortel et al., 2012). However, these correlations are to a large extent driven by subject characteristics, body weight, age, and height, which are important predictors of total energy expenditure (Plasqui et al., 2005). Based on previous reports, only 19% of the total energy expenditure is accounted for by physical activity in healthy subjects (Plasqui et al., 2005) and in patients with coronary heart disease (Ades et al., 2005) in a field setting. Previous research demonstrated that wearable devices, such as movement sensors, do not accurately represent energy expenditure measured by the DLW (Murakami et al., 2019; Willis et al., 2022). Several studies have also compared the DLW method to the total energy expenditure estimated by using accelerometers. The total energy expenditure using an accelerometer in patients with reduced pulmonary function, particularly chronic obstructive pulmonary disease, is underestimated by 11% compared with the DLW (Sato et al., 2021). Even the total energy expenditure of people who are not engaged in heavy physical activity is underestimated compared with that of accelerometer. A wrist-mounted motion sensor accounts for 78% of the variation of total energy expenditure using the DLW during the free-living period and 62% during the training period (Kinnunen et al., 2019). The total energy expenditure using an accelerometer in firefighters under normal working conditions is underestimated by 33% compared with the DLW (Touno et al., 2003). In the literature, the total energy expenditure by the accelerometer is underestimated compared with the DLW (Touno et al., 2003; Kinnunen et al., 2019; Murakami et al., 2019; Sato et al., 2021; Willis et al., 2022) despite the difference between wrist-mounted and waist/chest-worn accelerometers (Kinnunen et al., 2019). Therefore, the energy expenditure estimation for heavy equipment and static load activities has an underestimation bias, although accelerometer-based methods are convenient and applicable to real-world situations (Touno et al., 2003).

Various wearable devices to monitor heart rate (HR) have been developed, such as wristwatches, ear clips, and undershirts, and some devices can measure HR accurately. These devices can record HR noninvasively, with minimal technical effort and without the constraints of laboratory conditions (Karmen et al., 2019; Isakadze and Martin, 2020). The HR increases almost proportionally to exercise intensity and intra-individual oxygen uptake (% $\dot{V}O_2$ max), and its dynamics can influence a number of different factors, such as body temperature, food intake, body posture, and individual cardio-respiratory fitness level (Hill and Trowbridge, 1998). However, the use of HR wearable devices, most of which are wrist-worn devices, may underestimate energy expenditure and provide inaccurate measures of energy expenditure compared with reference standard criterion measures, including direct calorimetry and indirect calorimetry (Fuller et al., 2020; Chevance et al., 2022). Summarizing the total energy expenditure by most of the aforementioned devices, such as accelerometer, HR, GPS, and combined motion sensors, provides a more accurate estimation of energy expenditure at light-to-moderate intensities; by contrast, underestimation increases at very light and higher intensity activities (Aparicio-Ugarriza et al., 2015).

A previous review reported that adding indicators such as HR and heat flux values to acceleration values as a method of estimating daily energy expenditure and physical activity has not significantly improved the system (Van Remoortel et al., 2012). Nevertheless, focusing on the HR that can be measured with a wearable device can address the underestimation of energy expenditure by using only a tri-axial accelerometer because HR can relate intensity to energy expenditure. This study aimed to provide an energy expenditure estimation method using wearables to support food preparation and supply management, particularly for expedition rescue team workers with heavy equipment and static loads. By combining the characteristics of tri-axial accelerometers and HR monitor information, we evaluated the relationship between accelerometer-based physical activity and HR and provided an individually optimized energy expenditure estimation equation (model). Providing an equation/model to predict energy expenditure would promote the estimation of energy expenditure in the field, which is not measurable in the laboratory. As an application of this approach, the energy expenditure of rescue operations was estimated while mimicking large-scale disasters.

2 Materials and methods

2.1 Subjects

This study included 30 males (age: 34.2 ± 7.3 years; body mass index (BMI): 23.1 ± 1.7 kg/m^2) who participated in a 2-day disaster simulation training (i.e., rescue operations were unaware of the training protocol). Note that some of the firefighters of our previously reported paper were included in the present study (Koizumi et al., 2021). This study was approved by the local ethics committee of the University of Tsukuba (approval number: tai-28-66), conducted in accordance with the principles set forth in the Declaration of Helsinki, and the subjects were informed of and agreed to the details of this study. In addition, detailed explanations were given to each training director and participating organization-affiliated institution regarding the purpose and content of the experiment, and the experiment was started only after obtaining agreement.

2.2 Simulated rescue operations

Rescue operation training was conducted from November 30 to 1 December 2018, in Kanagawa Prefecture, Japan. The mean temperature of the 2-day training was 12.8°C (8.5°C–18.3°C), with humidity of 55.5%. The main activities were as follows: rescue activity training in a skyscraper, in a gas leak accident, in a collapsed building, at a sediment-related disaster site, in the event of many injured people on the subway, from vehicles, and in a tunnel collapse accident; fire-extinguishing activity training in an industrial complex fire.

2.3 Measurements

2.3.1 Accelerometer

Figure 1 shows flow chart of calculating energy expenditure. A tri-axial accelerometer (Active style Pro HJA-750C, 23 g, 40 mm × 52 mm × 12 mm; Omron Healthcare Co., Ltd., Kyoto, Japan) (Ohkawara et al., 2011; Nagayoshi et al., 2019; Nishida et al., 2020) was inserted into the chest pocket, located near the waist area (approximately 10 cm above the belt). The accelerometer was covered with a vinyl material and fixed in the chest pocket with adhesive tape to avoid the influence of liquid matter (e.g., water) in fire-extinguishing activities. The basic metabolic rate was calculated automatically when weight, height, age, and sex were entered, which was calculate using Ganpule's formula (Ganpile et al., 2007). The accelerometer was programmed to save the metabolic equivalent of task (MET) once every 10 s using a built-in Omron's algorithm (Oshima et al., 2010; Ohkawara et al., 2011). MET is defined as the ratio of the metabolic rate during an activity to the metabolic rate at rest (Jetté et al., 1990). Next, each transformed data point (METs) was converted into units of activity energy expenditure in kcal/min (Oshima et al., 2010; Ohkawara et al., 2011) to understand the amount of energy intake, and the data were divided and aggregated for each activity (see Section 2.3.2 Recording paper).

2.3.2 Recording paper

A recording paper was distributed, and rescue squads were asked to describe the activities that they performed during the disaster simulation training. This recording paper was used to classify the types and times of activities.

2.3.3 HR monitor

R–R intervals were measured by electrocardiogram signals using wearable HR sensors (WHS-1, myBeat, 14 g, 39 mm × 37 mm × 9 mm; Union Tool Co., Tokyo, Japan) (Arikawa et al., 2020). The HR sensor was placed on a shirt with attached conductive fiber electrodes (Kurabo Industries Ltd., Osaka, Japan, Figure 2). The HR per minute [beats per minute (bpm)] was estimated using the median of instantaneous HRs during a 1-min period to reduce the effect of the outliers presented in R–R interval data. If the absolute difference of the successive R–R intervals was larger than 200 ms, then the detected R–R interval was classified as an outlier using R version 3.6.0 (R Foundation for Statistical Computing, Vienna, Austria; http://www.R-project.org/). The percentage of the outlier R–R interval for each individual was also calculated.

2.3.4 Estimation equation (model)

A percent of HR reserve (%HRR) provides a good approximation (almost linear relation) of $\dot{V}O_2$ reserve (%$\dot{V}O_2$) was shown in the previous studies (Swain et al., 1998; Strath et al., 2000), and an HR-based equation can provide a reliable estimation of the METs (Swain et al., 1998; Strath et al., 2000). %HRR is defined as follows:

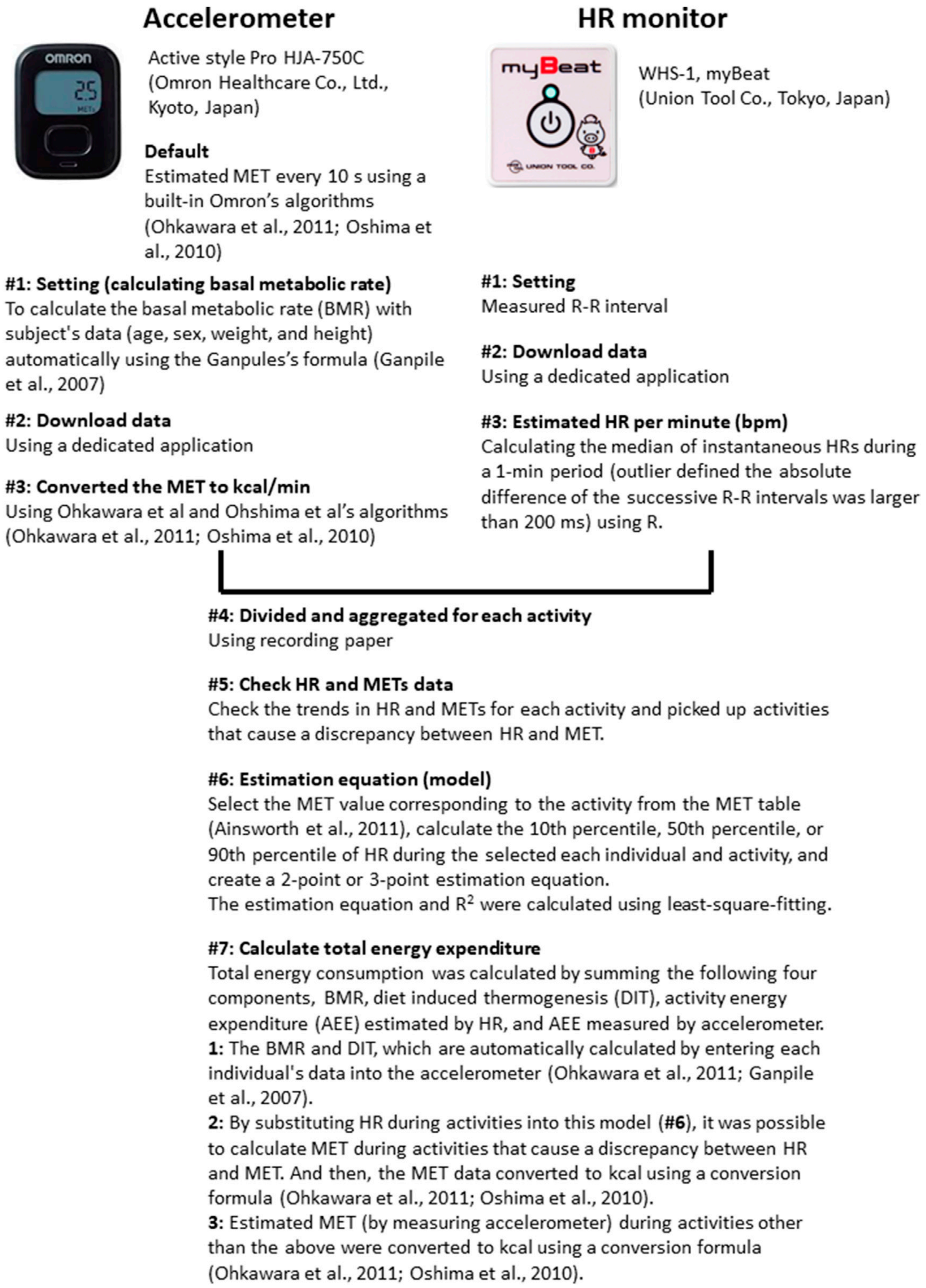
$$\%HRR = \frac{HR_{activity} - HR_{rest}}{HR_{max} - HR_{rest}} \times 100$$

where $HR_{activity}$ is the recorded HR during the activity; HR_{rest} is the HR while sitting at rest, and HRmax is the maximum HR as estimated by the Karvonen equation, (220 bpm—age) bpm. %$\dot{V}O_2$ reserve is defined as follows:

$$\%\dot{V}O_2R = \frac{\dot{V}O_{2act} - \dot{V}O_{2rest}}{\dot{V}O_{2\,max} - \dot{V}O_{2rest}} \times 100,$$

where $\dot{V}O_{2activity}$ is the recorded $\dot{V}O_2$ during the activity; $\dot{V}O_{2rest}$ is the $\dot{V}O_2$ while sitting at rest, and $\dot{V}O_{2\,max}$ is the maximum $\dot{V}O_2$ as estimated by the equation proposed by Jackson et al. (1990). By assuming %HRR ≈ %$\dot{V}O_2R$ and dividing the numerator and denominator of the %$\dot{V}O_2R$ by the $\dot{V}O_{2rest}$, we obtain the following equation:

$$\frac{HR_{act} - HR_{rest}}{HR_{max} - HR_{rest}} = \frac{\frac{\dot{V}O_{2act}}{\dot{V}O_{2rest}} - 1}{\frac{\dot{V}O_{2\,max}}{\dot{V}O_{2rest}} - 1} = \frac{METS_{act} - 1}{METS_{max} - 1}$$



Accelerometer

Active style Pro HJA-750C
(Omron Healthcare Co., Ltd.,
Kyoto, Japan)

Default
Estimated MET every 10 s using a built-in Omron's algorithms (Ohkawara et al., 2011; Oshima et al., 2010)

#1: Setting (calculating basal metabolic rate)
To calculate the basal metabolic rate (BMR) with subject's data (age, sex, weight, and height) automatically using the Ganpules's formula (Ganpile et al., 2007)

#2: Download data
Using a dedicated application

#3: Converted the MET to kcal/min
Using Ohkawara et al and Ohshima et al's algorithms (Ohkawara et al., 2011; Oshima et al., 2010)

HR monitor

WHS-1, myBeat
(Union Tool Co., Tokyo, Japan)

#1: Setting
Measured R-R interval

#2: Download data
Using a dedicated application

#3: Estimated HR per minute (bpm)
Calculating the median of instantaneous HRs during a 1-min period (outlier defined the absolute difference of the successive R-R intervals was larger than 200 ms) using R.

#4: Divided and aggregated for each activity
Using recording paper

#5: Check HR and METs data
Check the trends in HR and METs for each activity and picked up activities that cause a discrepancy between HR and MET.

#6: Estimation equation (model)
Select the MET value corresponding to the activity from the MET table (Ainsworth et al., 2011), calculate the 10th percentile, 50th percentile, or 90th percentile of HR during the selected each individual and activity, and create a 2-point or 3-point estimation equation.
The estimation equation and R^2 were calculated using least-square-fitting.

#7: Calculate total energy expenditure
Total energy consumption was calculated by summing the following four components, BMR, diet induced thermogenesis (DIT), activity energy expenditure (AEE) estimated by HR, and AEE measured by accelerometer.
1: The BMR and DIT, which are automatically calculated by entering each individual's data into the accelerometer (Ohkawara et al., 2011; Ganpile et al., 2007).
2: By substituting HR during activities into this model (**#6**), it was possible to calculate MET during activities that cause a discrepancy between HR and MET. And then, the MET data converted to kcal using a conversion formula (Ohkawara et al., 2011; Oshima et al., 2010).
3: Estimated MET (by measuring accelerometer) during activities other than the above were converted to kcal using a conversion formula (Ohkawara et al., 2011; Oshima et al., 2010).

FIGURE 1
Flow chart of calculating energy expenditure.

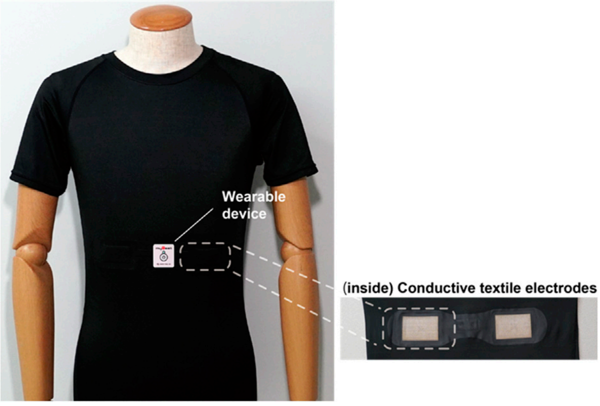


FIGURE 2
T-shirt attached with conductive fiber electrodes (backside). Stretchy shirts with electrodes were placed in the psoas. The size of the T-shirt was selected on the basis of the physique of the rescue operations.

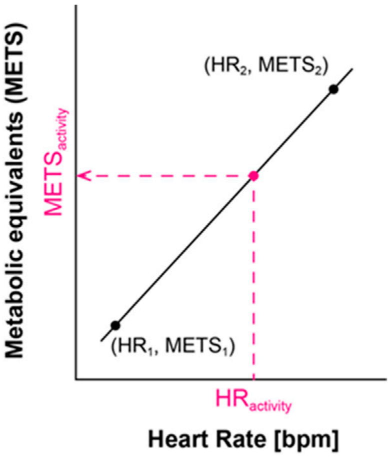


FIGURE 3
Estimation equation between HR and METs.

Based on the abovementioned relation, we obtain the following equation for MET estimation (Strath et al., 2000):

$$\mathrm{METS_{act}} = \frac{\mathrm{METS_{max}} - 1}{\mathrm{HR_{max}} - \mathrm{HR_{rest}}}\left(\mathrm{HR_{act}} - \mathrm{HR_{rest}}\right) + 1$$

Therefore, if we use two reference conditions with different METs, METS$_1$ and METS$_2$, instead of resting and the maximum $\dot{V}O_2$ condition, then we can obtain a more general equation for MET estimation.

$$\mathrm{METS_{act}} = \frac{\mathrm{MET_2} - \mathrm{MET_1}}{\mathrm{HR_2} - \mathrm{HR_1}}\left(\mathrm{HR_{act}} - \mathrm{HR_1}\right) + \mathrm{METS_1},\ (*)$$

where HR$_1$ and HR$_2$ are the HR recorded during the activity with METS$_1$ and METS$_2$, respectively.

We used two (minimum and maximum) or three (minimum, intermediate, and maximum) representative reference points recorded in each individual (Figure 3) to obtain the parameters, $(\mathrm{HR_1}, \mathrm{MET_1})$, $(\mathrm{HR_2}, \mathrm{MET_2})$, in the MET estimation equation [Eq (*)] optimized to each individual. The lowest HR, HR$_1$, corresponding MET$_1$ was calculated as follows: 1) the 10th percentile of HR during napping was calculated as HR$_1$, and 2) we defined the MET$_1$ as 0.95 MET based on the MET table (Garby et al., 1987). We selected the 10th percentile of HR instead of the observed lowest value to attenuate the effect of outliers of the HR measurement. The intermediate HR was calculated as follows: First, the 50th percentile of HR during withdrawal was calculated as the intermediate HR because withdrawal was performed by all groups and was considered to be less affected by tension as it was performed after all activities were completed, and then we defined the intermediate HR as 3.0 MET based on the MET table (Ainsworth et al., 2011). Moreover, the highest HR corresponding to the highest MET during all activities was calculated as follows: First, the 90th percentile of HR was calculated for each rescue operation activity. Although the highest HR would be induced by the activity with the highest MET, we selected the 90th percentile of HR instead of the observed highest value to attenuate the effect of outliers of the HR measurement. Second, activity's METs were defined on the basis of the MET table (Table 1) (Ainsworth et al., 2011), and then the activity with the highest HR was selected. Finally, the parameters, $(\mathrm{HR_2}, \mathrm{MET_2})$, were estimated for all participants using a least-square-fitting on the abovementioned values.

TABLE 1 Activities corresponding to the MET table.

Code	Major heading	Activity[a] (description)	METs
05147	home activities	withdrawal (implied walking, putting away household items, moderate effort)	3.0
11240	occupation	rescue narrow space, rescue activity, rescue search (fire fighter, general)	8.0
11244	occupation	rescue traffic, rescue transport (fire fighter, rescue victim, automobile accident, using pike pole)	6.8
11245	occupation	fire training (fire fighter, raising and climbing ladder with full gear, simulated fire suppression)	8.0
11246	occupation	gas rescue, rescue gas search, rescue gas transport (fire fighter, hauling hoses on ground, carrying/hoisting equipment, breaking down walls, wearing full gear)	9.0
11550	occupation	rescue sediment, rescue tunnel (shoveling, more than 7.3 kg/min, deep digging, vigorous effort)	8.8
17029	walking	rescue activity training in a skyscraper (carrying 22.7–33.6 kg load, upstairs)	10.0

[a]Activity represents the rescue operations conducted in this study. The code, major heading, description, and METs, are provided in the MET, table (Ainsworth et al., 2011).

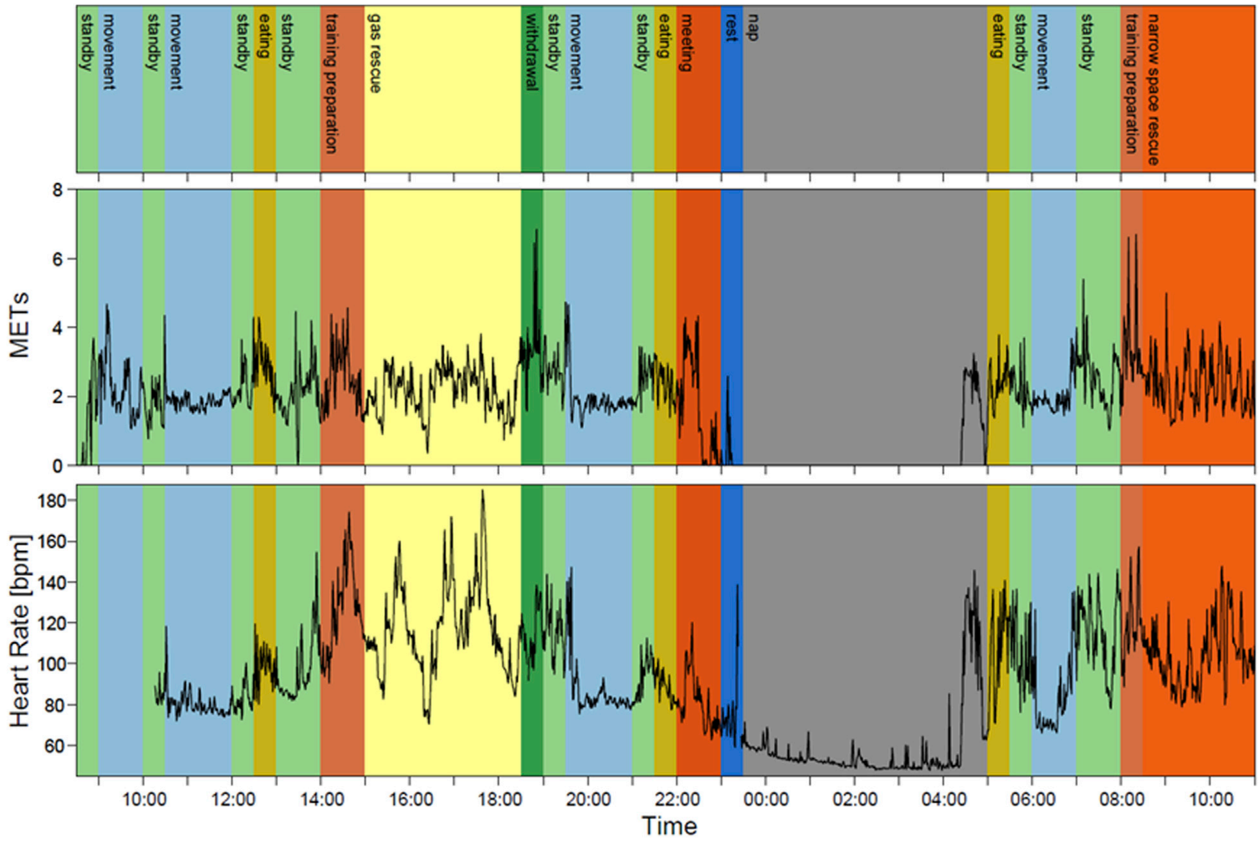


FIGURE 4
Typical result of METs and HR for one rescue worker during the 2-day disaster simulation training. The rescue worker engaged in rescue activity training in a gas leak accident (yellow) and excavation rescue in a narrow space (light orange). Colors were categorized for each activity: gray, nap; orange, meeting; light blue, movement; blue, rest; ocher, eating; light green, standby; green, withdrawal; yellow, gas leak accident; light orange, rescue narrow space; and brown, training preparation.

These two rescue squads were excluded from analysis because one rescue squad was only able to measure their HR for 15% during their nap time, and the activity of another rescue squad did not include withdrawal.

2.3.5 Statistical analysis

Data are presented as mean values and standard deviations.

3 Results

3.1 Typical example of measurements

We measured the physical activity and HR from 8:30 to 11:00 the next day. Figure 4 represents a typical example of METs and HR during simulated rescue operations (28 years, 173 cm, 72 kg).

TABLE 2 Heart rate during each activity.

Activity (METs)[a]	Heart rate (bpm)[b]	*n*
Nap (0.95 METs)	56.2 ± 13.5	27
Withdrawal (3.0 METs)	104.8 ± 17.2	27
Rescue training from vehicles (6.8 METs)	121.9 ± 24.1	5
Search activity (6.8 METs)	125.4 ± 20.7	10
Rescuer transport (6.8 METs)	132.9 ± 16.9	5
Rescue training in a collapsed building (8.0 METs)	116.7 ± 20.3	15
General firefighting (8.0 METs)	132.5 ± 14.6	5
Excavation rescue search activity (8.8 METs)	128.2 ± 13.8	9
Tunnel accident rescue activity (8.8 METs)	131.2 ± 15.7	4
Rescue activity training in a gas leak accident (9.0 METs)	131.2 ± 21.2	15
General rescue activity (9.0 METs)	135.1 ± 17.8	10
Rescue activity training in a skyscraper (10.0 METs)	154.9 ± 23.9	20

The results are presented as mean values ±SD.

[a]The table shows the activities extracted to create the equations for the relationship between HR, and METs; the figure is based on the MET, table (Garby et al., 1987; Ainsworth et al., 2011).

[b]HR, which was computed as the 10th percentile for nap, 50th percentile for withdrawal, and 90th percentile for other activities, is also shown. The number of rescue operations engaged in the activity is depicted as *n*.

The rescue worker engaged in rescue activity training in a gas leak accident (yellow) for 210 min, excavation rescue in narrow space (light orange) for 150 min, and nap (gray) for 330 min. The basic metabolic rate of the rescue worker, which is automatically calculated, was 1,600 kcal, and the total energy expenditure estimated by using the tri-axial accelerometer was 3,723 kcal (from 11:00 to 11:00 the next day). This example underestimated METs in the yellow and light-orange areas because the activities did not show high METs despite an increase in HR. Therefore, METs are more stable than HR during low-intensity activities, particularly during napping.

3.2 Calculation of HR for each activity

One rescue worker who conducted excavation rescue search and tunnel accident rescue activities could measure only the first 4 h (deficiency rate was 83.1%). Thus, the rescue worker was excluded from analysis. The percentage of missing HR data for each individual was 4.7% ± 4.4% (range: 0%–15.3%) for 27 rescue squads (age: 34.2 ± 7.5 years; BMI: 22.9 ± 1.5 kg/m^2). The HR was computed as the 10th percentile for naps, 50th percentile for withdrawal, and 90th percentile for other activities. The number of subjects in each activity is summarized in Table 2. Nap had the lowest HR, and rescue activity training in a skyscraper had the highest HR.

3.3 Basic metabolic rate and estimated total energy expenditure by the tri-axial accelerometer

The tri-axial accelerometer automatically calculated the basic metabolic rate, with an average of 1,545 ± 92 kcal, and the estimated total energy expenditure, with an average of 3,414 ± 229 kcal for 27 rescue squads (from 11:00 to 11:00 the next day).

3.4 Estimated total energy expenditure by combining the tri-accelerometer and HR

Figure 4 shows the estimation equation using two points (Figure 5A) created using the 10th percentile of HR during nap (0.95 METs) as the lowest HR and the 90th percentile of HR during rescue activity training in a gas leak accident (9.0 METs). Using three points (Figure 5B), the estimation equation was created using the abovementioned two points plus the intermediate 50th percentile of the HR during withdrawal (3.0 METs). The results showed that HR was higher during rescue training in a gas leak accident (9.0 METs) than during the excavation rescue in a narrow space (6.8 METs). Therefore, we developed an equation to estimate the highest HR for the latter (Figure 4). The energy expenditure calculated using the estimation equation replaced specific rescue activities, such as rescue activity training in a gas leak accident (yellow) and excavation rescue in a narrow space (light orange). The estimated total energy expenditure was 5,456 kcal using the two-point equation and 4,945 kcal using the three-point equation. The difference between the estimated total energy expenditure by the tri-axial accelerometer and that by combining the tri-axial accelerometer and HR was 1,733 kcal using the two-point equation and 1,222 kcal using the three-point equation.

The average value of R^2 in the estimation equation was 0.895 (95% confidence interval: 0.858–0.932) using the three-point equation (minimum, intermediate, and maximum). The average corrected total energy expenditure obtained by the tri-axial accelerometer and HR was 4,871 ± 486 kcal and 4,555 ± 391 kcal, respectively. The average difference between the estimated total energy expenditure by the tri-axial accelerometer and that by

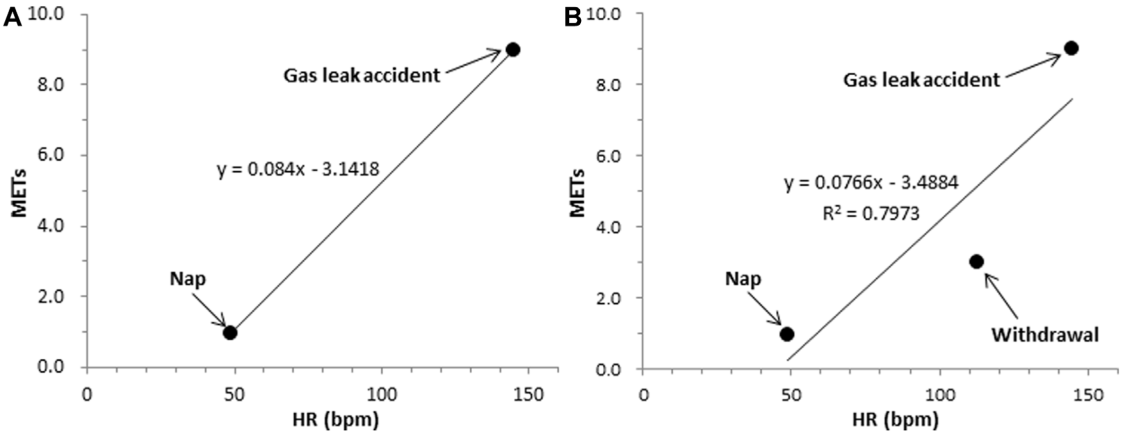


FIGURE 5
Estimation equation between HR and METs for a typical example of one rescue worker (Figure 4). **(A)** two points; **(B)** three points.

combining the tri-axial accelerometer and HR was 1,457 ± 486 kcal and 1,140 ± 410 kcal, respectively (accounting for an average increase of 143% and 133%, respectively).

4 Discussion

In this study, we developed a simple energy expenditure estimation method using wearables (tri-axial accelerometer and HR) for heavy load physical laborers to achieve a reliable estimation of energy expenditure for disaster relief and rescue workers as an expedition team. The relationship between the MET and individual HR in humans was also explored. If the minimum HR is known, then the maximum HR can be largely predicted; therefore, a model with some uniformity can be created without laboratory assessment. The strength of this study depends on the measurement of tri-axial acceleration, which can isolate each activity, capture a wide range of movement, as well as HR for rescue workers who experience various difficulties at the forefront, and estimate the total energy expenditure for each individual using a novel method in the field.

The MET concept, which expresses the intensity of physical activity by the time it corresponds to the resting metabolic rate, represents a simple, practical, and easily understood procedure to express the energy cost of physical activities as a multiple of the resting metabolic rate (Jetté et al., 1990); one MET represents the state of sitting and resting. In the present study, an accelerometer was used to store MET because energy expenditure can be calculated by multiplying the MET by a coefficient and individual body weight. The HR is high under the influence of mental and physical fatigue (Tanaka et al., 2011), stress, tension, and excitement (Cannon, 1929). Many researchers will conduct standardized laboratory assessments of resting and maximal HR before field testing using the submaximal test and/or Yo–Yo intermittent endurance test to determine maximal HR in the field (Sell and Ledesma, 2016). This approach is advantageous because the HR data in the field can be normalized to each individual, and other researchers can easily replicate those standardized tests. However, if only the minimum HR is known in humans, then the maximum HR can be largely predicted, that is, %HRR. Consequently, an estimation equation model with some uniformity can be created. Previous research demonstrated that estimating the total energy expenditure by using only an accelerometer is underestimated compared with the DLW (Touno et al., 2003; Kinnunen et al., 2019; Murakami et al., 2019; Sato et al., 2021; Willis et al., 2022). Therefore, given the nature of accelerometers, accurately measuring the total energy expenditure during heavy equipment or static load activities is difficult (Touno et al., 2003). In the present study, we combined METs and HR to improve underestimation.

METs differ depending on posture; therefore, we hypothesized that the METs during sitting and resting are different. Thus, 0.95 MET was derived from the original paper (Garby et al., 1987) of METs for a nap. We defined the minimum HR as the 10th percentile of nap time because the participating rescue squads operating in a blind training environment where they did not know what to do next slept very differently than usual and slept on their cots. In this study, the average HR during nap was 68.3 ± 15.3 bpm/min, which may also be affected by the abovementioned factors. The maximum HR was set at the 90th percentile of each rescue operation's activity because the same HR did not last forever, although the same activity was continued. Activity and activity time were classified on the basis of self-reported recording forms; however, the actual time spent waiting was also included, which may have reduced the HR during the activity. Although the training was conducted at relatively low temperatures in winter, the HR, which was set at the 90th percentile for training in winter conditions, may need to be higher because the load on the body is expected to be higher in summer than in winter. As rescue operations include various activities, not all activities can be applied to appropriate METs. However, a positive correlation was observed between METs and HR, indicating that the wearable device could accurately measure HR without burdening the subject. The HR values during multiple activities were consistent with those reported in previous studies (Rodahl, 1989; Parker et al., 2017). Average HRs during all tasks ranged from 110 to 130 bpm (Rodahl, 1989) and from 145 to 109 bpm for steep and flatlands, respectively (Parker et al., 2017). We defined withdrawal as an intermediate HR as it was an activity carried by all groups. In most rescue squad estimation equations, the intermediate point shifts downward. In this study, the METs may not be appropriate because withdrawal includes activities such as carrying heavy loads, moving, and feeling fatigue. Based on the error propagation law, the error in the estimated $\mathrm{METS}_{\mathrm{act}}$ depends on the accuracy of the measurements in METs and HR at the assumed maximum intensity. Improving the accuracy of those measurements with the aid of laboratory measurements is a future challenge.

A previous systematic review reported that tri-axial and multisensory devices tend to be more valid monitors compared with the DLW while adding indicators such as HR and heat flux values to acceleration values to estimate daily energy expenditure, and physical activity has only slightly improved the system (Van Remoortel et al., 2012). In addition, the combination of motion sensor and HR is valid for estimating free-living energy expenditure compared with the DLW, but it is less accurate for an individual assessment (Sliva et al., 2015). Kortelainen et al. (2021) proposed accurate energy expenditure predictions based on a few calibration measurements using a nonlinear (logistic) mixed model for energy expenditure and HR. They found that the logistic mixed model performed better than the linear mixed model when predicting energy expenditure at population level and with calibration. In this study, using the linear model, the result of the estimated equation with three points was 94% (−316 kcal) compared with the results of the two points. Therefore, no moderate-intensity activity that did not affect HR. In addition, using the tri-axial accelerometer method, the estimation equation with two points of total energy expenditure was underestimated by 43% compared with the combined accelerometer and HR. In the present study, we able to improve the underestimation of total energy expenditure, but we could not examine whether it was an overestimation or an accurate assessment. The novel estimation equation of the two points between the METs and HR is simple, and it has high validity as an estimation equation for energy expenditure in the field. Considering its application in the field, the two minimum and maximum estimation equations used in this study are sufficient. In estimating the energy expenditure per activity for a short period, rather than the average energy expenditure for 1 day or several days, combining tri-axial accelerometers and HR was easy and effective. Our research objective was to estimate energy expenditure in a real-world environment. Since our approach was a real-world data-driven study, rather than a traditional experimental design-driven study, there were many limitations on the measurement. Under such real-world, non-uniform, and *a priori* unforeseeable work environments, we believe that our method gave improved estimation results compared to the results of previous studies that estimated using only acceleration. Also, the results we refer to are the widely-accepted reference values, METs, estimated based on experimental studies in previous studies. Nevertheless, indeed, the values estimated based on experimental studies do not take into account factors such as individual differences, mental load, different types of exercise load, the type of heavy equipment, effects of circadian rhythm, effects of diet, and effects of sleep. It is difficult to conduct an experiment that takes all these factors into account. We also believe that in real-world, non-uniform, and *a priori* unforeseeable work environments, accurate estimation of individual tasks is difficult, whether using the DLW or other methods. We are sure that the accuracy can be improved by combining it with laboratory studies in the future. Beckner et al. (2023) reported sustained operations for the military personnel are often conducted in a state of negative energy balance and are associated with degraded cognitive performance and mood. In order to prevent such condition, the amount of energy ingested should also be considered in the future. The accuracy of the estimation equation can be further improved by accurately extracting the reference activities of METs. Moreover, estimating energy expenditure for a wide range of age groups, sex, and people with different body sizes, such as obese and overweight, may require an estimation equation with three or more points, instead of just two, in the future.

This study had four limitations. First, we focused only on HR because the measurements were conducted in winter. However, the temperature should also be considered because the HR can quickly increase depending on environmental conditions, particularly during summer. In addition, an equation that considers the circadian rhythm of the heartbeat should be developed. Second, the activities were calculated by applying them to the MET table; however, applying them to special rescue operation activities was difficult because appropriate METs could not be found. Therefore, the validity of METs remains to be examined using the Douglas bag and other methods. Third, we did not consider the diet induced thermogenesis, because it was automatically calculated by entering each individual's data into the accelerometer. Finally, the present study was conducted in the field, and no controls were used to verify the energy expenditure calculations. In the future, the energy expenditure remained to be examined using the Douglas bag and DLW.

In this study, a method to improve energy expenditure estimation for heavy-load physical labor, such as disaster relief and rescue operations, using a tri-axial accelerometer and HR monitor was proposed. Energy expenditure, which was underestimated by accelerometer-based energy expenditure methods, could be compensated by creating an individually optimized estimation equation between METs and individual HR. Furthermore, more detailed measurements were necessary for a large number of rescue operations in a wide range of ranks and firefighter activities in the future.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: All data files are available from the figshare database (https://figshare.com/articles/dataset/Estimation_equation_model_/21579168).

Ethics statement

The studies involving humans were approved by the local ethics committee of the University of Tsukuba (approval number: tai-28-66). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin because detailed explanations were given to each training director and participating organization-affiliated institution regarding the purpose and content of the experiment, and the experiment was started only after obtaining agreement.

Author contributions

HO: Conceptualization, Funding acquisition, Writing–original draft. YN: Investigation, Writing–review and editing. NK: Formal Analysis, Investigation, Writing–review and editing. HN: Investigation, Writing–review and editing. MK: Methodology, Writing–review and editing. KK: Data curation, Methodology, Writing–review and editing. NO: Conceptualization, Supervision, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Advanced Research Initiative for Human High Performance (ARIHHP, University of Tsukuba) under Grant [2017, 2018, 2019, 2020, and 2021].

Acknowledgments

We are extremely grateful to all the firefighters who cooperated with our research.

Conflict of interest

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

Publisher's note

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

References

Ades, P. A., Savage, P. D., Brochu, M., Tischler, M. D., Lee, N. M., and Poehlman, E. T. (2005). Resistance training increases total daily energy expenditure in disabled older women with coronary heart disease. *J. Appl. Physiol.* 98 (4), 1280–1285. doi:10.1152/japplphysiol.00360.2004

Ainsworth, B. E., Haskell, W. L., Herrmann, S. D., Meckes, N., Bassett, D. R., Jr, Tudor-Locke, C., et al. (2011). 2011 Compendium of Physical Activities: a second update of codes and MET values. *Med. Sci. Sports Exerc* 43 (8), 1575–1581. doi:10.1249/MSS.0b013e31821ece12

Aparicio-Ugarriza, R., Mielgo-Ayuso, J., Benito, P. J., Pedrero-Chamizo, R., Ara, I., González-Gross, M., et al. (2015). Physical activity assessment in the general population; instrumental methods and new technologies. *Nutr. Hosp.* 31 (Suppl. 3), 219–226. doi:10.3305/nh.2015.31.sup3.8769

Arikawa, T., Nakajima, T., Yazawa, H., Kaneda, H., Haruyama, A., Obi, S., et al. (2020). Clinical usefulness of new R-R interval analysis using the wearable heart rate sensor WHS-1 to identify obstructive sleep apnea: OSA and RRI analysis using a wearable heartbeat sensor. *J. Clin. Med.* 9 (10), 3359. doi:10.3390/jcm9103359

Beckner, M. E., Lieberman, H. R., Hatch-McChesney, A., Allem, J. T., Niro, P. J., Thompson, L. A., et al. (2023). Effects of energy balance on cognitive performance, risk-taking, ambulatory vigilance and mood during simulated military sustained operations (SUSOPS). *Physiol. Behav.* 258, 114010. doi:10.1016/j.physbeh.2022.114010

Berryman, C. E., Sepowitz, J. J., McClung, H. L., Lieberman, H. R., Frrina, E. K., McClung, J. P., et al. (2017). Supplementing an energy adequate, higher protein diet with protein does not enhance fat-free mass restoration after short-term severe negative energy balance. *Appl. Physiol.* 122, 1485–1493. doi:10.1152/japplphysiol.01039.2016

Bouten, C. V., Koekkoek, K. T., Verduin, M., Kodde, R., and Janssen, J. D. (1997). A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. *IEEE Trans. Biomed. Eng.* 44 (3), 136–147. doi:10.1109/10.554760

Cannon, W. B. (1929). *Bodily changes in pain, hunger, fear and rage: an account of recent research into the function of emotional excitement*. 2nd ed. New York: Appleton-Century-Crofts.

Chen, K. Y., Janz, K. F., Zhu, W., and Brychta, R. J. (2012). Re-defining the roles of sensors in objective physical activity monitoring. *Med. Sci. Sports Exerc* 44 (Suppl. 1), S13–S23. doi:10.1249/MSS.0b013e3182399bc8

Chevance, G., Golaszewski, N. M., Tipton, E., Hekler, E. B., Buman, M., Welk, G. J., et al. (2022). Accuracy and precision of energy expenditure, heart rate, and steps measured by combined-sensing Fitbits against reference measures: systematic review and meta-analysis. *JMIR Mhealth Uhealth* 10 (4), e35626. doi:10.2196/35626

Fuller, D., Clowell, E., Low, J., Orychock, K., Tobin, M. A., Simango, B., et al. (2020). Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. *JMIR Mhealth Uhealth* 8 (9), e18694. doi:10.2196/18694

Ganpile, A. A., Tanaka, S., Ishikawa-Takata, K., and Tabata, I. (2007). Interindividual variability in sleeping metabolic rate in Japanese subjects. *Eur. J. Clin. Nutr.* 61 (11), 1256–1261. doi:10.1038/sj.ejcn.1602645

Garby, L., Kurzer, M. S., Lammert, O., and Nielsen, E. (1987). Energy expenditure during sleep in men and women: evaporative and sensible heat losses. *Hum. Nutr. Clin. Nutr.* 41C, 225–233.

Hill, J. O., and Trowbridge, E. L. (1998). Childhood obesity: future directions and research priorities. *Pediatrics* 101 (3 Pt 2), 570–574. doi:10.1542/peds.101.3.570

Isakadze, N., and Martin, S. S. (2020). How useful is the smartwatch ECG? *Trends Cardiovasc Med.* 30 (7), 442–448. doi:10.1016/j.tcm.2019.10.010

Jackson, A. S., Blair, S. N., Mahar, M. T., Wier, L. T., Ross, R. M., and Stuteville, J. E. (1990). Prediction of functional aerobic capacity without exercise testing. *Med. Sci. Sports Exerc* 22 (6), 863–870. doi:10.1249/00005768-199012000-00021

Jetté, M., Sidney, K., and Blümchen, G. (1990). Metabolic equivalents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity. *Clin. Cardiol.* 13, 555–565. doi:10.1002/clc.4960130809

Karmen, C. L., Reisfel, M. A., McIntyre, M. K., Timmermans, R., and Frishman, W. (2019). The clinical value of heart rate monitoring using an Apple Watch. *Cardiol. Rev.* 27 (2), 60–62. doi:10.1097/CRD.0000000000000243

Kinnunen, H., Häkkinen, K., Schumann, M., Karavurta, L., Westerterp, K. R., and Kyröläinen, H. (2019). Training-induced changes in daily energy expenditure: methodological evaluation using wrist-worn accelerometer, heart rate monitor, and doubly labeled water technique. *PLoS One* 14 (7), e0219563. doi:10.1371/journal.pone.0219563

Koizumi, N., Negishi, Y., Ogata, H., Rakwal, R., and Omi, N. (2021). Estimating total energy expenditure for fire-fighters during large scale disaster response training using a tri-axial accelerometer. *Nutrients* 13 (8), 2789. doi:10.3390/nu13082789

Kortelainen, L., Helske, J., Finni, T., Mehtätalo, L., Tikkanen, O., and Kärkkäinen, S. (2021). A nonlinear mixed model approach to predict energy expenditure from heart rate. *Physiol. Meas.* 42 (3), 035001. doi:10.1088/1361-6579/abea25

Kramer, T. R., Moore, R. J., Shippee, R. L., Friedl, K. E., Martinez-Lopez, L., Chan, M. M., et al. (1997). Effects of food restriction in military training on T-lymphocyte responses. *Int. J. Sports Med.* 18 (Suppl. 1), S84–S90. doi:10.1055/s-2007-972704

Margolis, L. M., Murphy, N. E., Martini, S., Spitz, M. G., Thrane, I., McGraw, S. M., et al. (2014). Effects of winter military training on energy balance, whole-body protein balance, muscle damage, soreness, and physical performance. *Appl. Physiol. Nutr. Metab.* 39, 1395–1401. doi:10.1139/apnm-2014-0212

Martin, C. K., Das, S. K., Lindblad, L., Racette, S. B., McCrory, M. A., Weiss, E. P., et al. (2011). Effect of calorie restriction on the free-living physical activity levels of nonobese humans: results of three randomized trials. *J. Appl. Physiol.* 110, 956–963. doi:10.1152/japplphysiol.00846.2009

Midorikawa, T., Tanaka, S., Kaneko, K., Koizumi, K., Ishikawa-Takata, K., Futami, J., et al. (2007). Evaluation of low-intensity physical activity by triaxial accelerometry. *Obesity* 15 (12), 3031–3038. doi:10.1038/oby.2007.361

Murakami, H., Kawakami, R., Nakae, S., Yamada, Y., Nakata, Y., Ohkawara, K., et al. (2019). Accuracy of 12 wearable devices for estimating physical activity energy expenditure using a metabolic chamber and the doubly labeled water method: validation study. *JMIR Mhealth Uhealth* 7 (8), e13938. doi:10.2196/13938

Nagayoshi, S., Oshima, Y., Ando, T., Aoyama, T., Nakae, S., Usui, C., et al. (2019). Validity of estimating physical activity intensity using a triaxial accelerometer in healthy adults and older adults. *BMJ Open Sport Exerc Med.* 5 (1), e000592. doi:10.1136/bmjsem-2019-000592

Ndahimana, D., and Kim, E. K. (2017). Measurement methods for physical activity and energy expenditure: a review. *Clin. Nutr. Res.* 6 (2), 68–80. doi:10.7762/cnr.2017.6.2.68

Nishida, Y., Tanaka, S., Nakae, S., Yamada, Y., Morino, K., Kondo, K., et al. (2020). Validity of the use of a triaxial accelerometer and a physical activity questionnaire for estimating total energy expenditure and physical activity level among elderly patients with type 2 diabetes mellitus: CLEVER-DM study. *Ann. Nutr. Metab.* 76 (1), 62–72. doi:10.1159/000506223

Ohkawara, K., Oshima, Y., Hikihara, Y., Ishikawa-Takata, K., Tabata, I., and Tanaka, S. (2011). Real-time estimation of daily physical activity intensity by a triaxial accelerometer and a gravity-removal classification algorithm. *Br. J. Nutr.* 105, 1681–1691. doi:10.1017/S0007114510005441

Oshima, Y., Kawaguchi, K., Tanaka, S., Ohkawara, K., Hikihara, Y., Ishikawa-Takata, K., et al. (2010). Classifying household and locomotive activities using a triaxial accelerometer. *Gait Posture* 31, 370–374. doi:10.1016/j.gaitpost.2010.01.005

Parker, R., Vitalis, A., Walker, R., Riley, D., and Pearce, H. G. (2017). Measuring wildland fire fighter performance with wearable technology. *Appl. Ergon.* 59, 34–44. doi:10.1016/j.apergo.2016.08.018

Planer, D., Leibowitz, D., Hadid, A., Erlich, T., Sharon, N., Paltiel, O., et al. (2012). The effect of prolonged physical activity performed during extreme caloric deprivation on cardiac function. *PLoS One* 7 (2), e31266. doi:10.1371/journal.pone.0031266

Plasqui, G., Joosen, A. M., Kester, A. D., Goris, A. H., and Westerterp, K. R. (2005). Measuring free-living energy expenditure and physical activity with triaxial accelerometry. *Obes. Res.* 13 (8), 1363–1369. doi:10.1038/oby.2005.165

Plasqui, G., and Westerterp, K. R. (2007). Physical activity assessment with accelerometers: an evaluation against doubly labeled water. *Obesity* 15 (10), 2371–2379. doi:10.1038/oby.2007.281

Rodahl, K. (1989). *Physiology of work*. London: CRC Press.

Sato, H., Nakamura, H., Nishida, Y., Shirahata, T., Yogi, S., Akagami, T., et al. (2021). Energy expenditure and physical activity in COPD by doubly labelled water method and an accelerometer. *ERJ Open Res.* 7 (2), 00407–02020. doi:10.1183/23120541.00407-2020

Sell, K. M., and Ledesma, A. B. (2016). Heart rate and energy expenditure in division I field hockey players during competitive play. *J. Strength Cond. Res.* 30 (8), 2122–2128. doi:10.1519/JSC.0000000000001334

Sliva, A. M., Santos, D. A., Matias, C. N., Júdice, P. B., Magalhães, J. P., Ekelund, U., et al. (2015). Accuracy of a combined heart rate and motion sensor for assessing energy expenditure in free-living adults during a double-blind crossover caffeine trial using doubly labeled water as the reference method. *Eur. J. Clin. Nutr.* 69 (1), 20–27. doi:10.1038/ejcn.2014.51

Strath, S. J., Swartx, A. M., Bassett, D. R., Jr, O'Brien, W. L., King, G. A., and Ainsworth, B. E. (2000). Evaluation of heart rate as a method for assessing moderate intensity physical activity. *Med. Sci. Sports Exerc* 32 (9 Suppl. l), S465–S470. doi:10.1097/00005768-200009001-00005

Swain, D. P., Leutholtz, B. C., King, M. E., Haas, L. A., and Branch, J. D. (1998). Relationship between % heart rate reserve and % VO_2 reserve in treadmill exercise. *Med. Sci. Sports Exerc* 30 (2), 318–321. doi:10.1097/00005768-199802000-00022

Tanaka, M., Mizuno, K., Yamaguti, K., Kuratsune, H., Fujii, A., Baba, H., et al. (2011). Autonomic nervous alterations associated with daily level of fatigue. *Behav. Brain Funct.* 7, 46. doi:10.1186/1744-9081-7-46

Touno, M., Hasina, R. H., Ebine, N., Peng, H. Y., Yoshitake, Y., Tanaka, H., et al. (2003). Measurement of total energy expenditure in Japanese Firefighters under normal working condition using the doubly labelled water method. *Jpn. J. Phys. Fit. Sports Med.* 52, 265–274. doi:10.7600/jspfsm1949.52.265

Van Remoortel, H., Giavedoni, S., Raste, Y., Burtin, C., Louvaris, Z., Gimeno-Santos, E., et al. (2012). Validity of activity monitors in health and chronic disease: a systematic review. *Int. J. Behav. Nutr. Phys. Act.* 9, 84. doi:10.1186/1479-5868-9-84

Willis, E. A., Creasy, S. A., Saint-Maurice, P. F., Keadle, S. K., Pontzer, H., Schoeller, D., et al. (2022). Physical activity and total daily energy expenditure in older US adults: constrained versus additive models. *Med. Sci. Sports Exerc* 54 (1), 98–105. doi:10.1249/MSS.0000000000002759

Yamada, Y., Yokoyama, K., Noriyasu, R., Osaki, T., Adachi, T., Itoi, A., et al. (2009). Light-intensity activities are important for estimating physical activity energy expenditure using uniaxial and triaxial accelerometers. *Eur. J. Appl. Physiol.* 105 (1), 141–152. doi:10.1007/s00421-008-0883-7

TYPE Original Research
PUBLISHED 06 January 2025
DOI 10.3389/fphys.2024.1498399

Check for updates

OPEN ACCESS

EDITED BY
Colin K. Drummond,
Case Western Reserve University, United States

REVIEWED BY
Neal William Pollock,
Laval University, Canada
Claus-Martin Muth,
Universitaetsklinikum Ulm, Germany

*CORRESPONDENCE
Frank Pernett,
✉ frank.pernett@miun.se

RECEIVED 18 September 2024
ACCEPTED 29 November 2024
PUBLISHED 06 January 2025

CITATION
Pernett F, Mulder E, Johansson F, Sieber A, Bermudez R, Lossner M and Schagatay E (2025) Toward a hyperventilation detection system in freediving: a proof of concept using force sensor technology.
Front. Physiol. 15:1498399.
doi: 10.3389/fphys.2024.1498399

Toward a hyperventilation detection system in freediving: a proof of concept using force sensor technology

Frank Pernett[1,2]*, Eric Mulder[1], Filip Johansson[1], Arne Sieber[1,3], Ricardo Bermudez[4], Marcus Lossner[5] and Erika Schagatay[1,2]

[1]Environmental Physiology Group, Department of Health Sciences, Mid Sweden University, Östersund, Sweden, [2]Swedish Winter Sports Research Centre, Department of Health Sciences, Mid Sweden University, Östersund, Sweden, [3]Oxygen Scientific GmbH, Graz, Austria, [4]Sensing Systems Corporation, Dartmouth, United States, [5]Independent hardware and software engineer, Atlanta, GA, United States

Background and aim: Hyperventilation before breath-hold diving (freediving) is widely accepted as a risk factor for hypoxic syncope or blackout (BO), but there is no practical way to address it before dives. This study explores the feasibility of using a force sensor to predict end-tidal carbon dioxide ($P_{ET}CO_2$) to assess hyperventilation in freedivers.

Methods and results: Twenty-one freedivers volunteered to participate during two national competitions. The divers were instructed to breathe normally and perform three dry apneas of 1, 2, and 3-min duration at 2-min intervals in a sitting position. Before and after the apneas, $P_{ET}CO_2$ was recorded. The signal from the force sensor, attached to a chest belt, was used to record the frequency and amplitude of the chest movements, and the product of these values in the 60 s before the apnea was used to predict $P_{ET}CO_2$. The mean $P_{ET}CO_2$ was below 35 mmHg before all apneas. The mean amplitude of the signal from the force sensor increased from apnea 1 to apnea 3 ($p < 0.001$), while the respiratory rate was similar (NS). The product of the respiratory rate and amplitude from the force sensor explained 34% of the variability of the $P_{ET}CO_2$ in the third apnea.

Conclusion: This study shows that a force sensor can estimate hyperventilation before static apnea, providing a basis for further research. More studies are needed to confirm its effectiveness in preventing issues. Freedivers may hyperventilate without noticing it, and such a system could improve awareness of this condition. Additional underwater tests are essential to determine whether this system can enhance safety in freediving.

KEYWORDS

tidal volume, breath-hold, apnea, blackout, wearable technology

Abbreviations: BO, blackout; eMv, estimated minute ventilation.

Introduction

Breath-hold divers, also referred to as freedivers, often employ a breathing pattern known as hyperventilation to extend their apnea duration by reducing the alveolar carbon dioxide pressure ($PACO_2$). Volitional hyperventilation is a conscious effort to increase the breathing rate and depth, which increases alveolar ventilation, leading to a slight increase in alveolar oxygen pressure (PAO_2), a reduction in $PACO_{2,}$ lowering of arterial CO_2 levels (hypocapnia), and an elevation in pH (West and Luks, 2021). In contrast, metabolism-driven hyperventilation is an automatic, homeostatic response to increased metabolic activity, such as during exercise, to expel excess CO_2 and stabilize blood gas levels (Forster et al., 2012). Hyperventilation before breath-hold diving, despite a slight increase in arterial oxygen pressure (PaO_2), leads to a greater risk of losing consciousness underwater (Craig, 1961; 1976; Edmonds and Walker, 1999; Lippmann and Pearn, 2012), as the control of ventilation relies on chemoreceptors that respond to changes in $PaCO_2$ and hydrogen ion (H^+) levels. As $PaCO_2$ decreases due to hyperventilation, the ventilatory drive is compromised, leading to a delayed urge to breathe. Consequently, this results in an extended apnea duration (Hill, 1973; Lin et al., 1974; Bain et al., 2017; Pernett et al., 2023). A longer apnea duration increases the level of hypoxia during a breath-hold, which exposes the freediver to an increased risk of losing consciousness underwater, known as blackout (BO) or hypoxic syncope (Lindholm and Gennser, 2005; Kumar and Ng, 2010). The risk of severe oxygen desaturation has recently been found to be exacerbated during repeated series of apnea after short-time hyperventilation of 15 s (Pernett et al., 2023).

Hyperventilation is reported as a risky practice before breath-holding among recreational swimmers (Boyd et al., 2015), spearfishers (Lippmann, 2019), and snorkelers (Dunne et al., 2021). This breathing pattern increases apnea duration and desaturation, increasing the risk of BO, with the potential consequence of drowning if not promptly addressed. In addition, hyperventilation can reduce cerebral blood flow by 2% for each 1 mmHg of decline in $PaCO_2$ (Raichle and Plum, 1972). Experienced freedivers exhibit enhanced tolerance to hypoxia as training seems to diminish their hypoxic ventilatory response (Schneeberger et al., 1986; Ferretti et al., 1991; Lindholm and Lundgren, 2006). This suggests that trained freedivers, when engaging in hyperventilation, may experience pronounced hypoxemia since they depend on the hypoxic stimulus to terminate the breath-hold.

Despite the evidence contradicting the benefits of hyperventilation, there remains a significant gap in knowledge concerning the prevalence and role of this practice among competitive freedivers, snorkelers, and spearfishers. Some insights into this issue have emerged from blood gas analyses in studies characterized by relatively modest sample sizes. In elite freedivers, documented pre-diving $PaCO_2$ levels vary, with reported values of 29 mmHg (Molchanova et al., 2020), 26 mmHg (Muth et al., 2003), and 21 mmHg (Scott et al., 2021). In contrast, non-elite breath-hold divers and Ama divers exhibit pre-diving values within the normal range, registering $PaCO_2$ levels of 38 ± 3 mmHg (mean ± SD; Bosco et al., 2018) and 42 ± 2 mmHg (mean ± SD; Qvist et al., 1993), respectively. However, evaluating pre-apnea $PaCO_2$ or $P_{ET}CO_2$ in non-laboratory settings, during diving, poses practical challenges. Blood gas analysis, while providing precise data, demands specific expertise and is invasive. Similarly, measuring exhaled gases requires equipment susceptible to damage in aquatic environments. An alternative strategy involves the measurement of tidal volume (Vt) and respiratory rate (RR) to estimate the minute ventilation at rest.

Various studies have explored methods employing stretch, piezoelectric, optical, pressure, electromagnetic, or acoustic sensors; accelerometers; and electrical impedance techniques for estimating Vt and RR (Panahi et al., 2020; Monaco and Stefanini, 2021). These techniques essentially aim to monitor alterations in thoracic and abdominal movements to infer Vt. However, applying these sensors in the underwater environment is complex. Our laboratory has constructed a unique underwater monitor involving a force sensor in a buckle attached to a chest belt, allowing detailed chest movement recording (Sieber et al., 2022). The main goal of the current study was to determine whether respiratory data obtained using this custom-built force sensor could predict $P_{ET}CO_2$ before a static breath-hold to assess the practice of hyperventilation in freedivers. Additionally, the secondary aim was to visually assess the signal quality when the sensor was used on divers in the pool.

Methods

Device description

The concept involved designing a U-shaped buckle (Figure 1A) that is both water- and pressure-proof and equipped with integrated strain gauges. These gauges enable the detection of pulling forces exerted on the buckle, facilitating the monitoring of alterations in chest circumference during underwater activities (Figure 1C). The description of this force sensor has been previously documented (Sieber et al., 2022). The custom buckle was crafted through a conventional biomechanical engineering design process employing machine drawing techniques. The choice of stainless steel as the buckle material was made to ensure its suitability for use in saltwater environments. The sensor was calibrated by applying a force of 10 N to one of its legs, producing a slight bending of the section connecting the two legs (Figure 1A). The applied force was correlated with the electrical signal from the sensor. To ensure water resistance, the entire buckle, including the strain gauges, was coated with a multipurpose rubber coating (Plasti Dip International, Blaine, MN). The strap from a commercially available Polar heart rate belt (Polar T34, Polar Electro, OY, Finland) was used to place the buckle on the chest (Figure 1B).

An improved version of a data logger, constructed by our laboratory previously, was used to read out the signals of the sensor-equipped buckle (Mulder et al., 2021; Figures 1A, D). Due to the low amplitude of the signals of the sensor-equipped buckle, it was necessary to employ an amplifier and a high-resolution analog-to-digital converter. We opted for the AD7192 analog-to-digital converter by Analog Devices, which is specifically designed for strain gauge signal acquisition. This integrated circuit combines a programmable gate array with a maximum 128x amplification, a 24-bit sigma–delta ADC, and a filtering stage, which effectively suppresses noise, particularly from 50 or 60 Hz power lines. Further improvements to the data logger included a

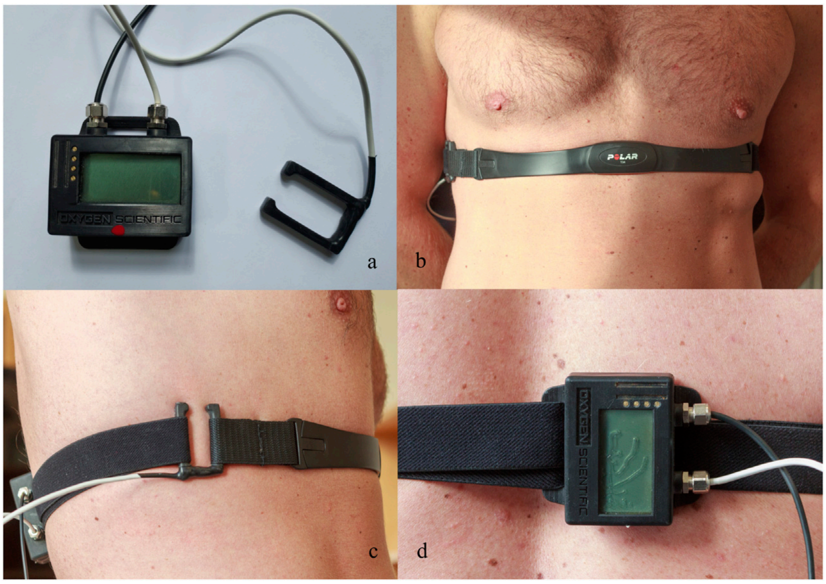


FIGURE 1
Custom-made buckle equipped with four strain gauges aligned on the sensor to measure the strain created by the applied force along with the data logger **(A)**. Frontal **(B)** and lateral **(C)** view of the buckle in its operational position attached to the chest strap. Details of the data logger in operational position on the back of the participant **(D)**.

USB port, a Bluetooth module, a digital pressure sensor, and a 3 × 16 character LC display.

Vital capacity calibration

The calibration procedure aimed to test the accuracy of the force sensor to estimate the vital capacity (VC). Details about the VC calibration are presented in Supplementary Material. The equation for the predicted VC was VC = 1.6 + (0.4396 × amplitude). The difference between the measured VC and the predicted VC was 0.00 ± 0.7 L (Supplementary Figure S1B).

Participants

The study included 21 participants (5 female and 16 male) with a mean ± SD age of 44 ± 7 years, a height of 178 ± 9 cm, a weight of 73 ± 10 kg, and a lung vital capacity of 5.82 ± 1.22 L. All participants were trained freedivers. Their training load was 5 ± 7 h per week. The study was conducted during two freediving national competitions. The divers competed in four pool disciplines. The participants received written and oral information on the protocol, after which they signed an informed consent document. The protocol was approved by the Swedish Research Ethics Authorities (EPM; #2019-05147) and complied with the Helsinki Declaration of 2004, apart from preregistration in a database.

Study design

The study involved a dry static apnea ramp test with durations of 1, 2, and 3 min, respectively, spaced by 2 min of recovery (Figure 2).

Procedures

A field laboratory was setup within the same pool area where the competitions took place. Participants were required to have a minimum of 12 h of rest following maximal performance and at least 2 h of fasting before initiating the test. Height, weight, and slow vital capacity were measured in triplicate in standing conditions, and the largest volume was used (Compact Expert, Vitalograph, Buckingham, United Kingdom). The participants filled out a questionnaire with information on the training load and personal best achievements in different freediving disciplines in the last 12 months. The participants performed a series of three apneas with fixed duration in dry conditions and the seated position (Figure 2). A researcher carried out a 2-min countdown before starting. At 30 s before apnea, a nose clip was applied, and 20 s before apnea, a mouthpiece was offered to breathe through. Ten seconds before the apnea, the countdown continued second by second.

Participants were instructed to exhale completely and then take a large, but not maximal, inhalation before starting the apnea voluntarily; this technique results in a volume of approximately 80 - 85% of the vital capacity (Schagatay and Holm, 1996). The participants were instructed to avoid hyperventilation. An experimenter closely monitored peripheral arterial oxygen saturation S_pO_2 and was ready to interrupt the apnea should it fall below 65%. The room temperature was 26.9°C ± 2.0°C.

Measurements

$P_{ET}CO_2$ was measured before and after every apnea via an infrared-based gas measurement module (LifeSense LS1-9R,

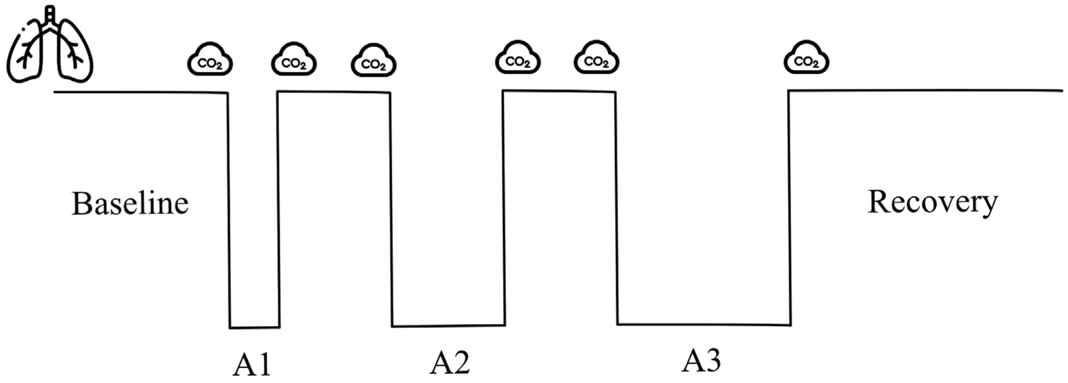


FIGURE 2
Apnea test protocol, involving apneas of 1, 2, and 3 min duration (A1–A3). Icons represent the time of vital capacity and exhaled CO_2 measurements. Baseline (3-min). Between apnea breathing intervals (2-min). Recovery (5-min).

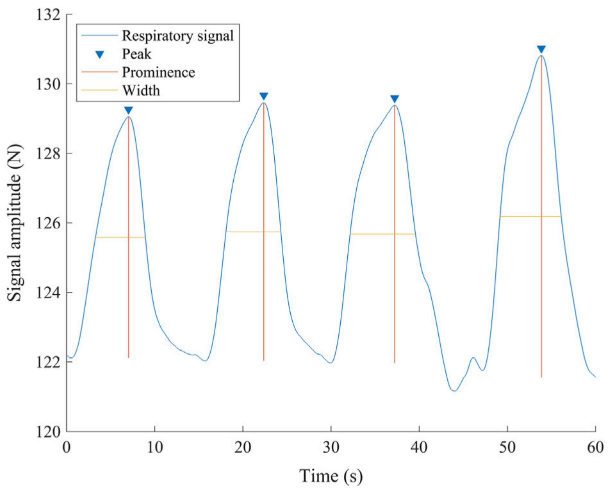


FIGURE 3
Representative respiratory signal from one participant 60 s before the breath-hold. The inverted triangles show the detected peaks of the signal, corresponding to the end of inhalation, and are used to calculate the respiratory rate. The vertical lines show the amplitude of every wave (prominence) used as a surrogate of tidal volume. The horizontal lines show the duration measured at the midpoint of the prominence (width). N, newton; s, seconds.

Nonin Medical Inc., Plymouth, United States). The diver breathed through a disposable mouthpiece with a bacterial filter connected to a T-valve with two one-way valves (AFT21, Biopac Systems, Goleta, United States). S_pO_2 and heart rate (HR) were measured using a reflectance sensor (800R, Nonin Medical Inc., Plymouth, United States) placed on the forehead 1 cm over the left eye and connected to a clinical monitor (LifeSense, Nonin Medical Inc., Plymouth, United States). Breathing movements were measured continuously using the prototype force sensor (Sieber et al., 2022).

Six male freedivers were also outfitted with the prototype force sensor prior to engaging in their freediving competition performances. This was done to assess the signal quality in the underwater environment.

Data analysis

$P_{ET}CO_2$ used was the highest value measured after the last exhalation before the apnea. The data from the data logger were extracted and analyzed with custom-made scripts using MATLAB (R2022b, MathWorks Inc., Natick, United States). Within the final minute leading up to the last exhalation before commencing the breath-hold, both breathing frequency and signal amplitude were extracted for analysis. The number of peaks in the respiratory signal represented the RR, and the prominence of the signal was used as the surrogate of Vt. With those values, the estimated minute ventilation (eMv) was calculated as the product of the RR and the amplitude of the prominence (RR x prominence; Figure 3).

Statistical analysis

The statistical analysis was carried out using SPSS 27 software (IBM Corp, Armonk, United States). The data were tested for normality using Shapiro–Wilk test and are reported as the mean ± SD. Outliers were defined as cases with a studentized deleted residual greater than three standard deviations (SD). A Spearman's correlation test was run to assess the relationship between VC and $P_{ET}CO_2$ and the amplitude of the signal from the force sensor. A one-way repeated measures analysis of variance (ANOVA) was used to compare VC and amplitude of the respiratory signal and compare $P_{ET}CO_2$, amplitude, and eMv before every apnea. The Bonferroni correction for multiple comparisons was applied. Significance was observed at $p < 0.05$. A linear regression was run to predict PE_TCO_2 from the eMV before every apnea. The Bland–Altman method was used to assess the agreement between the measured VC and predicted VC and between measured PE_TCO_2 and predicted PE_TCO_2 (Bland and Altman, 1986). The accepted clinical limits of agreement (LOA) for capnography are ≤5 mmHg (Wu et al., 2003), but the LOA between PE_TCO_2 and $PaCO_2$ could be as large as + 31 mmHg (İşat et al., 2023). When comparing two methods for measuring PE_TCO_2, the LOA could be 11 mmHg (Tamashiro et al., 2023). We set the accepted LOA to ≤10 mmHg. Effect sizes were estimated by the

TABLE 1 Pre-apnea respiratory values and data from respiratory buckle signal.

	$P_{ET}CO_2$ (mmHg)		RR (bpm)	Amplitude (N)	Vt (L)	eMv (N/min)
	Measured	Predicted				
A1#	31 ± 5	31 ± 3	10 ± 3	2.9 ± 1.8	2.87 ± 2.39	25.6 ± 12.6
A2	31 ± 7	33 ± 4	8 ± 2*	3.7 ± 2.5*	3.23 ± 2.70	28.7 ± 16.1
A3	29 ± 6*	30 ± 4	9 ± 3	4.0 ± 2.4*	3.36 ± 2.66	32.4 ± 14.1*

Values are presented as the mean ±1 SD. #n = 20. $P_{ET}CO_2$, end-tidal exhaled pressure of carbon dioxide; RR, respiratory rate; bpm, breaths per minute; N, newton; Vt, calculated tidal volume; L, liters, eMV, estimated minute ventilation calculated as the product of the RR and the amplitude of the prominence in the respiratory signal; A1–A3, apnea 1 to apnea 3. *Significantly different from A1.

partial eta squared (η_p^2) and the generalized eta squared (η_G^2) and are presented with a 90% confidence interval (CI). An effect size of 0.01–0.05 was considered small, 0.06–0.13 was considered medium, and 0.14 and above was considered large (Cohen, 1988; Bakeman, 2005; Lakens, 2013).

End-tidal carbon dioxide ($P_{ET}CO_2$) data before the first apnea were missing for one participant; therefore, analyses for apnea 1 were conducted with data from 20 participants, as indicated in the results.

Results

All participants completed the apnea protocol as intended, except four participants, who were unable to reach the full 3-min duration during the third apnea. These divers were included in the analysis, resulting in an average duration of 174 ± 13 s for A3.

Respiratory values

$P_{ET}CO_2$ was lower before the last apnea (A3) than before the first apnea (A1, p = 0.034, $\eta_p^2 = 0.16$, 90% CI [0.01–0.31], and $\eta_G^2 = 0.03$; Table 1). The RR was lower in A2 than in A1 (p = 0.034, $\eta_p^2 = 0.20$, 90% CI [0.02–0.34], and $\eta_G^2 = 0.05$; Table 1).

Respiratory signal

The signal from the device was clear, and RR and amplitude were easily detectable (Figure 3). The amplitude was larger in A2 and A3 than in A1 (p < 0.001, $\eta_p^2 = 0.36$, 90% CI [0.12–0.52], and $\eta_G^2 = 0.05$; Table 1). The product of the amplitude and the respiratory rate (eMV) was higher in A3 than in A1 (p < 0.001, $\eta_p^2 = 0.35$, 90% CI [0.14–0.49], and $\eta_G^2 = 0.04$; Table 1).

Correlation analysis of estimated tidal volume with $P_{ET}CO2$

For A1, the correlation did not reach significance ($r_s = -0.371$ and p = 0.054, Figure 4A), while there was a moderate negative correlation for A2 ($r_s = -0.500$ and p = 0.010; Figure 4B) and for A3 ($r_s = -0.512$ and p = 0.009; Figure 4C).

$P_{ET}CO_2$ prediction

A linear regression analysis revealed a significant predictive relationship between eMv and $P_{ET}CO_2$ in A1 (F (1, 18) = 6.629 and p = 0.019; Figure 4A), A2 (F (1, 19) = 13.994 and p = 0.001; Figure 4B), and A3 (F (1, 19) = 9.600 and p = 0.006; Figure 4C). eMv accounted for 27% of the explained variability in A1, 42% in A2, and 34% in A3. The linear regression equation for every apnea (Figure 4) was used to calculate the predicted $P_{ET}CO_2$ (pred$P_{ET}CO_2$) before the three apneas (Table 1). The difference between $P_{ET}CO_2$ and pred$P_{ET}CO_2$ was −0.00 ± 4.5 mmHg for A1, −1.15 ± 5.0 mmHg for A2, and −0.00 ± 5.2 mmHg for A3 (Figure 5).

Underwater respiratory signal

The quality of the respiratory signal recorded in water before starting apneic performance was satisfactory (Figure 6).

Discussion

Our results indicate that hyperventilation before breath-holding may be estimated using the signal from the force sensor. However, our method underestimates $P_{ET}CO_2$ values when mean values exceed 35 mmHg. Measurements appear more reliable when $P_{ET}CO_2$ is in the hypocapnic range of 25–35 mmHg, with reduced accuracy as $P_{ET}CO_2$ approaches normocapnia. These findings suggest that the prediction is more appropriate for mild hypocapnia but may be less reliable during normocapnia.

The successful application of the device for underwater performance, with good signal quality, is promising for future development. Although swimming motions, arm and leg movements, and chest compression at depth could affect the quality of the signal, the force sensor has the potential to identify involuntary breathing movements that signal the physiological breaking point (Agostoni, 1963).

Hyperventilation

We also found that freedivers hyperventilate without noticing as they keep RR within normal or even in the lower range of normal values, which explains the previous observations in our group (unpublished work). The hyperventilation is, thus, solely due to

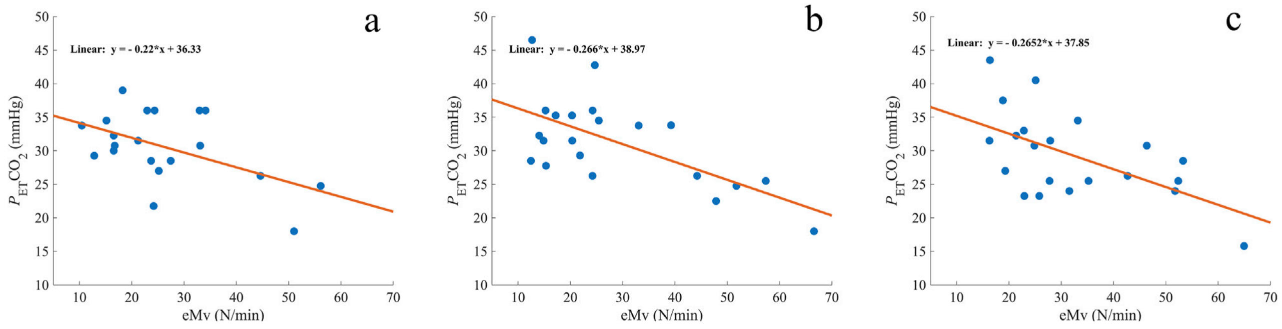


FIGURE 4
Comparison of measured end-tidal CO$_2$ ($P_{ET}CO_2$) with the estimated minute ventilation (eMV) before A1 **(A)**, A2 **(B)**, and A3 **(C)**. The orange line represents the regression line, and the corresponding formula is expressed on each graph. mmHg, millimeters of mercury; *n* = 20 **(A)** and 21 **(B, C)**.

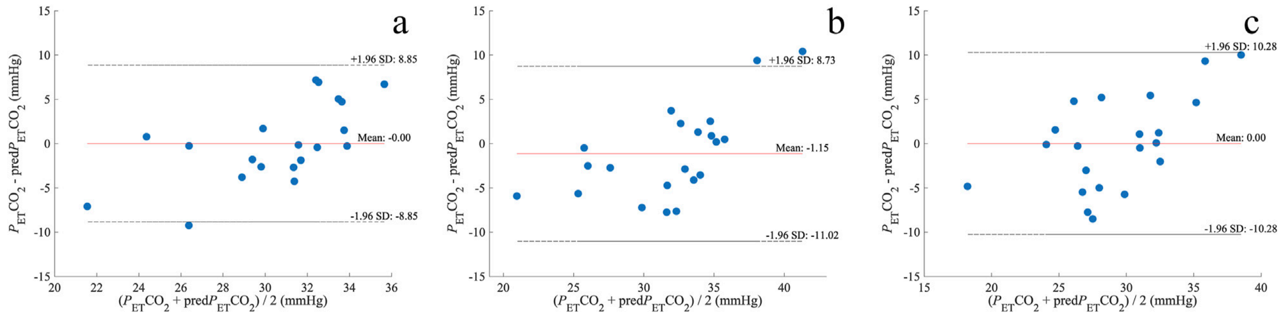


FIGURE 5
Bland–Altman plots of the difference between $P_{ET}CO_2$ and pred$P_{ET}CO_2$ before A1 **(A)**, A2 **(B)**, and A3 **(C)**. The dotted lines represent the upper limit of agreement (mean + 1.96 SD) and lower limit of agreement (mean − 1.96 SD); mmHg, millimeters of mercury; *n* = 20 **(A)** and 21 **(B, C)**.

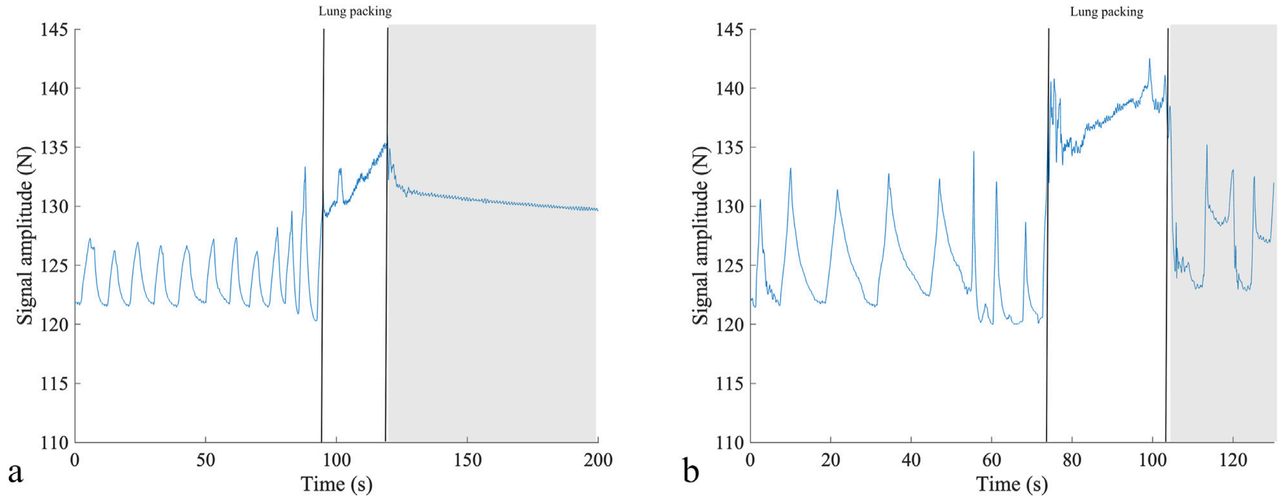


FIGURE 6
Respiratory signal from two participants before competition performance in static apnea **(A)** and dynamic apnea without fins **(B)**, depicting the differences in the breathing pattern as the diver in **(A)** shows shallower breaths but at an increased breathing frequency compared to diver **(B)**. The two vertical lines show the period of lung packing, and the gray rectangle shows the beginning of the apneic performance. N, newton; s, seconds.

increasing Vt. This emphasizes the challenge of quantifying the depth of breathing, a parameter that is less easily observed than RR both by the diver and observer. Quantifying the extent of hyperventilation is critical as severe hypocapnia correlates with reduced cerebral blood flow. In healthy participants, a 31% decrease in cerebral blood flow at a $PaCO_2$ level of 26 ± 2 mmHg has been reported (Fortune et al., 1995). Even moderate hyperventilation can cause a 20% reduction in brain blood flow (Reivich, 1964). Additionally, patients with brain hypertension also demonstrated up to a 34% increase in brain tissue hypoxia when $P_{ET}CO_2$ values fell below 25 mmHg (Carrera et al., 2010). Some participants in our study, particularly before the third apnea, experienced severe hypocapnia ($PE_TCO_2 \leq 25$ mmHg), and divers who initiate a dive with severe hypocapnia could be at a higher risk of BO. Hyperventilation alone could be a contributing factor to a transient loss of consciousness (Immink et al., 2014). At present, the relationship between the severity of hypocapnia and BO remains unclear.

Estimating lung volumes

During quiet breathing, Vt is mainly determined by the diaphragm contraction, which induces small changes in the vertical volume of the lung (West and Luks, 2021). As the force sensor detects changes in the circumference of the thorax, it is expected to be less sensitive at lower Vt, such as in A1. However, hyperventilation typically entails a more pronounced movement of the diaphragm and accessory muscles. Consequently, this amplifies the thoracic diameter, thereby enhancing the potential to detect an increase in chest circumference using the force sensor. In our study, the estimated Vt constituted nearly 58% of the VC. We acknowledge the limitation at low volumes, which explains why we cannot estimate VC or Vt with 100% accuracy and why the correlation with $P_{ET}CO_2$ was not significant in A1. This limitation applies to all the techniques used to estimate Vt from wearables as estimating it based on the movements of the chest wall is challenging (Monaco and Stefanini, 2021). As our intention is not to use it in a clinical setting but to monitor athletes for high respiratory activity, we consider that our results are suitable for exploring practical applications in different freediving situations, including saltwater and depth. This study acts as a proof-of-concept for applying breath analysis to estimate $P_{ET}CO_2$ levels during various underwater performances.

Additionally, the respiratory signal proved instrumental in detecting thoracic changes associated with "lung packing"—a maneuver employed by freedivers to enhance their total lung capacity (Örnhagen et al., 1998; Figure 6). This maneuver was initially described as glossopharyngeal breathing in post-polio patients (Dail et al., 1955).

Limitations

Our results apply only to dry static apneas in the sitting position, so the device should be further tested in underwater scenarios.

Additionally, despite most of the measurements being within the limits of agreement, there was a tendency to underpredict $P_{ET}CO_2$ when it was close to normal values. This means that during normal ventilation, the changes in the thoracic circumference were small and did not exert enough force in the sensor, so the amplitude of the signal was lower than expected. As we measured the changes in chest circumference in only one place, we could have missed information when ventilation was shallow or was only affecting the upper part of the chest.

Conclusion

This study demonstrates the potential of using a force sensor to estimate hyperventilation before breath-holding under static conditions, providing a foundation for further exploration. While the prediction model accounts for a moderate proportion of the variability in $P_{ET}CO_2$, additional validation is required to establish its utility in preventive applications. Freedivers may hyperventilate even at seemingly regular or reduced breathing frequencies, emphasizing the importance of refining this approach. Further research, including underwater assessments, is essential to evaluate the feasibility of this system for improving safety in freediving.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Swedish Ethical Review Authority. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

FP: conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing–original draft, and writing–review and editing. EM: conceptualization, validation, writing–original draft, and writing–review and editing. FJ: investigation, methodology, and writing–review and editing. AS: investigation, methodology, writing–original draft, and writing–review and editing. RB: investigation, methodology, validation, writing–original draft, and writing–review and editing. ML: methodology, writing–original draft, and writing–review and editing. ES: conceptualization, formal analysis, funding acquisition, project administration, resources, supervision, writing–original draft, and writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Funding was obtained through a donation from the Francis family in memory of their son/brother, who drowned from hypoxic blackout while snorkeling and holding his breath to dive underwater, and by a grant from the Swedish Research Council for Sport Science (CIF) Funding number P2019-0200.

Acknowledgments

The authors express gratitude to all participating freedivers and the competition organizers. Special recognition is given to Piero Giobbi, whose absence is profoundly felt. The authors also extend their gratitude to Valdemar Karlsson for granting permission to conduct tests during the competitions and offering invaluable assistance in organizing the field laboratory.

Conflict of interest

Author AS is the CEO of Oxygen Scientific GmbH. Author RB is the owner of Sensing Systems Corporation.

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2024.1498399/full#supplementary-material

References

Agostoni, E. (1963). Diaphragm activity during breath holding: factors related to its onset. *J. Appl. Physiol.* 18, 30–36. doi:10.1152/jappl.1963.18.1.30

Bain, A. R., Ainslie, P. N., Barak, O. F., Hoiland, R. L., Drvis, I., Mijacika, T., et al. (2017). Hypercapnia is essential to reduce the cerebral oxidative metabolism during extreme apnea in humans. *J. Cereb. Blood Flow Metabolism* 37, 3231–3242. doi:10.1177/0271678X16686093

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behav. Res. Methods* 37, 379–384. doi:10.3758/BF03192707

Bland, J. M., and Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 327, 307–310. doi:10.1016/s0140-6736(86)90837-8

Bosco, G., Rizzato, A., Martani, L., Schiavo, S., Talamonti, E., Garetto, G., et al. (2018). Arterial blood gas analysis in breath-hold divers at depth. *Front. Physiol.* 9, 1558. doi:10.3389/fphys.2018.01558

Boyd, C., Levy, A., McProud, T., Huang, L., Raneses, E., Olson, C., et al. (2015). Fatal and nonfatal drowning outcomes related to dangerous underwater breath-holding behaviors - New York State, 1988-2011. *MMWR Morb. Mortal. Wkly. Rep.* 64, 518–521.

Carrera, E., Schmidt, J. M., Fernandez, L., Kurtz, P., Merkow, M., Stuart, M., et al. (2010). Spontaneous hyperventilation and brain tissue hypoxia in patients with severe brain injury. *J. Neurol. Neurosurg. Psychiatry* 81, 793–797. doi:10.1136/jnnp.2009.174425

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2 edition. New York: Lawrence Erlbaum Associates Publishers.

Craig, A. B., Jr. (1976). Summary of 58 cases of loss of consciousness during underwater swimming and diving. *Med. Sci. Sports* 8, 171–175. doi:10.1249/00005768-197600830-00007

Craig, A. B., Jr (1961). Causes of loss of consciousness during underwater swimming. *J. Appl. Physiol.* 16, 583–586. doi:10.1152/jappl.1961.16.4.583

Dail, C. W., Affeldt, J. E., and Collier, C. R. (1955). Clinical aspects of glossopharyngeal breathing; report of use by one hundred postpoliomyelitic patients. *J. Am. Med. Assoc.* 158, 445–449. doi:10.1001/JAMA.1955.02960060003002

Dunne, C. L., Madill, J., Peden, A. E., Valesco, B., Lippmann, J., Szpilman, D., et al. (2021). An underappreciated cause of ocean-related fatalities: a systematic review on the epidemiology, risk factors, and treatment of snorkelling-related drowning. *Resusc. Plus* 6, 100103. doi:10.1016/j.resplu.2021.100103

Edmonds, C. W., and Walker, D. G. (1999). Snorkelling deaths in Australia, 1987-1996. *Med. J. Aus* 171, 591–594. doi:10.5694/j.1326-5377.1999.tb123809.x

Ferretti, G., Costa, M., Ferrigno, M., Grassi, B., Marconi, C., Lundgren, C. E. G., et al. (1991). Alveolar gas composition and exchange during deep breath-hold diving and dry breath holds in elite divers. *J. Appl. Physiol.* 70, 794–802. doi:10.1152/jappl.1991.70.2.794

Forster, H. V., Haouzi, P., and Dempsey, J. A. (2012). Control of breathing during exercise. *Compr. Physiol.* 2, 743–777. doi:10.1002/cphy.c100045

Fortune, J. B., Feustel, P. J., DeLuna, C., Graca, L., Hasselbarth, J., Kupinski, A. M., et al. (1995). Cerebral blood flow and blood volume in response to O2 and CO2 changes in normal humans. *J. Trauma* 39, 463–472. doi:10.1097/00005373-199509000-00012

Hill, P. M. N. (1973). Hyperventilation, breath holding and alveolar oxygen tensions at the breaking point. *Respir. Physiol.* 19, 201–209. doi:10.1016/0034-5687(73)90078-9

Immink, R. V., Pott, F. C., Secher, N. H., and Van Lieshout, J. J. (2014). Hyperventilation, cerebral perfusion, and syncope. *J. Appl. Physiol.* 116, 844–851. doi:10.1152/japplphysiol.00637.2013

İşat, G., Öztürk, T. C., Onur, Ö. E., Özdemir, S., Akoğlu, E. Ü., Akyıl, F. T., et al. (2023). Comparison of the arterial PaCO2 values and ETCO2 values measured with sidestream capnography in patients with a prediagnosis of COPD exacerbation. *Avicenna J. Med.* 13, 182–186. doi:10.1055/S-0043-1771179

Kumar, K. R., and Ng, K. (2010). Don't hold your breath: anoxic convulsions from coupled hyperventilation-underwater breath-holding. *Med. J. Aus* 192, 663–664. doi:10.5694/j.1326-5377.2010.tb03673.x

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front. Psychol.* 4, 863. doi:10.3389/fpsyg.2013.00863

Lin, Y. C., Lally, A., Moore, T. O., and Hong, S. K. (1974). Physiological and conventional breath hold breaking points. *J. Appl. Physiol.* 37, 291–296. doi:10.1152/jappl.1974.37.3.291

Lindholm, P., and Gennser, M. (2005). Aggravated hypoxia during breath-holds after prolonged exercise. *Eur. J. Appl. Physiol.* 93, 701–707. doi:10.1007/s00421-004-1242-y

Lindholm, P., and Lundgren, C. E. G. (2006). Alveolar gas composition before and after maximal breath-holds in competitive divers. *Undersea Hyperb. Med.* 33, 463–467.

Lippmann, J. (2019). Snorkelling and breath-hold diving fatalities in Australia, 2001 to 2013. Demographics, characteristics and chain of events. *Diving Hyperb. Med.* 49, 192–203. doi:10.28920/dhm49.3.192-203

Lippmann, J. M., and Pearn, J. H. (2012). Snorkelling-related deaths in Australia, 1994-2006. *Med. J. Aus* 197, 230–232. doi:10.5694/mja11.10988

Molchanova, N., Rybakov, V., Kalinin, E., and Kamenshchikova, A. (2020). Features of gas exchange in advanced level freediver during monofin practice in pool. *SportRxiv*. doi:10.31236/OSF.IO/8JNBR

Monaco, V., and Stefanini, C. (2021). Assessing the tidal volume through wearables: a scoping review. *Sensors (Basel)* 21, 4124. doi:10.3390/S21124124

Mulder, E., Schagatay, E., and Sieber, A. (2021). First evaluation of a newly constructed underwater pulse oximeter for use in breath-holding activities. *Front. Physiol.* 12, 649674. doi:10.3389/fphys.2021.649674

Muth, C. M., Radermacher, P., Pittner, A., Steinacker, J., Schabana, R., Hamich, S., et al. (2003). Arterial blood gases during diving in elite apnea divers. *Int. J. Sports Med.* 24, 104–107. doi:10.1055/s-2003-38401

Örnhagen, H., Schagatay, E., Andersson, J., Bergsten, E., Gustafsson, P., and Sandström, S. (1998). "Mechanisms of "buccal pumping" ("lung packing") and its pulmonary effects," in *24th annual scientific meeting, European underwater baromedical society*. Editor M. Gennser (Stockholm, Sweden), 80–83.

Panahi, A., Hassanzadeh, A., and Moulavi, A. (2020). Design of a low cost, double triangle, piezoelectric sensor for respiratory monitoring applications. *Sens. Biosensing Res.* 30, 100378. doi:10.1016/J.SBSR.2020.100378

Pernett, F., Bergenhed, P., Holmström, P., Mulder, E., and Schagatay, E. (2023). Effects of hyperventilation on oxygenation, apnea breaking points, diving response, and spleen contraction during serial static apneas. *Eur. J. Appl. Physiol.* 123, 1809–1824. doi:10.1007/s00421-023-05202-7

Qvist, J., Hurford, W. E., Yang, S. P., Radermacher, P., Falke, K. J., Ahn, Do W., et al. (1993). Arterial blood gas tensions during breath-hold diving in the Korean ama. *J. Appl. Physiol.* 75, 285–293. doi:10.1152/jappl.1993.75.1.285

Raichle, M. E., and Plum, F. (1972). Hyperventilation and cerebral blood flow. *Stroke* 3, 566–575. doi:10.1161/01.str.3.5.566

Reivich, M. (1964). Arterial Pco2 and cerebral hemodynamics. *Am. J. Physiol.* 206, 25–35. doi:10.1152/AJPLEGACY.1964.206.1.25

Schagatay, E., and Holm, B. (1996). Effects of water and ambient air temperatures on human diving bradycardia. *Eur. J. Appl. Physiol.* 73, 1–6. doi:10.1007/BF00262802

Schneeberger, J., Murray, W. B., Mouton, W. L., and Stewart, R. I. (1986). Breath holding in divers and non-divers-a reappraisal. *S Afr. Med. J.* 69, 822–824.

Scott, T., van Waart, H., Vrijdag, X. C., Mullins, D., Mesley, P., and Mitchell, S. J. (2021). Arterial blood gas measurements during deep open-water breath-hold dives. *J. Appl. Physiol.* 130, 1490–1495. doi:10.1152/japplphysiol.00111.2021

Sieber, A., Mulder, E., Pernett, F., Sehic, I., and Schagatay, E. (2022). "Novel sensor for assessment of involuntary breathing movements (respiratory contractions) during breath-hold diving [poster presentation]," in *46th annual European underwater and baromedical society conference*. (Prague).

Tamashiro, S., Nakayama, I., Gibo, K., and Izawa, J. (2023). Comparison of mainstream end tidal carbon dioxide on Y-piece side versus patient side of heat and moisture exchanger filters in critically ill adult patients: a prospective observational study. *J. Clin. Monit. Comput.* 37, 399–407. doi:10.1007/S10877-022-00901-6/

West, J. B., and Luks, A. M. (2021). in *West's respiratory physiology: the essentials*. Editors J. B. West and A. M. Luks 11th Edn (Philadelphia: Wolter Kluwer).

Wu, C. H., Chou, H. C., Hsieh, W. S., Chen, W. K., Huang, P. Y., and Tsao, P. N. (2003). Good estimation of arterial carbon dioxide by end-tidal carbon dioxide monitoring in the neonatal intensive care unit. *Pediatr. Pulmonol.* 35, 292–295. doi:10.1002/PPUL.10260

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 09 January 2025
DOI 10.3389/fpubh.2024.1426168

OPEN ACCESS

EDITED BY
Mohammad Yavarimanesh,
University of San Diego, United States

REVIEWED BY
Colin K. Drummond,
Case Western Reserve University,
United States
Abhishek Chakraborty,
University of Calcutta, India

*CORRESPONDENCE
Mohammed Alhameed
✉ malhameed@jazanu.edu.sa

RECEIVED 30 April 2024
ACCEPTED 25 November 2024
PUBLISHED 09 January 2025

CITATION
Kishor Kumar Reddy C, Kaza VS, Madana Mohana R, Alhameed M, Jeribi F, Alam S and Shuaib M (2025) Detecting anomalies in smart wearables for hypertension: a deep learning mechanism.
Front. Public Health 12:1426168.
doi: 10.3389/fpubh.2024.1426168

Detecting anomalies in smart wearables for hypertension: a deep learning mechanism

C. Kishor Kumar Reddy[1], Vijaya Sindhoori Kaza[1], R. Madana Mohana[2], Mohammed Alhameed[3]*, Fathe Jeribi[3], Shadab Alam[3] and Mohammed Shuaib[3]

[1]Stanley College of Engineering and Technology for Women, Hyderabad, India, [2]Department of Artificial Intelligence and Data Science, Chaithanya Bharathi Institute of Technology, Hyderabad, Telangana, India, [3]Department of Computer Science, College of Engineering and Computer Science, Jazan University, Jazan, Saudi Arabia

Introduction: The growing demand for real-time, affordable, and accessible healthcare has underscored the need for advanced technologies that can provide timely health monitoring. One such area is predicting arterial blood pressure (BP) using non-invasive methods, which is crucial for managing cardiovascular diseases. This research aims to address the limitations of current healthcare systems, particularly in remote areas, by leveraging deep learning techniques in Smart Health Monitoring (SHM).

Methods: This paper introduces a novel neural network architecture, ResNet-LSTM, to predict BP from physiological signals such as electrocardiogram (ECG) and photoplethysmogram (PPG). The combination of ResNet's feature extraction capabilities and LSTM's sequential data processing offers improved prediction accuracy. Comprehensive error analysis was conducted, and the model was validated using Leave-One-Out (LOO) cross-validation and an additional dataset.

Results: The ResNet-LSTM model showed superior performance, particularly with PPG data, achieving a mean absolute error (MAE) of 6.2 mmHg and a root mean square error (RMSE) of 8.9 mmHg for BP prediction. Despite the higher computational cost (~4,375 FLOPs), the improved accuracy and generalization across datasets demonstrate the model's robustness and suitability for continuous BP monitoring.

Discussion: The results confirm the potential of integrating ResNet-LSTM into SHM for accurate and non-invasive BP prediction. This approach also highlights the need for accurate anomaly detection in continuous monitoring systems, especially for wearable devices. Future work will focus on enhancing cloud-based infrastructures for real-time analysis and refining anomaly detection models to improve patient outcomes.

KEYWORDS

deep learning, machine learning, smart health monitoring, smart wearables, hypertension

1 Introduction

1.1 Smart health monitoring

One of the most significant developments in the healthcare industry in the current digital age is smart health care. Traditional medicine based on bioengineering has started to gradually digitalize information due to scientific theory and technological advancements. The healthcare system continuously monitors a patient by examining a variety of data and extrapolating a positive outcome from previous instances of such continuous monitoring. In Intensive Care Units (ICUs), continuous monitoring of patients is standard practice, allowing healthcare providers to access critical information in real-time. This monitoring can be lifesaving for conditions such as diabetes, asthma attacks, heart failure, and hypertension. Smart medical devices can connect to smartphones, enabling the seamless transmission of important patient data to clinicians. These gadgets also record data on blood pressure, weight, blood sugar, and oxygen levels. Smart health care makes it possible for people from a variety of backgrounds (such as doctors, nurses, caregivers for older family members, and patients) to find suitable information and results, appropriate information, and solutions, reduce medical errors, improve care, and reduce expenses at the right time in the health-care department/facilities (1). Several methods are used in smart health care, together with the usage of devices such as mobiles, computers, and televisions, along with various networks, like wide area networks (WANs), local area networks (LANs), and body area networks (BANs). The parameters that are most frequently tracked include blood heat, heart rate, blood pressure, and motion detection.

This research focuses specifically on arterial blood pressure (BP) monitoring, which plays a critical role in managing conditions like hypertension and cardiovascular diseases. The continuous monitoring provided by smart devices enhances real-time assessments, especially in individuals who may lack awareness of their vital signs or have varying levels of clinical knowledge (2). Smart medical devices, such as wearables, help both patients and healthcare providers to access relevant information. These devices track important metrics like blood pressure, heart rate, and oxygen levels, thus offering more accessible and timely interventions, particularly for those in remote areas or with limited access to healthcare (3). SHM empowers diverse users, from patients to healthcare professionals, by reducing medical errors, improving patient care, and cutting healthcare costs by providing real-time, continuous data transmission through networks like LANs and BANs.

Table 1 presents a sample of commonly used wearable sensor technologies, focusing specifically on their applications related to arterial blood pressure (BP) monitoring and other cardiovascular assessments. While this table does not encompass the full breadth of wearable devices available, it highlights technologies particularly relevant to BP prediction and management, as well as related clinical applications such as arrhythmia detection and heart failure management.

1.2 Hypertension: conditions for detecting hypertension

Hypertension is a significant global health concern, affecting millions and contributing to a higher risk of cardiovascular diseases. Blood pressure (BP) is a dynamic physiological measure that fluctuates minute by minute, influenced by various environmental and physiological factors (4). Continuous monitoring of BP helps detect trends that might indicate early signs of hypertension or cardiovascular strain. Home BP monitoring is gaining prominence as it offers valuable insights into BP fluctuations throughout the day and night, potentially unveiling conditions like "white coat" hypertension or irregularities linked to stress. This monitoring in diverse contexts allows for more informed decisions in treatment and management, reducing risks such as heart disease or hypertension-induced mortality. Effective BP control is especially crucial for reducing cardiovascular risks, particularly in older adults. While BP measurement has been shown to be an effective predictor of outcomes in cardiovascular disease, a better understanding of BP levels and variability could enhance risk stratification. It may facilitate the detection of "white coat" hypertension and assess excessive BP responses to various stresses. Variations in BP levels between day and night can also provide important information regarding the cardiovascular system (5). Furthermore, the comorbidity of mental illnesses and hypertension is linked to a higher cardiovascular mortality than hypertension alone, as hypertensive patients are more prone to experience anxiety. Effective BP control can decrease the risk of cardiovascular disease (CVD) and mortality in older individuals. BP varies over both short and long periods, including days, months, quarters, or years (6). Home BP monitoring, highly advised as a supplement to standard BP measurement in recent hypertension guidelines, plays a major role in the management of hypertension.

1.3 What are medical anomalies and why are they different?

Medical anomalies refer to deviations from typical physiological patterns, which may indicate underlying conditions or pathologies. These deviations can be congenital (present at birth) or acquired over time. For instance, variations in blood pressure (BP) could point to cardiovascular disorders, while other anomalies might suggest arrhythmias or metabolic imbalances.

A critical distinction must be made between anomalies and artifacts in medical data. Anomalies reflect genuine physiological irregularities that could suggest disease or abnormal conditions. In contrast, artifacts are errors or distortions in the data—often resulting from sensor misreading or environmental factors—that do not represent real physiological conditions. For example, sudden fluctuations in BP readings could be caused by movement or sensor misalignment rather than an actual BP variation. The machine learning (ML) system presented in this paper focuses primarily on detecting anomalies—deviations in physiological signals such as BP that may indicate an abnormal health state. However, distinguishing between true anomalies and artifacts is also essential to ensure accuracy in diagnoses. This research integrates signal processing techniques within deep learning models to filter out artifacts and enhance the detection of clinically relevant anomalies. To handle medical imaging tasks like classification and segmentation, anomaly detection is one potential methodology that can make use of semi-supervised and unsupervised methods.

Figure 1 illustrates the essential phases of processing medical data using machine learning for anomaly detection. The figure outlines a step-by-step flow from data acquisition to prediction and diagnosis, highlighting how each phase is related to anomaly detection:

Abbreviations: BP, Blood Pressure; SHM, Smart health monitoring; PPG, Photoplethysmograms; ECG, Electrocardiograms; LSTM, Long Short-Term Memory; MAE, Mean absolute error; RMSE, Root Mean Squared Error; MoN, Model of normality; FLOPS, Floating-point operations per second.

TABLE 1 Summary of wearable sensor technologies and clinical applications.

Sensor technology	Device type	Measurements	Clinical applications
PPG	Smartwatch or Band	Heart Rate Variability (HRV), Heart Rate (HR), Blood pressure (BP) without a cuff, oxygen saturation (SaO2), Heart Rate, Sleep Stages, Pulse-based Rhythm Detection, and Stroke Volume	Prediction of arterial blood pressure; Evaluation of risk in both healthy and cardiovascularly ill individuals; screening for and treatment of hypertension; identification and diagnosis of arrhythmias; tracking of sleep; Management of heart failure
ECG	Smart Ring	both single- and multiple-lead ECGs, ongoing or only when necessary observation, interval assessments (such as QTc), detection of arrhythmias, Changes in electrolyte abnormalities	Prediction of arterial blood pressure; Evaluation of risk in both healthy and CVD individuals; screening for and treatment of hypertension; identification and diagnosis of arrhythmias; diagnosis of acute coronary syndrome; extended QTc diagnosis Management of heart failure
Accelerometer	Chest Strap	Steps taken, force of impact, speed, amount of idle time, and exercise	Monitoring physical activity; Assessing risk in both healthy and CVD-afflicted individuals; Cardiopulmonary telerehabilitation; management of heart failure
Barometer	Wristband	Stair count	Monitoring physical activity; Assessing risk in both healthy and CVD-afflicted individuals; Cardiopulmonary telerehabilitation; management of heart failure
GPS	Smart Clothing	Travel distance and burned calories	Monitoring physical activity; Assessing risk in both healthy and CVD-afflicted individuals; Cardiopulmonary telerehabilitation; management of heart failure
Biometric Sensors	Smart Earbuds	Constant Monitoring of Electrolytes and Blood Glucose	monitoring blood sugar levels continuously; managing heart failure
		Non-invasive electrolyte levels in saliva and sweat and state of hydration	
Biomechanical	Smart Shoes	Ballistocardiograms, Seismocardiograms, Dielectric sensors	Weight, body vibrations, lung fluid volume, stroke volume, and cardiac output

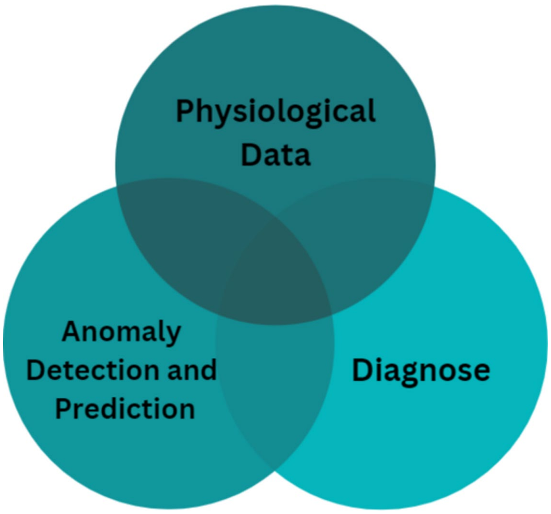


FIGURE 1
Key phases of processing medical data and their connection to anomaly detection.

a Prediction: Machine learning algorithms predict the future state of physiological signals such as BP, helping clinicians anticipate adverse events or trends (e.g., a gradual increase in BP).
b Diagnosis: By analyzing physiological signals, machine learning models can help identify pathological symptoms (e.g., hypertensive crises or arrhythmias).

These two tasks—prediction and diagnosis—are closely linked to anomaly detection since they enable identification of abnormal patterns in the data. The model extracts distinct features from the data, thereby improving diagnostic accuracy and delivering insights into patient health.

Key challenges in medical anomaly detection include:

a Test sensitivity: High sensitivity is required to detect subtle deviations accurately, ensuring that no abnormality is overlooked during diagnosis.
b Patient-specific factors: An effective model must account for individual differences in physiological baselines, ensuring that anomalies are detected based on personalized norms rather than generalized data.

Given these challenges, medical anomaly detection typically falls under supervised learning, where models are trained on labelled data (normal vs. abnormal) to identify anomalies. This contrasts with other domains, where anomaly detection is often an unsupervised task due to the absence of predefined labels (7).

1.4 Why use deep learning for medical anomalies?

Deep learning (DL) has emerged as a potent instrument in biomedical research because of its capacity to handle the intricate problems related to the identification of medical anomalies (8). The key advantages of deep learning, particularly in this context, include:

a Non-linearity modeling: Medical data is often non-linear, with complex relationships between variables. Deep learning models, such as the ResNet-LSTM, are capable of capturing these non-linear relationships, making it easier to distinguish between normal and abnormal physiological states.

b Handling data discrepancies: Medical data often contains inconsistencies or noise, whether due to artifacts or natural variability in patient signals. Deep learning models can manage these discrepancies by learning patterns from large datasets, thereby filtering out irrelevant variations and focusing on clinically significant changes.

In this paper, we leverage a ResNet-LSTM architecture for its ability to model both spatial and temporal features. This enables the model to uncover long-term dependencies in physiological data without the need for explicit feature engineering. Specifically, this approach helps identify BP anomalies by analyzing patterns in photoplethysmogram (PPG) and electrocardiogram (ECG) signals over time. The ResNet component effectively extracts spatial features from the data, while the LSTM component captures temporal relationships, enhancing the model's predictive power in detecting deviations. By applying this deep learning framework, the model is able to provide continuous, real-time monitoring of physiological signals, making it a robust tool for identifying true anomalies while minimizing the influence of artifacts. Figure 2 provides a hierarchical taxonomy of current deep learning techniques used in anomaly detection, illustrating how different models (including ResNet-LSTM) fit within the broader landscape of anomaly detection approaches.

1.5 Objective of paper

This paper's main objective is to use DL and ML techniques to investigate the transformative potential of SHM. It focuses specifically on the application of neural network architectures, ResNet-LSTM in particular, for the prediction of arterial blood pressure. The aim of this paper is to assess the efficacy of SHM in providing reliable, affordable, and timely health monitoring services, especially in remote areas. The study aims to contribute to the paradigm shift in health data assessment and anomaly detection by integrating intelligent sensors that can monitor health in real-time. It emphasizes the significance of continuous monitoring through wearables.

The first section of the paper introduces the problems that the genesis and transmission of diseases present to the healthcare sector, highlighting the need for creative solutions. The concept of SHM and its potential to transform the assessment of health data is then explored in depth. In the research methodology section, it is explained how physiological signals like PPG and ECG are used to predict arterial blood pressure using deep learning, specifically ResNet-LSTM. A thorough analysis of the ResNet-LSTM network's performance, including MAE and RMSE values, is provided in the results section. The network's accuracy across all BP prediction scenarios is demonstrated by numerical values. Interpreting the results, the discussion highlights the importance of accurate anomaly detection and wearables for continuous monitoring. Throughout, the research emphasizes the practical implications of the research in addressing current healthcare challenges and promoting personalized, effective health monitoring solutions.

Table 2 summarizes the types of anomaly detection techniques used in your paper, including deep learning, machine learning, statistical methods, and hybrid approaches. The precise method used,

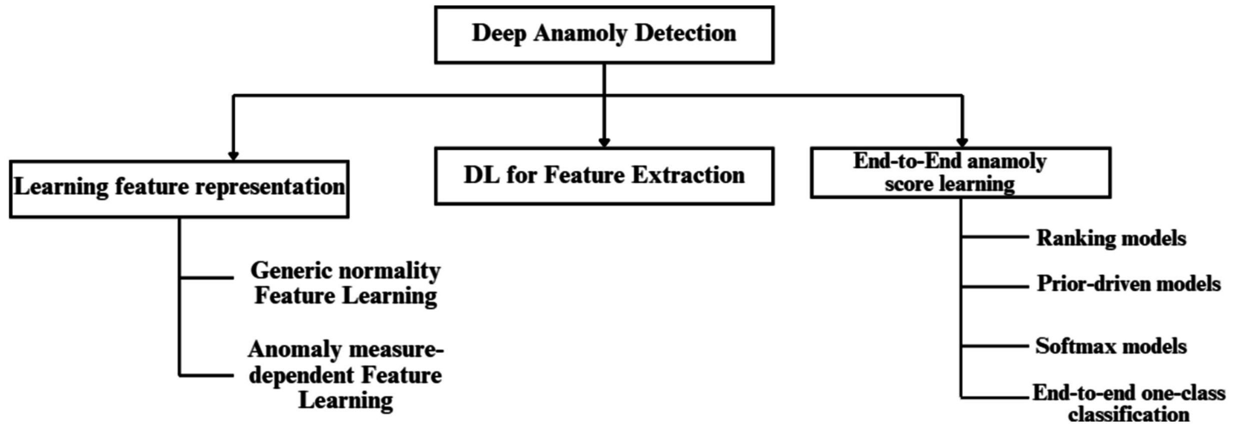

Deep Anamoly Detection
Learning feature representation
DL for Feature Extraction
End-to-End anamoly score learning
Generic normality Feature Learning
Anomaly measure-dependent Feature Learning
Ranking models
Prior-driven models
Softmax models
End-to-end one-class classification

FIGURE 2
A hierarchical taxonomy of current deep anomaly detection techniques.

TABLE 2 Summary of anomaly detection techniques.

Technique	Type	Data analysis	Online/ Offline	Reciprocity	Adaptability	Data processing
ResNet-LSTM	Deep Learning	Physiological Signals (ECG, PPG)	Online	Temporal	Non-adjustable	Central
WaveNet+LSTM	Machine Learning	Physiological Signals (ECG, PPG)	Offline	Temporal	Non-adjustable	Central
Clustering Algorithm	Statistical Method	Physiological Signals (ECG, PPG)	Offline	Spatial	Non-adjustable	Distributed
Transfer Learning	Hybrid	Image Features	Offline	-	Adjustable	Central

the kind of data analysis, reciprocity, online/offline nature, flexibility, and data processing strategy are given for each technique.

The current methods for detecting anomalies and sensor faults are briefly explained in the following section. The suggested method for detecting sensor anomalies is presented in Section 3. In Section 4, experiments and findings are covered along with a comparison of the suggested strategy with related approaches. Conclusion and potential future work are presented in Sections 5 and 6, respectively.

2 Literature survey

Many important factors, such as data processing algorithms, communication networks, sensor selection, contact-based versus contactless techniques, and other design considerations, must be carefully considered to create a dependable remote monitoring system. Numerous review papers offer perceptive evaluations of smart technology by examining it from multiple perspectives.

While Ohta et al. (37) focused on developing a health monitoring system especially for senior citizens who live alone to ease their anxieties and encourage independent living, while Tamura et al. (38) investigated the development of a home health monitoring system that did not forbid activities like bathing, sleeping, or urinating. These studies paved the way for the creation of intelligent wearables for health detection that add new features on a regular basis. Deep learning techniques are utilized by researchers for medical anomaly identification.

Clifford et al. (4) showcased the potential of computational methods in cardiology through their work on the categorization of heart sound recordings as normal or abnormal. In automated rehabilitation, Wang et al. (9) used deep back propagation–LSTM networks for upper-limbs EMG signal categorization. In the context of infectious diseases, Singh et al. (10) created a multi-objective differential evolution-based convolutional neural network for COVID-19 patient classification from chest CT images. Chang et al. (3) demonstrated the precise classification of genetic alterations in gliomas using deep-learning convolutional neural networks for applications other than healthcare.

Using image, audio, and inspection robot sensors, Hea et al. (39) investigated the connection between technology and infrastructure maintenance and developed a non-invasive method for fault diagnosis and detection in water distribution systems. Motwani et al. (1) provided a comprehensive analysis of machine learning-based ubiquitous and intelligent healthcare monitoring frameworks and provided insights into novel and developing treatments for patients with chronic illnesses. Several review articles covering a range of fields discussed anomaly detection. García-Macías and Ubertini (Springer) integrated SHM systems, focusing on data fusion and unsupervised learning to identify damage. Aliyu et al.'s paper, "Anomaly Detection in Wearable Location Trackers for Child Safety," focused on microprocessors and microsystems. Churová et al. proposed an anomaly detection method for real-world data (11).

The study carried out by Hamieh et al. (12) sheds light on a noteworthy and demanding application of remote monitoring: mental health. They highlight the value of using unsupervised learning to spot relapses in individuals with psychotic disorders, demonstrating the potential benefits of objective, non-intrusive monitoring for early intervention and improved patient outcomes. Further research into AI-powered mental health monitoring systems that safeguard user privacy and provide carers and clinicians with useful information is made possible by this study. Jahan et al. (13) introduce us to a new field by using smartwatch technology for activity recognition within the context of religious rites like salat. This study shows how adaptable remote monitoring can be, going beyond traditional applications in fitness and health to satisfy cultural and religious demands. Think about the benefits that come from tailoring activity detection algorithms to various practices so that individuals can meaningfully monitor their participation and adherence.

The application of remote monitoring in human behavior analysis is elaborated upon by Bozdog et al. (14). Their research demonstrates how anomalies can be detected and intricate human behavior patterns can be deciphered using wearable sensors and machine learning. This opens the door to applications like risk prediction, personalized coaching, and even environmental adaptation based on real-time behavioral data. Our primary concerns as we use these technologies to understand human behavior should be ethics and user privacy. The resource Kalpana et al. (40) was helpful as it gathered an extensive overview of deep learning methods for anomaly identification in human activity recognition, which helped in formulating the proposed model and understand the scope and need for this research work. This provides a comprehensive summary of current research trends and identifies areas that warrant additional research. By highlighting the benefits and drawbacks of various algorithms, this paper lays the foundation for researchers to build on current understanding and push the boundaries of human activity detection accuracy and interpretability. By adding to the body of knowledge, these combined efforts promote the development of remote monitoring systems and increase their efficacy in a range of applications. All the existing works have been summarized for a better understanding in Table 3.

3 Proposed methodologies

The algorithmic strategies for detecting medical anomalies are:

a *Unsupervised anomaly detection*: It does not involve any supervision signal that would indicate whether a sample is normal or not during the learning process. Unsupervised methods are therefore intriguing to the machine learning field since they do not require labelled datasets. The following subsections introduce two popular unsupervised deep anomaly architectures: Autoencoders (AEs) and Generative Adversarial Networks (GANs). AEs have been extensively used for automatic feature learning ever since they were first introduced as a pre-training method for deep neural networks. The model is trained to reconstruct the input using a learnt compressed representation that is stored at the core of the architecture because the AEs are symmetrical. Assume that the current input (I) is a dataset made up of samples, and that the encoder and decoder networks are denoted, respectively. Next, the compressed form is provided as follows: Formally, let us assume that the encoder and decoder networks are denoted, that the dataset comprises samples, and that the current input is p. Next, it is decided what the compressed representation using Equation 1.

$$l = h(i) \tag{1}$$

and the reconstruction is performed using Equation 2.

$$y = g.(l) \tag{2}$$

To minimize the reconstruction loss, $\mathrm{K, pK\,and\,L}(\mathrm{x}, \mathrm{g}(\mathrm{h}(\mathrm{x}))$ this model has been trained.

b *Supervised anomaly detection*: Due to its high demand in diagnostic application because of its high sensitivity and durability supervised learning is being applied widely for medical anomaly detection. It also has proven to be better performing than unsupervised methods. In This approach, a supervised signal is presented which indicates whether the samples are from the normal category or abnormal. Thus, making the job to behave as a binary classifier, and training the models using binary cross-entropy loss. Multi-task learning (MTL) which is a subtype of supervised learning, helps to transfer pertinent knowledge collected from various linked tasks, among them. For example, the difficulties brought on by subject-specific differences can be solved using a secondary subject identification task. As a result, the model develops the ability to classify anomalies while also learning to recognize similarities and differences among participants. The deep learning architectures that have been explored thus far are feedforward designs, meaning that data moves from input to output in a single direction. Their capacity to model temporal signals is hence constrained. Recurrent Neural Networks are used to overcome this restriction.

c *Recurrent neural networks (RNNs)*: Recurrence is a crucial characteristic for tasks like time-series modeling since it essentially means that the output of the current time step is once more used as an input to the subsequent time step. The modeling of sequential medical data, such as EEG and phonocardiographic data, is also essential for obtaining the temporal evolution of the signal. For modeling long-term dependencies, simple RNN architectures are ineffective due to BPTT-caused disappearing gradients. Several variations of RNN models have been created to mitigate this issue. However, because RNNs have a lot of vanishing gradients, they cannot accurately represent long-term dependencies; for this reason, LSTM networks are developed.

3.1 How do smart watches analyze heart rate?

The heart rate monitor of the smartwatch uses an easy and economic optical method PPG. It is employed to find changes in blood volume in the tissue's microvascular network. This technique uses a combination of green LED and infrared light along with photosensitive diodes to illuminate the skin and measure the absorption of the green light accordingly (15) as depicted in Figure 3.

TABLE 3 Summary and insights obtained from existing literature.

Reference	Methods	Techniques	Results	Problems identified
García-Macías and Ubertini (Springer)	Structural health monitoring (SHM) systems	Data fusion, unsupervised learning for damage identification	Incorporation of SHM systems, emphasis on data fusion and unsupervised learning	Customization of activity detection algorithms
Aliyu et al.	Wearable location trackers: detecting anomalies	Microprocessors, microsystems	Centered on wearable location trackers' anomaly detection for kid safety	Privacy concerns
Churová et al. (2020) (11)	Real-world data anomaly detection technique	Not specified	Proposed real-world data anomaly detection technique	Privacy concerns
Hamieh et al. (2023) (12)	Unsupervised learning for mental health monitoring	Not specified	Identification of relapses in people with psychotic disorders, potential for early intervention	Privacy concerns in mental health monitoring
Jahan et al. (2023) (13)	Smartwatch technology for activity recognition in religious rites	Not specified	Flexible remote monitoring beyond health and fitness, meeting cultural and religious requirements	Customization of activity detection algorithms
Bozdog et al. (2021) (14)	Remote monitoring in human behavior analysis	Wearable sensors, machine learning	Potential for understanding intricate patterns in human behavior, applications in risk prediction and coaching	Ethical and privacy concerns in human behavior analysis
Kalpana et al. (2022)	Deep learning methods for anomaly identification in human activity recognition	Not specified	Overview of current research trends, advantages, and disadvantages of various algorithms	Areas that show promise for further investigation

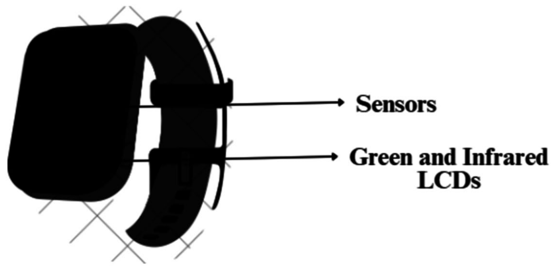


FIGURE 3
Photosensitive diodes (sensors), green LED and infrared lights on the base of a smartwatch.

Here, this combination of infrared light and green LED is taken because the Red Blood Cells (RBCs) reflect the red light and absorb the green light. This helps the sensors compute the amount of blood flowing through the wrist at any given time. The green LED is generally used when the user is performing any exercise. Normally, the watch uses infrared light to calculate the heart rate every 10 min. Furthermore, if the watch is loosely worn or the skin is perfused the LED increases its brightness and sampling rate to measure the exact heart rate. When the watch measures the heart rate every 10 min, it switches to the green LED in case the infrared light fails to provide an adequate reading (16). These lights flash hundreds of times per minute to get a hold of the blood flow which helps the device calculate the heart rate precisely.

3.2 Architecture of the anomaly detection model and its role in data processing of the MIMIC database

The MIMIC database (17), which contains a variety of data gathered from ICU patients, is used to calculate the relationship between PPG and ABP and assess how the model responds to abrupt variations in blood pressure. In this application, LSTM and CNN are combined. Since CNN can extract deep features and LSTM can learn from past experiences, these models are ideal for anomaly detection. Due to fully connected layers and connectionless nodes processing a single input between layers, the 2D CNN and LSTM model offers a superior classification (18). A temporal sequence is used as the input for an LSTM and is connected to the nodes from a directed graph along with a typical order.

3.2.1 Convolutional neural network (CNN)

CNN is used in many different applications, such as image classification, object recognition, and medical image analysis. CNN is mainly used to extract local characteristics from higher-level inputs. These characteristics are then forwarded to lower layers for help with more complex features. Its three layers are pooling, fully connected (FC), and convolutional (FC) (1).

A Convolutional layer:

A collection of kernels for generating a tensor of feature mappings is present in the convolutional layer of the CNN layer. The kernels use the striding process to entwine the entire input to produce the output volume's dimensions as integers, while the convolutional layer reduces the dimensions of the input volume. To retain the size of the input volume using low-level characteristics while padding an input volume of zeros, the striding procedure is required. The convolutional layer's function is described as in Equation 3.

$$G(x,y) = (M * N)(i,j) = \sum\sum M(x+i, y+j)\, N(i,j) \tag{3}$$

where G is the result of a 2D feature map, N is a 2D filter of size i × j, and M is the input matrix.

B Rectified linear unit (ReLU) layer:

The convolutional layer's operation is indicated by M*N. Feature maps can be made more nonlinear by using the ReLU layer. ReLU uses a threshold input of zero to calculate activation. The mathematical expression for it is as follows as given in Equation 4.

$$\mathrm{f}(\mathrm{x}) = \max(0,\mathrm{x}) \tag{4}$$

C Pooling layer:

The pooling layer performs a down sample of the specified input scale to minimize the number of factors. Max pooling is the most popular technique since it yields the highest result for a certain input region. Using the characteristics gathered from the previous two layers, the FC layer computes the judgments made by CNN. It serves as a classifier, the FC layer.

- Why use ResNet?

Res Nets help in preserving a low error rate in the deeper layers of the network hence, making them one of the most efficient Neural Network Architectures. It employs a method known as skip connections (12). The salient characteristic of this technique is that regularization bypasses any layer that reduces or impedes the architecture's performance. To form a residual block, this connection skips some layers between the activations of one layer and those of subsequent layers. These residual blocks are stacked together to create ResNets. leads to training the deep neural network without any vanishing or exploding gradient disruptions. This network fits the residual mapping by letting the network do the fitting. Hence, instead of saying G(x), initial mapping, let the network fit as defined by Equation 5.

$$H(i) := G(i) - H(i) := G(x) + \mathrm{i} \tag{5}$$

3.2.2 Long short – term memory (LSTM)

A particular kind of recurrent neural network called an LSTM solves disappearing and exploding gradient problems by using memory blocks rather than the standard RNN units. The LSTM's cell state also stores the long-term states, enabling it to link data gathered in the past and present. Three distinct types of gates make up the internal structure of the LSTM, as shown in Figure 4:

x_p- denotes the current input;

C_p and C_{p-1} – denote the new and previous cell states, respectively; and

h_p and h_{p-1} - denote the current and previous outputs, respectively.

$$i_p = \sigma\left(W_i \cdot \left[h_{p-1}, x_p\right] + b_i\right) \tag{6}$$

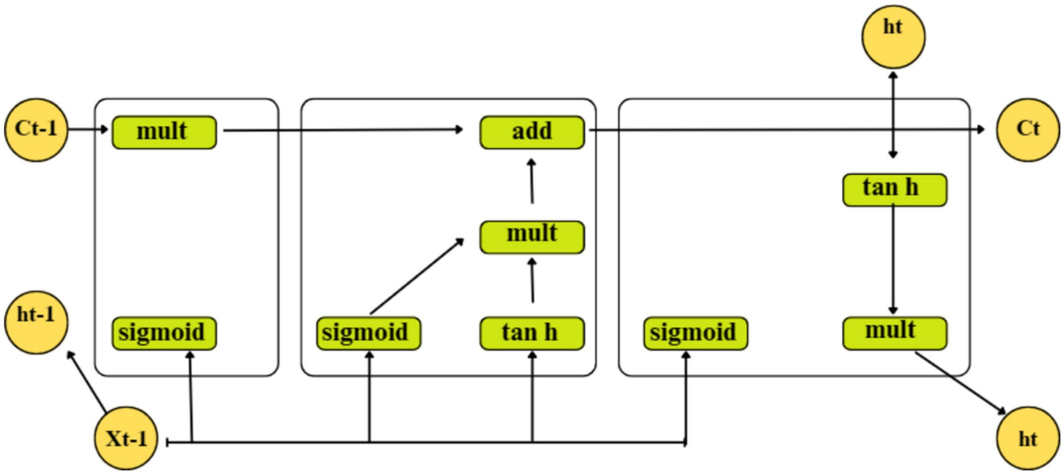


FIGURE 4
Architecture of LSTM input gate.

$$\tilde{C}_p = \tanh\left(W_i.\left[h_{p-1},x_p\right]+b_i\right) \tag{7}$$

$$C_p = f_p\, C_{p-1} + i_p\tilde{C}_p \tag{8}$$

$$f_p = \sigma\left(W_f.\left[h_{p-1},x_p\right]+b_f\right) \tag{9}$$

Where, $\tilde{C}_p$ represents the current moment information and refers to a tan h output,

C_{p-1} represents the long-term memory information,

W_i denotes a sigmoid output and the weighted matrices of the input gate,

b_i represents the LSTM bias of the input gate,

W_f represents the weight matrix,

b_f represents the offset, and

σ represents the sigmoid function.

Here, x_p and h_{p-1} are passed through a sigmoid layer using Equation 6. to determine which portion of the data needs to be added. Moreover, after x_p and h_{p-1} have passed through the tanh layer, Equation 7 is utilized to extract new information, and Equation 8 integrates long-term memory and current memory into C_p. The information from a previous cell that can be forgotten is determined using Equation 9.

$$O_p = \sigma\left(W_o.\left[h_{p-1},x_p\right]+b_o\right) \tag{10}$$

$$h_p = O_p \tanh\left(C_p\right) \tag{11}$$

Where, W_o represents the weighted matrices of the output gate, and.

b_o represents the LSTM bias of the output gate.

Using Equations 10, 11, the output gate determines the states necessary for x_p and h_{p-1} inputs to continue. To obtain the final output, the state decision vectors that transfer new information, Ct, across the tanh layer are located and multiplied.

3.2.3 Combined CNN-LSTM network

This design uses the LSTM as a classifier and the CNN to extract complex features from images (19). The suggested network has a total of 20 layers, as seen in Figure 5. These layers consist of an FC layer, an LSTM layer, five pooling layers, twelve convolutional layers, and an output layer that applies the SoftMax function. Two or three 2D CNNs, a pooling layer, and a dropout layer with a 25% dropout rate connect each convolutional block. The 3×3 sized kernel convolutional layer is activated by the ReLU function and prepared for feature extraction. The max-pooling layer's 2×2 size kernels are used to reduce the size of the input image. The LSTM layer uses the function map transferred in the last stage of the model to extract time information.

3.3 How the hybrid network identifies high-risk ABP conditions

There are two approaches to detect hypertension and monitor the blood pressure. The first method treats the model as though it were a regression task—that is, as though it produces continuous values. As a result, the systolic and diastolic values of blood pressure are calculated using the PPG and ECG signals (20). By using both signals or features computed from PPG and ECG (or, in some cases, only PPG signal is used) signals as input, various machine learning techniques, such as linear regression models and artificial neural networks (ANN) for regression tasks, are used to estimate BP values.

The second approach treats the model as if it generates discrete values or labels, i.e., like a classification task. In this method, the models try to compute the level of hypertension the patient belongs to, based on clinical and socio-demographic data (21). This approach differs from the first is that, the first approach utilizes raw signals or features extracted from the input data used in ML model, whereas the second approach makes use of continuous clinical data.

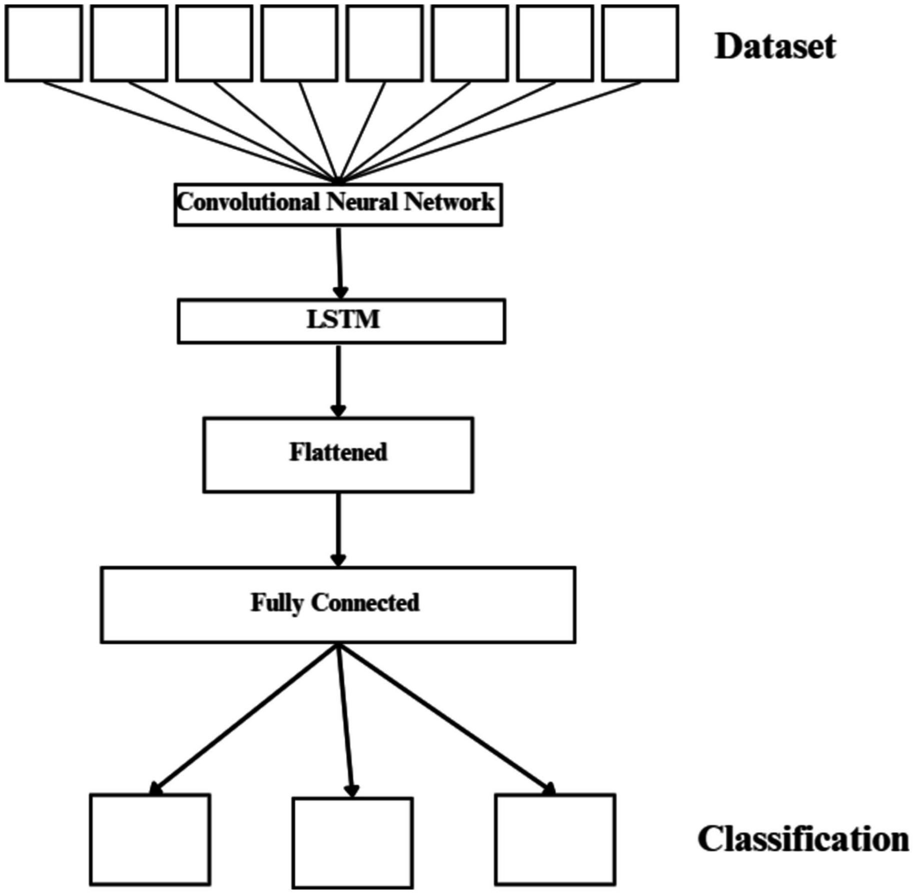


FIGURE 5
Proposed hybrid network.

(a) Regression task (first approach)

Linear regression formula is defined in Equations 12, 13:

$$y = mx + b \tag{12}$$

Where:

- y is the predicted blood pressure value.
- m is the slope.
- x is the input signal.
- b is the y-intercept.

(b) Classification task (second approach)

SoftMax function:

$$\text{SoftMax}(x)_i = \frac{e^{xi}}{\sum_j e^{xj}} \tag{13}$$

where:

- e is the base of the natural logarithm.
- x_i is the input value for class i.
- The function outputs a probability distribution over multiple class.

The following subsection provides insights on how to use these methods along with transfer learning in the hybrid network.

3.4 How to integrate transfer learning in the hybrid network

Figure 6 illustrates the flowchart for anomaly detection, which feeds N anomaly-free images into the deep feature extractor of the transfer learning model. The MoN, which learns/extracts normality from the input images, is created using the learnt or extracted features. Consequently, a transfer learning model is used to extract the features for a given input image. A similarity measure is then used to compare the extracted features to the MoN, and the anomaly is identified if the resulting anomaly score is greater than the decision threshold.

A Transfer learning model

We use EfficientNet, which was trained on the ImageNet dataset, for transfer learning. It employs a state-of-the-art scaling method that uniformly scales each dimension (depth, width, and resolution) using a compound scaling coefficient. The balanced scaling of the model leads to improved performance. The baseline network of EfficientNet, called "EfficientNet-B0," maximizes FLOPS and precision (22). Next, the baseline network was scaled with different compound coefficients to create the "EfficientNet-B1 through B7" EfficientNet scaled versions. Using a multi-objective Neural Architecture Search (NAS) that improves accuracy and FLOPS, the core network of EfficientNet was built.

B Model of normality (MoN)

The representations that do not fit its specification are marked as anomalies since MoN learns normality from the characteristics that are extracted. Therefore, all regular variants must be included in the MoN creation for the designated purposes. A MoN is constructed for each data class by averaging the learnt features taken from the N normal pictures, which are only used to create MoNs and are not included in the evaluation set.

C Similarity measure

The MoN's subjective similarity to an image can be expressed in terms of a distance measure specified on the learnt feature space, since each input image's deep-learned features function as a unique identifier. In order to accomplish this, we use Euclidean distance to calculate the similarity between MoN and features taken from the test photos.

Euclidean distance formula is given in Equation 14:

$$\text{Distance}\left(\text{MoN,Image}\right)=\sqrt{\sum_{i}\left(MoN_i - Image_i\right)^2} \tag{14}$$

Since it directly impacts detection efficiency, the decision threshold is a critical component of distance-based anomaly detection algorithms. Figure 7 shows a flowchart that illustrates the threshold-setting procedure. By adjusting this level appropriately, it is possible to significantly increase detection accuracy while reducing false positive rate. Here, based on the vectors K_{max} and K_{mean}, we suggest a clear way for setting the working-point threshold in Table 4.

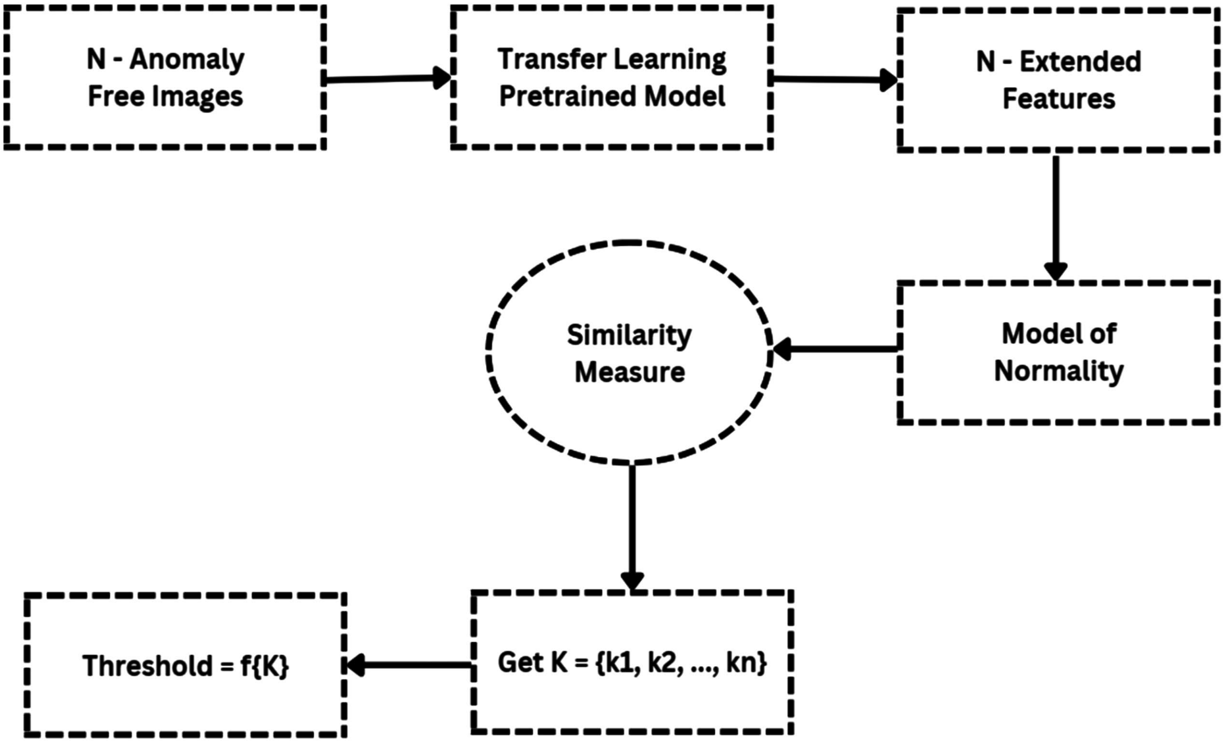


FIGURE 7
Working-point threshold setting process.

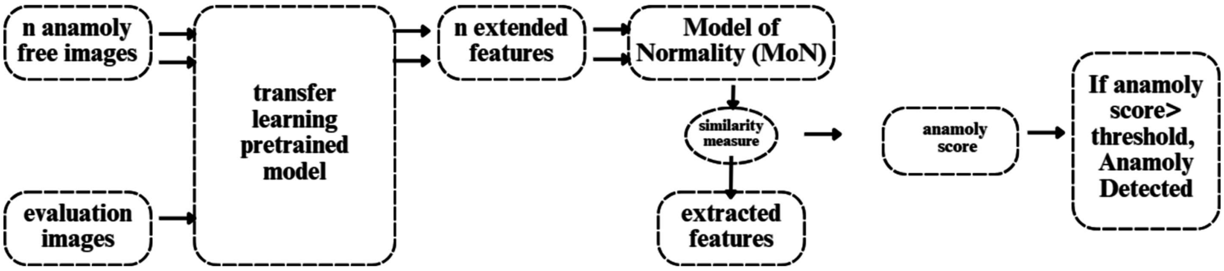


FIGURE 6
Anomaly detection.

TABLE 4 Working-point threshold based on the vectors K_{max} and K_{mean}.

Threshold (T)	K_1 - K_{max}, K_2 – K_{mean}
T_1	max(K_1)
T_2	max(K_1) - std.(K_1)
T_3	mean(K_1) + std.(K_1)
T_4	max(K_2)
T_5	max(K_2) - std.(K_2)
T_6	mean(K_2) + std.(K_2)

3.5 About dataset

The graph illustrated in Figure 8, shows the distribution of heart rate zones recorded during a specific date (2019-04-11) as part of SHM using wearable sensors, from the dataset (23). Each heart rate zone, categorized based on intensity levels such as "Out of Range," "Fat Burn," "Cardio," and "Peak," is represented by a bar in the graph. The height of each bar corresponds to the duration (in minutes) spent in the respective heart rate zone, while the color coding helps distinguish between different zones.

This graph is particularly relevant to our discussion on the use of physiological signals, such as ECG and PPG, in predicting arterial BP through neural network architectures (24, 25). In the context of ResNet-LSTM network's superior performance in health monitoring, this graph provides valuable insights into the distribution of heart rate zones, which are indicative of the intensity levels of physical activity or exertion (26–28). Understanding these heart rate patterns can contribute to the accurate prediction of BP and overall health monitoring (29). Furthermore, the graph aligns with the significance of anomaly detection and the need for accurate monitoring through wearables. By analyzing heart rate data and identifying anomalies or irregularities in heart rate patterns, healthcare professionals can intervene in a timely manner to address potential health concerns (30, 31). Overall, this graph serves as a visual representation of the physiological data collected through SHM, supporting the discussion on leveraging innovative technologies for real-time health monitoring and prediction (32–34).

The MIMIC (Medical Information Mart for Intensive Care) dataset is a critical resource in this research, providing a diverse collection of real-world clinical signals that support the development and evaluation of the proposed blood pressure (BP) prediction models. MIMIC contains rich physiological data, including Electrocardiogram (ECG), Photoplethysmography (PPG), and Arterial Blood Pressure (ABP) measurements from a wide variety of patients with different medical conditions (35, 36). This diversity allows models like the ResNet-LSTM to generalize across various patient profiles and medical scenarios, improving the accuracy and reliability of BP anomaly detection. In the study, the dataset was used to train and validate machine learning models designed to predict systolic and diastolic BP. The results, displayed in Tables 5, 6, show that the ResNet-LSTM model outperforms other architectures in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). This comprehensive dataset played a crucial role in the testing and validation of the proposed model, ensuring that the models were exposed to real-world complexities, thus enhancing their predictive power and applicability in clinical settings.

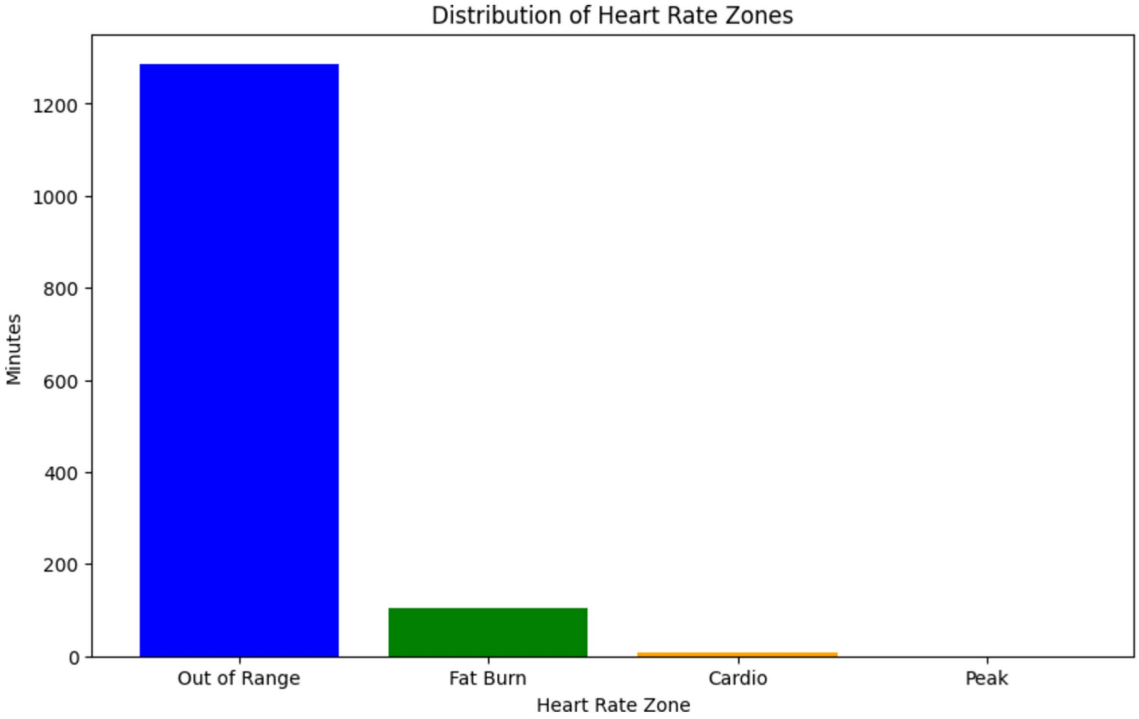


FIGURE 8
Distribution of heart rate zones as per our readings.

TABLE 5 Errors on the total BP prediction using the MIMIC database.

Tested set	MAE	RMSE	MAE	RMSE
Dataset	Photoplethysmography dataset		Photoplethysmography + Electroencephalogram Dataset	
Fully connected	18.6	27.2	18.3	25.7
Long Short-Term Memory	8.6	13.3	5.9	9.3
WaveNet	12.3	18.3	11.3	17.5
WaveNet + Long Short-Term Memory	10.0	15.6	5.7	9.0
ResNet + Long Short-Term Memory	6.2	8.9	3.3	5.0

TABLE 6 LOO outcomes using ResNet-LSTM on the MIMIC database.

Tested set	RMSE	MAE	MAE D	MAE S	RMSE D	RMSE S
Direct SBP/DBP prediction						
Photoplethysmography (50 pat)	-	-	10.746	23.598	12.344	27.643
Photoplethysmography (40 pat)	-	-	11.106	24.223	12.642	28.247
Electrocardiogram (40 pat)	-	-	9.548	20.367	10.848	23.070
Entire BP prediction						
Photoplethysmography (50 pat)	19.155	15.342	10.684	21.467	12.349	25.383
Photoplethysmography (40 pat)	19.560	15.679	10.818	22.409	12.411	26.246
Electrocardiogram (40 pat)	18.018	14.609	10.105	22.099	11.529	24.587

4 Results

The results of BP prediction achieved using the distinct settings and networks are summarized in Table 5. Performance was improved in each setup when PPG was used. The ResNet + LSTM network, which accurately predicted BP values, was the best one (Table 6). On the validation set, the network overall MAEs were considered. Since the networks are designed with the primary objective of generating BP values in mind, direct BP prediction appears to be the optimal strategy. But when networks need to infer the entire signal, they need to learn information that will not be used. Table 7 illustrates that the ResNet + LSTM is the optimal network in both scenarios and also illustrates the neural network's complexity for anomaly detection.

The errors in the total BP prediction for various setups using the MIMIC database (17) are displayed, as shown in Supplementary Figure S5. MAE and RMSE values are shown in the figure to illustrate how different configurations—Fully connected, LSTM, WaveNet, WaveNet+LSTM, and ResNet+LSTM—perform in terms of performance. As stated in the research of Paviglianiti et al. (5) titled "A Comparison of Deep Learning Techniques for Arterial Blood Pressure Prediction" published in Cognitive Computation, statistical comparisons with current models were conducted for more thorough examination. Using a custom dataset from the works of Paviglianiti et al. (5) and ECG, PPG, and ABP readings taken from the MIMIC database, our model is trained and tested. This collection of clinical signal data is a priceless resource that provides an accurate representation of physiological parameters observed in daily life. With the use of this vast and varied dataset, our models were able to learn and generalize across a broad range of medical conditions and patient profiles. We ensured that our models were exposed to the nuances and complexities found in actual patient data by utilizing clinical signals from the MIMIC database, which increased their predictive power.

TABLE 7 Comparison of complexity of various neural networks.

Neural network	Cost estimation (FLOPs)		Complexity order
	PPG + ECG	PPG	
Fully connected	~2,500	~625	$O(L \times (V_{dim})^2)$
Long Short-Term Memory	~2,500	~625	$O(L \times (V_{dim})^2)$
WaveNet	~7,500	~1850	$O(L \times (V_{dim})^2 \times k_{size})$
WaveNet + Long Short-Term Memory	~7,500	~1850	$O(L \times (V_{dim})^2 \times k_{size})$
ResNet + Long Short-Term Memory	~17,500	~4,375	$O(L \times (V_{dim})^2 \times k_{size})$

The MIMIC Database's Leave-One-Out (LOO) results are shown in Supplementary Figure S6, with an emphasis on the optimal neural network architecture, ResNet-LSTM. The figure shows whole BP prediction scenarios as well as MAE and RMSE values for direct systolic/diastolic blood pressure (SBP/DBP) prediction using PPG and ECG signals.

In addition, we validated our models' performance using the Pulse Transit Time PPG dataset. To ascertain whether our models could be used outside of the training set, this additional dataset was essential. Carefully comparing the results to this independent dataset demonstrated the robustness and reliability of our models. Consequently, we were able to assess our models' efficacy against state-of-the-art techniques and gain a better understanding of how well they predicted arterial blood pressure.

A thorough comparison of neural network complexities is shown in Figure 9. The graph shows the cost estimation and complexity order (FLOPs) for several neural network architectures, such as ResNet+LSTM, WaveNet, WaveNet+LSTM, and Fully connected. Understanding each architecture's computational efficiency is made easier with the help of this visual representation.

Figure 10's-line plot illustrates the trend of MAE and RMSE in several configurations, such as Fully Connected, LSTM, WaveNet, WaveNet+LSTM, and ResNet+LSTM. The x-axis represents the different setups, and the y-axis shows the error values. Plot

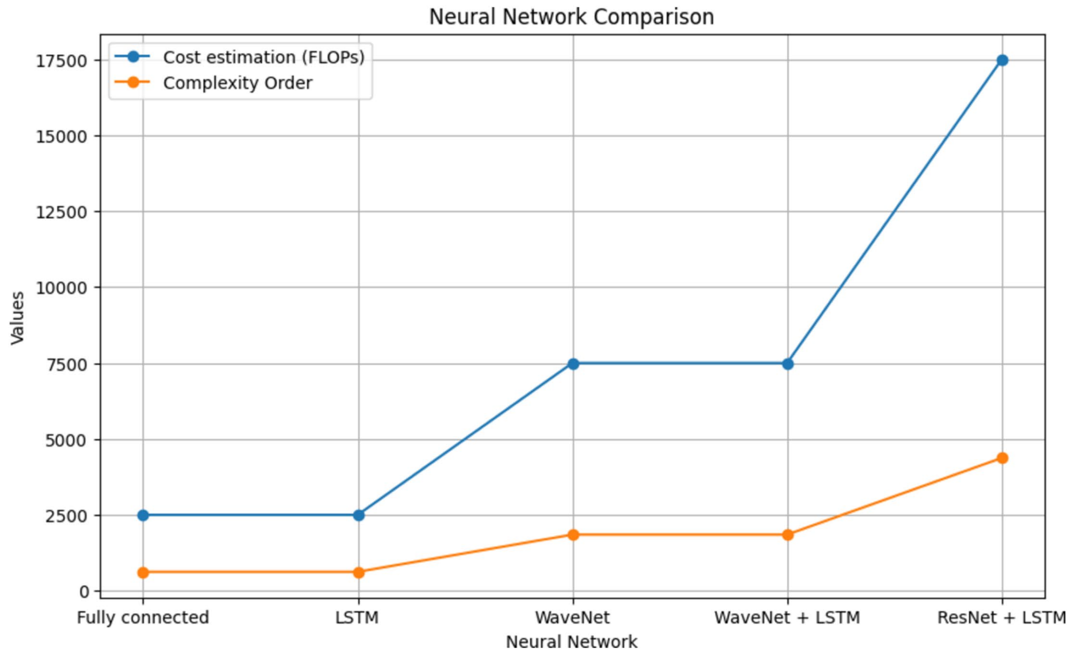


FIGURE 9
Neural network comparison.

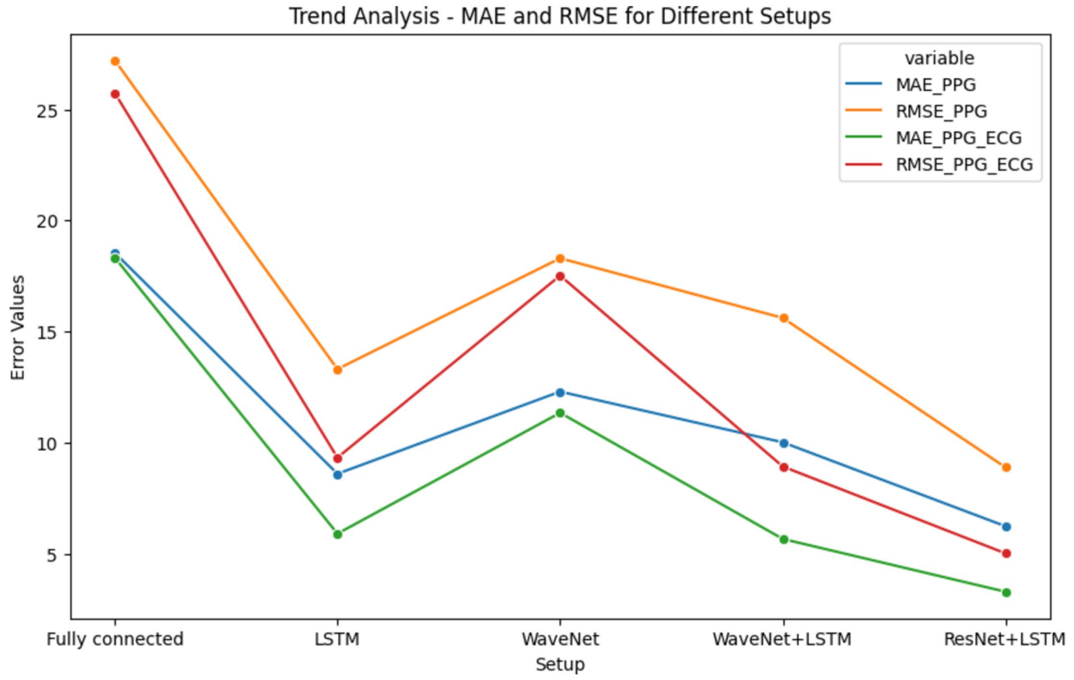


FIGURE 10
Trend analysis of MAE and RMSE for different setups.

performance for each setup in terms of error distribution and prediction accuracy is shown. The MAE and RMSE for direct Systolic Blood Pressure and Diastolic Blood Pressure predictions are compared in the bar chart shown in Figure 10 for several tested sets, such as PPG (50 pat), PPG (40 pat), and ECG (40 pat). The graphic aids in evaluating how well the neural network (ResNet+LSTM) predicts blood pressure based solely on physiological signals.

The stacked bar chart as depicted in Figures 11, 12, illustrates the MAE and RMSE for entire blood pressure prediction across different tested sets. The chart is divided into segments representing MAE and RMSE values for direct SBP and DBP predictions using the ResNet+LSTM neural network. It offers a visual comparison of the errors associated with different physiological signals and tested sets.

The heatmap depicted in Figure 13, provides a comprehensive overview of the computational complexity (Cost estimation in FLOPs) associated with different neural network architectures. Each cell in the heatmap corresponds to a specific neural network's complexity for predicting arterial blood pressure using PPG signals. Darker shades represent higher computational costs.

4.1 Results from our readings

The line graph illustrated in Figure 14, depicts the variation in heart rate values over time, captured at regular intervals during a specific monitoring session. Each data point on the graph represents a recorded heart rate value at a particular timestamp, with the x-axis indicating time and the y-axis representing heart rate values.

The graph in Figure 14 is pertinent to the abstract's exploration of SHM and the utilization of wearable sensors for real-time health monitoring. It reflects the continuous monitoring of physiological parameters, such as heart rate, which is crucial for assessing overall health status and detecting anomalies or irregularities. It provides insights into the temporal dynamics of heart rate, which is a key physiological signal used in BP prediction models. By analyzing trends and fluctuations in heart rate values over time, healthcare professionals can infer patterns of physical activity, stress, or other factors that may impact BP levels. Changes in heart rate patterns, as depicted by the line graph, can serve as indicators of potential health concerns or deviations from normal physiological states, prompting timely interventions or further investigation.

The histogram depicted in Figure 15, provides a visual representation of the distribution of heart rate values, allowing us to identify the central tendency and spread of heart rate data. By observing the shape of the histogram and the location of its central peak, we can gain insights into the typical range of heart rate values recorded during the monitoring session. Additionally, the spread of the histogram can indicate the variability or dispersion of heart rate values around the central tendency. Overall, the histogram provides a summary of the overall heart rate distribution, aiding in the interpretation of physiological responses and patterns observed during the monitoring session.

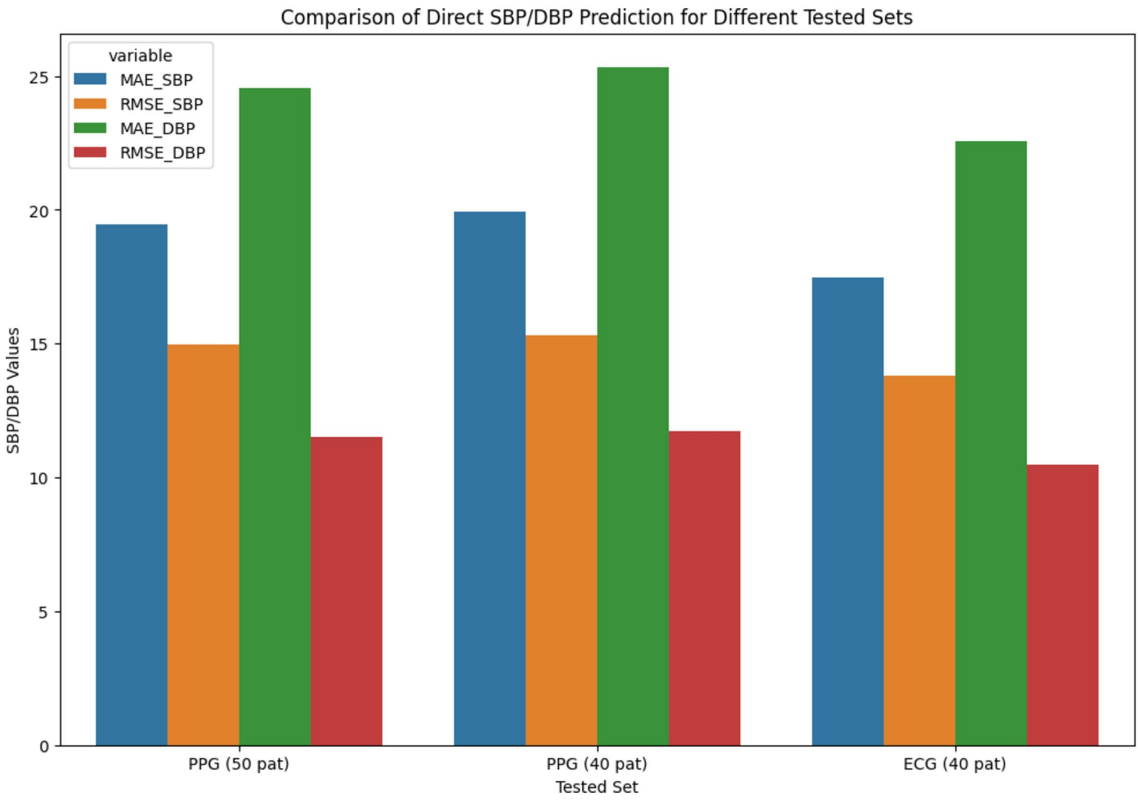


FIGURE 11
Comparison of direct SBP/DBP prediction for different tested sets.

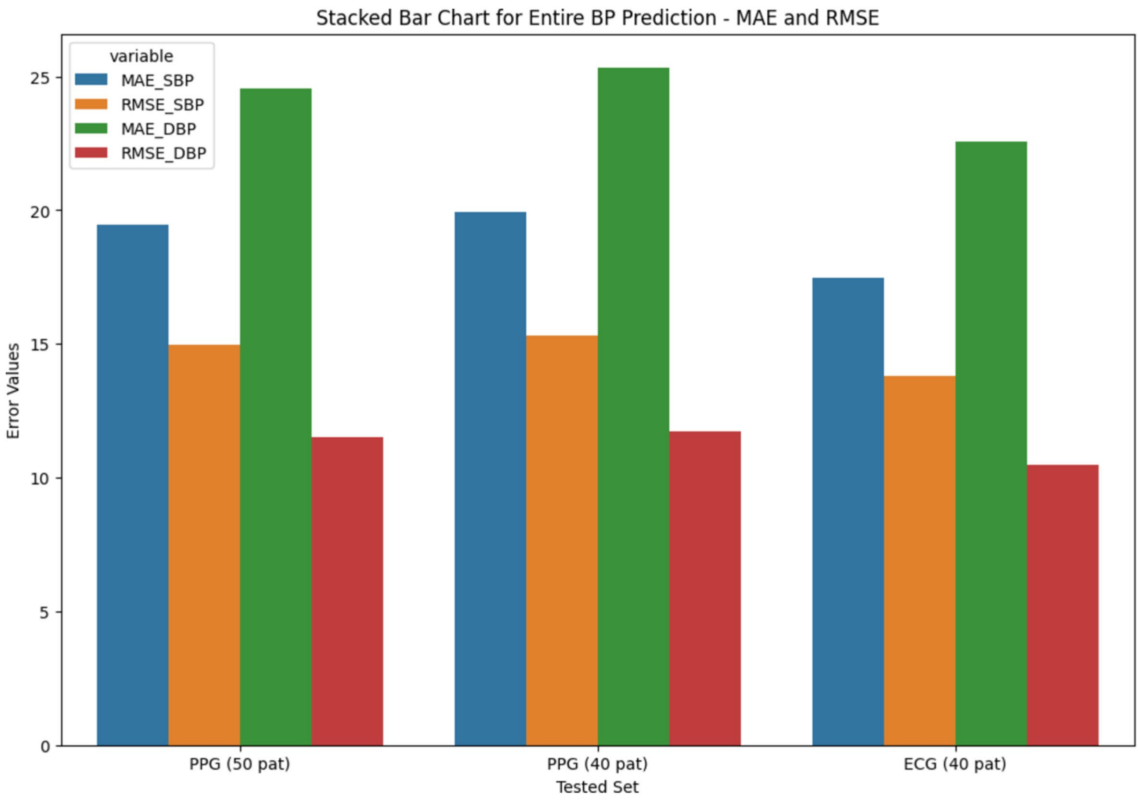


FIGURE 12
Stacked bar chart for entire BP prediction based on MAE and RMSE.

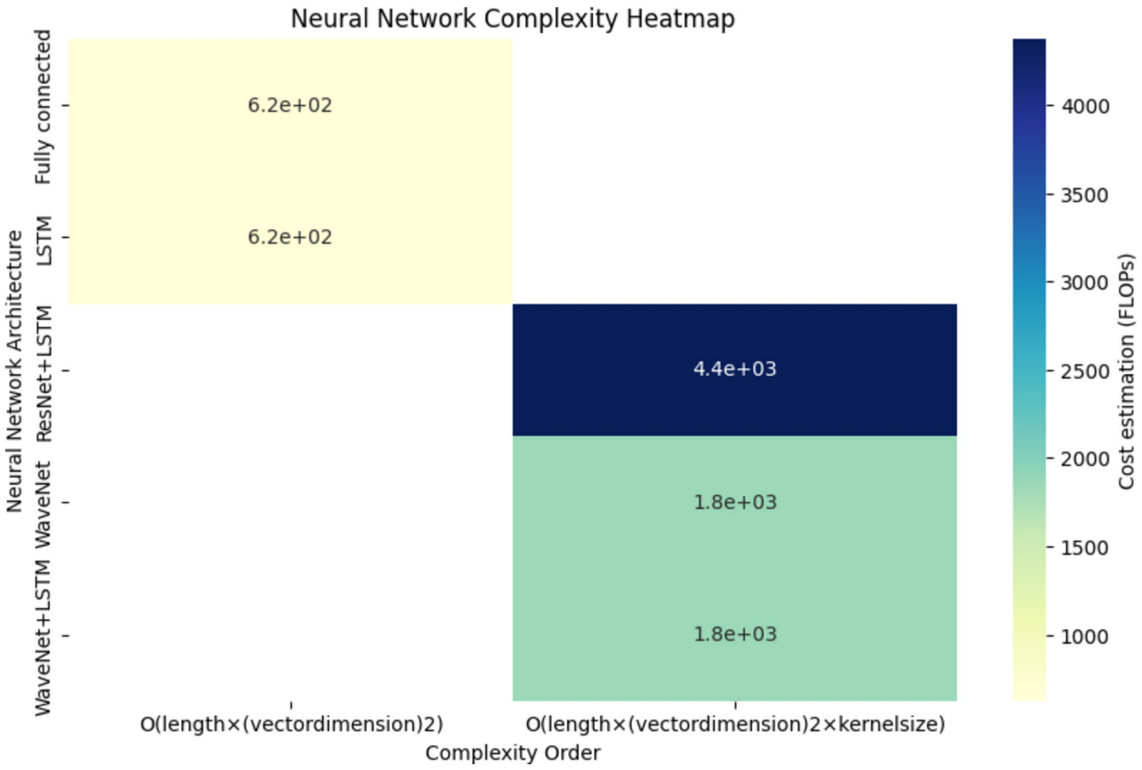


FIGURE 13
Neural network complexity heatmap.

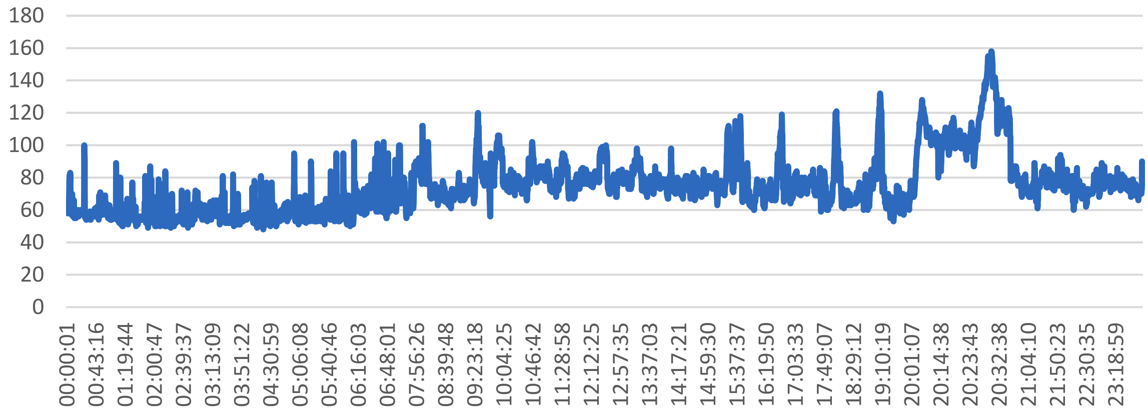


FIGURE 14
Timestamp vs. heart rate line graph from our readings.

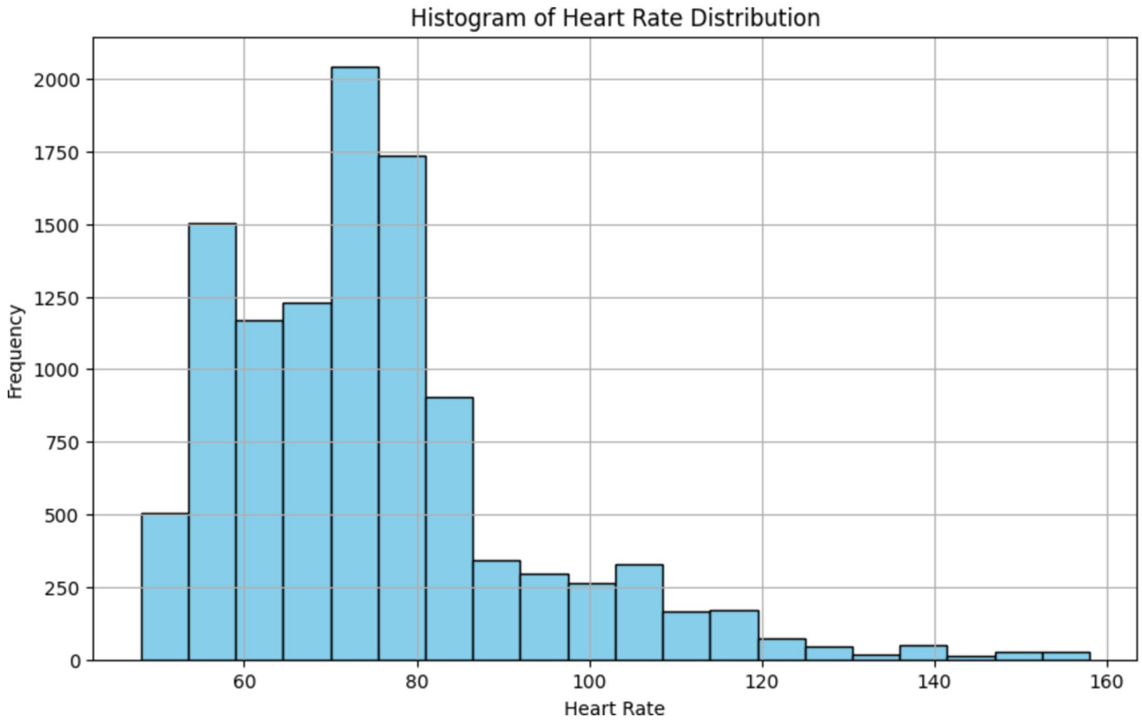


FIGURE 15
Histogram of heart rate distribution from our readings.

The box plot depicted in Figure 16, provides a visual representation of the variability in heart rate values, allowing us to assess the spread and dispersion of the data. By observing the box plot, we can identify the median (central tendency) of the heart rate values, as well as the interquartile range (IQR) which represents the spread of the middle 50% of the data. Additionally, any outliers or extreme values beyond the whiskers of the box plot can indicate potential anomalies or irregularities in the heart rate data. By visually inspecting the box plot, healthcare professionals can identify any outliers or extreme values that may require further investigation or intervention, supporting the overarching goal of continuous monitoring and early detection of health issues.

By analyzing the correlation matrix illustrated in Figure 17, healthcare professionals can identify any correlations or dependencies between different physiological signals, which are crucial for accurately predicting arterial blood pressure and overall health monitoring. Understanding the relationships between physiological signals can aid in the development of effective prediction models and personalized healthcare interventions.

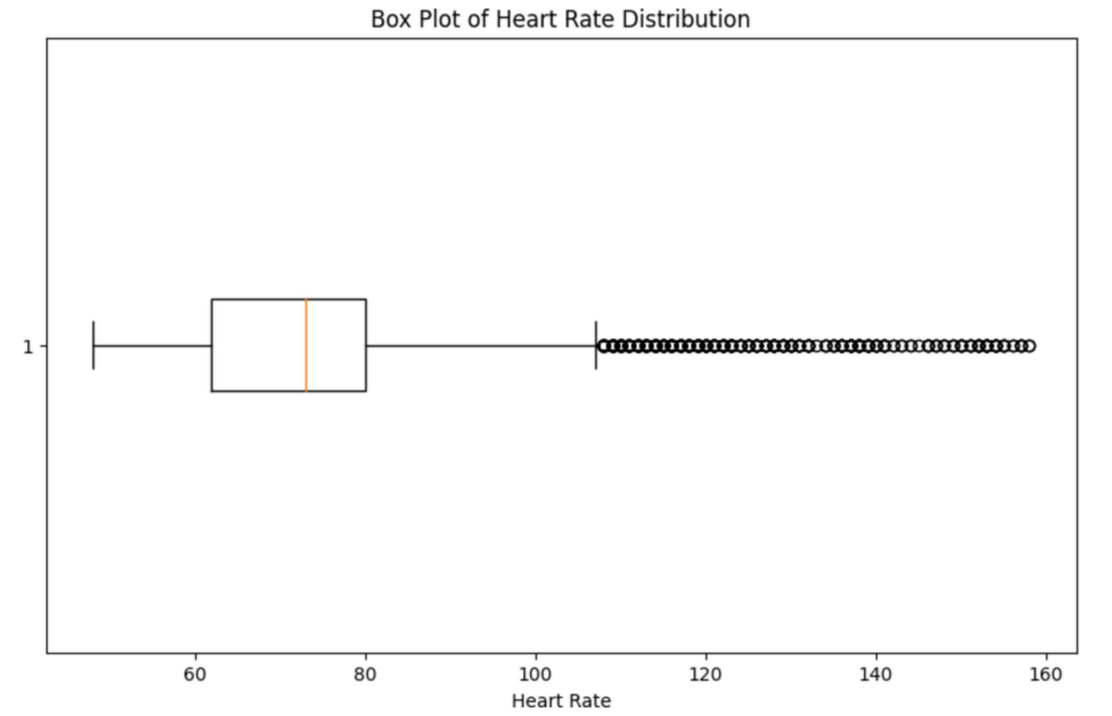


FIGURE 16
Box plot of heart rate distribution from our readings.

5 Conclusion

An approach to evaluate personal healthcare is to continuously monitor physiological markers. Understanding the underlying causes of illness states can be greatly aided by identifying patterns that can be discovered by recognizing outliers or irregularities in heart rates and other characteristics. The vast amount of data collected by wearable device sensors contains irregularities, hence finding anomalies requires accurately automated algorithms. Across the globe, there is a gradual but continuous transition from hospital-based care to patient-centric care. This will gradually pave the way for a data influx, along with the rise in popularity of wearable technology. Any illness status tracking requires continuous wearables-related data points, which are outside the purview of medical care. Such daily longitudinal data collection over extended times can lead to data buildup. The methods used to obtain and disseminate the data determine how the data will be used analytically, as has been described throughout this current article. It is crucial to build wearables-related software for precise health monitoring as well as cutting-edge data collection, analysis, and visualization. Exact clarifications that can be linked to activity of the user and everyday involvement are necessary for both solo and hybrid systems for anomaly identification. The transparency of algorithms used to calculate step or sleep data is another crucial area that should be supported. Enhancement of anomaly detection algorithms take place continuously. To increase the predictability huge datasets have also been made accessible simultaneously. Thus, there is a connection between wearable technology and data analysis and important sectors like the cloud and data security. A wearable device's ability to communicate with hand-held devices like smartphones and the cloud is facilitated by the internet connection, particularly Wi-Fi.

6 Discussion

Numerous detection techniques have been suggested to identify anomalies due to their clinical importance and the effects they have on diagnosis and treatment. Wearable gadget clinical investigations are also becoming more common. Several essential conditions must be met to get therapeutically useful outputs from wearables-related data. To generate suggestions, users must make decisions and align their goals, which both could call for platforms in addition to mobile apps. It is important to test forth current recommendations regarding the correlation of different device-derived data. For instance, when exercise and sleep are connected, the underlying physiological imbalance can be hidden.

By building a dedicated cloud infrastructure for data analysis and storage, wearables can be used more effectively in healthcare. The security of these devices should also be considered. Device-to-device connectivity and an online cloud infrastructure subject to strict regulations are prerequisites for the digital healthcare framework. Eventually, these components might help with the use of wearable big data and accurate anomaly detection in pathology research. Apart from these factors, the next improvements ought to concentrate on the cutting-edge cloud infrastructure designed

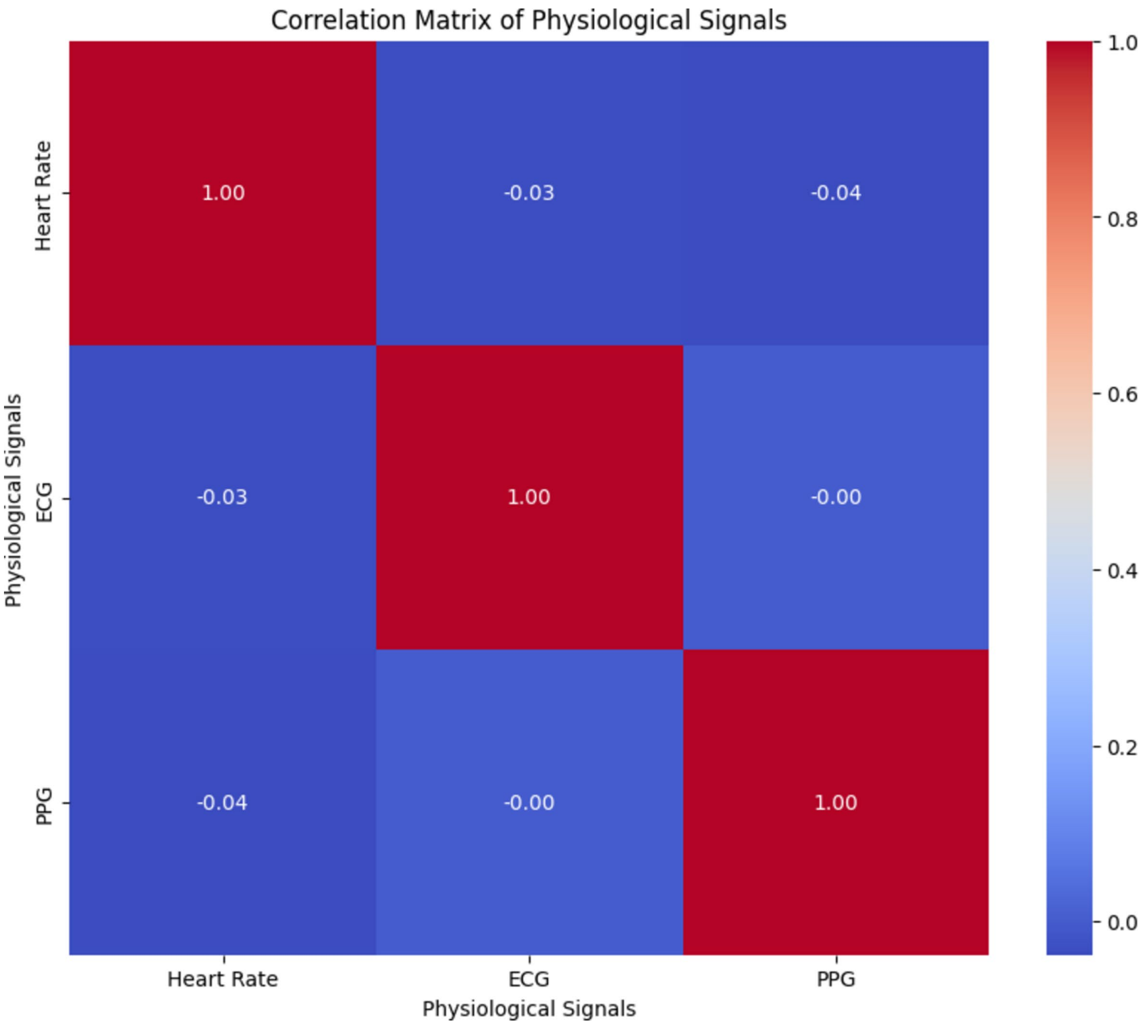


FIGURE 17
Correlation matrix of physiological signals from our readings.

specifically for the analysis and archiving of health data generated by wearables. To ensure scalability, security, and smooth integration with healthcare systems, this infrastructure should effectively manage the substantial amount of data generated by wearables. Moreover, the creation of a networked wearable ecosystem may open new opportunities for health monitoring synergies, enabling the cooperative use of fitness trackers, smartwatches, and medical sensors to provide a thorough understanding of a person's wellbeing.

Work on improving anomaly detection models should consider contextual integration of different wearable data streams. Comprehending the interplay among variables like ambient circumstances, user behavior, and physiological metrics can offer a more comprehensive perspective on a person's health state. This strategy is in line with the development of user-centric decision support systems, which customize insights and suggestions based on data from wearables to personal objectives and health goals. To create a seamless data flow between wearables and healthcare providers, collaboration with current health platforms and electronic health records (EHRs) is essential. Ensuring compliance with interoperability standards promotes timely interventions based on anomaly detection results and comprehensive patient care. In addition, the incorporation of behavioral analytics into models for anomaly detection can provide a more profound comprehension of patterns concerning user behavior, way of life, and compliance with health advice. This can improve the accuracy of anomaly detection by accounting for individual behavioral differences.

Experiential validation studies should be carried out as wearable technology in healthcare continues to advance to evaluate the practicality and user-friendliness of anomaly detection systems. Algorithm modifications based on user and healthcare professional feedback can enhance the technology's overall usefulness. The responsible and efficient integration of wearable technology in healthcare will be further aided by the creation of ethical frameworks, the adoption of explainable AI techniques, and the development of strategies for long-term health monitoring using wearables.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: our dataset: https://github.com/VsinK14/SHM, MIMIC dataset: https://archive.physionet.org/physiobank/database/mimicdb/.

Ethics statement

The datasets used for this study are available publicly; therefore, Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

CK: Conceptualization, Investigation, Project administration, Resources, Supervision, Writing – original draft. VK: Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – original draft. RM: Data curation, Formal analysis, Investigation, Validation, Writing – review & editing, Resources. MA: Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft. FJ: Formal analysis, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. SA: Conceptualization, Investigation, Methodology, Supervision, Visualization, Writing – review & editing. MS: Formal analysis, Methodology, Visualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The authors gratefully acknowledge the funding of the Deanship of Graduate Studies and Scientific Research, Jazan University, Saudi Arabia through Project Number GSSRD-24.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2024.1426168/full#supplementary-material

References

1. Motwani A, Shukla PK, Pawar M. Ubiquitous and smart healthcare monitoring frameworks based on machine learning: a comprehensive review. *Artif Intell Med*. (2022) 134:102431. doi: 10.1016/j.artmed.2022.102431

2. Dertat A. (2017) Autoencoders, Towards data science. Available online at: https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

3. Chang P, Grinband J, Weinberg BD, Bardis M, Khy M, Cadena G, et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *Am J Neuroradiol*. (2018) 39:1201–7. doi: 10.3174/ajnr.A5667

4. Clifford G. D., Liu C., Moody B., Springer D., Silva I., Li Q., et al. "Classification of normal/abnormal heart sound recordings: the physionet/computing in cardiology challenge 2016," in 2016 Computing in Cardiology Conference (CinC) IEEE. (2016), pp. 609–612.

5. Paviglianiti A, Randazzo V, Villata S, Cirrincione G, Pasero E. A Comparison of Deep Learning Techniques for Arterial Blood Pressure Prediction. *Cognit Comput*. (2021) 14:1689–710. doi: 10.1007/s12559-021-09910-0

6. Hasan AM, Jalab HA, Meziane F, Kahtan H, Al-Ahmad AS. Combining deep and handcrafted image features for MRI brain scan classification. *IEEE Access*. (2019) 7:79959–67. doi: 10.1109/ACCESS.2019.2922691

7. General Adversarial Networks. (2024). Geeks for Geeks. Available online at: https://www.geeksforgeeks.org/generative-adversarial-network-gan/

8. Jat Avnish Singh, Grønli Tor-Morten. (2022). Smart watch for smart health monitoring: a literature review. International work conference on bioinformatics and biomedical engineering.

9. Wang Y, Wu Q, Dey N, Fong S, Ashour AS. Deep back propagation–long short-term memory network based upper-limb sEMG signal classification for automated rehabilitation. *Biocybern Biomed Eng*. (2020) 40:987–1001. doi: 10.1016/j.bbe.2020.05.003

10. Singh D, Kumar V, Vaishali KM. Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks. *Eur J Clin Microbial Infect Dis*. (2020) 39:1379–89. doi: 10.1007/s10096-020-03901-z

11. Churová V, Vyškovský R, Maršálová K, Kudláček D, Schwarz D. Anomaly detection algorithm for real-world data and evidence in clinical research: implementation, evaluation, and validation study. *JMIR Med Inform*. (2021) 9:e27172. doi: 10.2196/27172

12. Hamieh S., Heiries V., Al Osman H., Godin C. (2023) Relapse detection in patients with psychotic disorders using unsupervised learning on smartwatch signals. In ICASSP 2023–2023 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 1–2).

13. Jahan I, Al-Nabhan NA, Noor J, Rahaman M, Al Islam AA. Leveraging a smartwatch for activity recognition in Salat. *IEEE Access*. (2023) 11:97284–317. doi: 10.1109/ACCESS.2023.3311261

14. Bozdog I. A., Daniel-Nicusor T., Antal M., Antal C., Cioara T., Anghel I., et al. (2021) Human behavior and anomaly detection using machine learning and wearable sensors. In 2021 IEEE 17th international conference on intelligent computer communication and processing (ICCP) (pp. 383–390).

15. Aliyu MB, Ahmad AAIS. Anomaly detection in wearable location trackers for child safety. *Microprocess Microsyst*. (2022) 91:104545. doi: 10.1016/j.micpro.2022.104545

16. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int J Uncertain Fuzziness Knowl*. (1998) 6:107–16. doi: 10.1142/S0218488598000094

17. MIMIC Dataset. (2016). The MIMIC Database. Available online at: https://archive.physionet.org/physiobank/database/mimicdb/

18. Chen G. A gentle tutorial of recurrent neural network with error backpropagation. *arXiv*. (2016):1–10. doi: 10.48550/arXiv.1610.02583

19. Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Z Med Phys*. (2019) 29:102–27. doi: 10.1016/j.zemedi.2018.11.002

20. Pang G. (2020) Survey on Deep Learning for anomaly detection, Towards data science. Available online at: https://towardsdatascience.com/a-comprehensive-survey-on-deep-learning-for-anomaly-detection-b1989b09ae38

21. Fernando T, Gammulle H, Denman S, Sridharan S, Fookes C. Deep learning for medical anomaly detection - a survey. *ACM Comput Surv*. (2021). doi: 10.1145/3464423

22. Galvez R.L., Bandala A.A., Dadios E.P., Vicerra R.R.P., Maningo J.M.Z., Object detection using convolutional neural networks IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, Institute of Electrical and Electronics Engineers Inc. (2019), pp. 2023–2027.

23. Our Dataset. (2024). Dataset. Available online at: https://github.com/VsinK14/SHM

24. Rostamian A, O'Hara JG. Event prediction within directional change framework using a CNN-LSTM model. Berlin: Springer (2022).

25. García-Macías E, Ubertini F. Integrated SHM systems: damage detection through unsupervised learning and data fusion In: Structural health monitoring based on data science techniques. Berlin: Springer (2021)

26. Fastforward Labs. (2020). Deep Learning for Anomaly Detection. Available online at: https://ff12.fastforwardlabs.com/

27. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. *Pattern Recogn*. (2018) 77:354–77. doi: 10.1016/j.patcog.2017.10.013

28. Jithin S, Sunny CPK, Karnani PK, Sandeep C, Pingle FL, Anekoji M, et al. Anomaly detection framework for wearables data: a perspective review on data concepts, data analysis algorithms and prospects. *Sensors*. (2022) 22:756. doi: 10.3390/s22030756

29. Allen J. Photoplethysmography and its application in clinical physiological measurement. *Physiol Meas*. (2007) 28:R1–R39. doi: 10.1088/0967-3334/28/3/r01

30. Avcı K. A novel method for classifying liver and brain tumors using convolutional neural networks, discrete wavelet transforms and long short-term memory networks. *Sensors*. (2019) 19:1992. doi: 10.3390/s19091992

31. Goldstein M, Uchida S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS One*. (2016) 11:e0152173. doi: 10.1371/journal.pone.0152173

32. Islam Z, Islam M, Asraf A. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. *Inform Med Unlocked*. (2020) 20:100412. doi: 10.1016/j.imu.2020.100412

33. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: full training or fine-tuning? *IEEE Trans Med Imag*. (2016) 35:1299–312. doi: 10.1109/TMI.2016.2535302

34. Hea R, Xua P, Chen Z, Luo W, Zhineng S, Mao J. A non-intrusive approach for fault detection and diagnosis of water distribution systems based on image sensors, audio sensors and an inspection robot. *Energy and Buildings*. (2021). doi: 10.1016/j.enbuild.2021.110967

35. Aburakhia Sulaiman, Tayeh Tareq, Myers Ryan, Shami Abdallah. (2020) A transfer learning framework for anomaly detection using model of normality. 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). Vancouver, BC, Canada.

36. Raoof SS, Durai MAS. A comprehensive review on smart health care: applications, paradigms, and challenges with case studies. *Contrast Media Mol Imaging*. (2022). doi: 10.1155/2022/4822235

37. Ohta S., Nakamoto H., Shinagawa Y., Tanikawa T. A health monitoring system for elderly people living alone. *J Telemed Telecare*. (2002) 8:151–156. doi: 10.1177/1357633X0200800305

38. Tamura-Chen W, Tamura T. Seamless Healthcare Monitoring: Advancements in Wearable, Attachable, and Invisible Devices, 1st ed. Cham, Switzerland: Springer (2018).

39. Hea R, Xua P, Chen Z, Luo W, Zhineng S, Mao J. A non-intrusive approach for fault detection and diagnosis of water distribution systems based on image sensors, audio sensors and an inspection robot. *Energy and Buildings*. (2021).

40. Kalpana R., Nageshwar Rao D., Patro AK. "A survey on deep learning techniques for anomaly detection in human activity recognition," *Smart and Sustainable Technologies: Rural and Tribal Development Using IoT and Cloud Computing: Proceedings of ICSST 2021*. (2022) 337–347.

frontiers | Frontiers in Physiology

TYPE Original Research
PUBLISHED 13 February 2025
DOI 10.3389/fphys.2025.1491401

Check for updates

OPEN ACCESS

EDITED BY
Mohammad Yavarimanesh,
University of San Diego, United States

REVIEWED BY
Colin K Drummond,
Case Western Reserve University, United States
Younes Rezaee Danesh,
Yüzüncü Yıl University, Türkiye

*CORRESPONDENCE
Stefan Mendt,
✉ stefan.mendt@charite.de

†These authors share first authorship

‡These authors share senior authorship

RECEIVED 04 September 2024
ACCEPTED 16 January 2025
PUBLISHED 13 February 2025

CITATION
Mendt S, Zout G, Rabuffetti M, Gunga H-C, Bunker A, Barteit S and Maggioni MA (2025) Laboratory comparison of consumer-grade and research-established wearables for monitoring heart rate, body temperature, and physical acitivity in sub-Saharan Africa.
Front. Physiol. 16:1491401.
doi: 10.3389/fphys.2025.1491401

Laboratory comparison of consumer-grade and research-established wearables for monitoring heart rate, body temperature, and physical acitivity in sub-Saharan Africa

Stefan Mendt [1†*], Georgi Zout [1†], Marco Rabuffetti [2], Hanns-Christian Gunga [1], Aditi Bunker [3], Sandra Barteit [3‡] and Martina Anna Maggioni [1,4‡]

[1]Charité - Universitätsmedizin Berlin, Institute of Physiology, Center for Space Medicine and Extreme Environments Berlin, Berlin, Germany, [2]IRCCS Fondazione Don Carlo Gnocchi, Milano, Italy, [3]Heidelberg Institute of Global Health, Heidelberg University Hospital, Heidelberg University, Heidelberg, Germany, [4]Department of Biomedical Sciences for Health, Università degli Studi di Milano, Milano, Italy

Background: Consumer-grade wearables are becoming increasingly popular in research and in clinical contexts. These technologies hold significant promise for advancing digital medicine, particularly in remote and rural areas in low-income settings like sub-Saharan Africa, where climate change is exacerbating health risks. This study evaluates the data agreement between consumer-grade and research-established devices under standardized conditions.

Methods: Twenty-two participants (11 women, 11 men) performed a structured protocol, consisting of six different activity phases (sitting, standing, and the first four stages of the classic Bruce treadmill test). We collected heart rate, (core) body temperature, step count, and energy expenditure. Each variable was simultaneously tracked by consumer-grade and established research-grade devices to evaluate the validity of the consumer-grade devices. We statistically compared the data agreement using Pearson's correlation *r*, Lin's concordance correlation coefficient (LCCC), Bland-Altman method, and mean absolute percentage error.

Results: A good agreement was found between the wrist-worn Withings Pulse HR (consumer-grade) and the chest-worn Faros Bittium 180 in measuring heart rate while sitting, standing, and slow walking on a treadmill at a speed of 2.7 km/h ($r \geq 0.82$, $|\text{bias}| \leq 3.1$ bpm), but this decreased with increasing speed ($r \leq 0.33$, $|\text{bias}| \leq 11.7$ bpm). The agreement between the Withing device and the research-established device worn on the wrist (GENEActiv) for measuring the number of steps also decreased during the treadmill phases (first stage: $r = 0.48$, bias = 0.6 steps/min; fourth stage: $r = 0.48$, bias = 17.3 steps/min). Energy expenditure agreement between the Withings device and the indirect calorimetry method was poor during the treadmill test ($|r| \leq 0.29$, $|\text{bias }| \geq 1.7$ MET). The Tucky thermometer under the armpit (consumer-grade) and the Tcore sensor on the forehead were found to be in poor agreement in measuring (core) body temperature during resting phases ($r \leq 0.53$, $|\text{bias}| \geq 0.8$°C) and deteriorated during the treadmill test.

Conclusion: The Withings device showed adequate performance for heart rate at low activity levels and step count at higher activity levels, but had limited overall accuracy. The Tucky device showed poor agreement with the Tcore in all six different activity phases. The limited accuracy of consumer-grade devices suggests caution in their use for rigorous research, but points to their potential utility in capture general physiological trends in long-term field monitoring or population-health surveillance.

KEYWORDS

fitness tracker, accelerometer, physiological parameters, global health, heat stress, SSA

1 Introduction

According to the Intergovernmental Panel on Climate Change, the global average annual temperature is expected to rise by 1.5°C between 2030 and 2052 (compared to pre-industrial levels) due to greenhouse gas emissions and other human activities (Masson-Delmotte et al., 2018). A rise in global temperature causes extreme weather events, which pose an increased risk to nature, the economic, and to human health (Eitelwein et al., 2024). For example, prolonged periods of unusually low rainfall (droughts) threaten food and water security, heatwaves will lead to a significant increase in temperature-related diseases and deaths, wildfires will increase air pollution, and flooding will increase crop damage and the risk of disease. The World Health Organization estimates that climate change will cause around 250,000 additional deaths per year due to malnutrition, malaria, diarrheal diseases, and heat stress alone (World Health Organization, 2023). Climate change also adversely affect the ability to work and reduce labor productivity (Kjellstrom et al., 2009), particularly in low-income regions such as sub-Saharan Africa (SSA), where subsistence agriculture is crucial for the livelihood of small rural communities (Asare-Nuamah, 2021; Ayal, 2021). For example, the simulated effects of climate change on agricultural production in the eastern and coastal regions of Kenya predicts a at least 50% rest/hour work intensity during the planting season and a up to 50% rest/hour work intensity during the maize harvesting period for the years 2050 and 2100 (Yengoh and Ardö, 2020). As smallholder farmers use a lot of human labor, an increase in environmental temperature has a considerable impact on their health. In addition to increased cardiovascular stress and impaired physical and cognitive functions, physical exertion due to labor increases the incidence of heatstroke (Bouchama et al., 2022).

Research on the effects of environmental heat-related stress on health and work ability in low- and middle-income countries primarily relies on data from hospitals, surveys, and of Health and Demographic Surveillance Systems (Diboulo et al., 2012; Egondi et al., 2012; Katiyatiya et al., 2014; Park et al., 2018; Chavaillaz et al., 2019; Frimpong et al., 2020; Barteit et al., 2023; Sapari et al., 2023). The ability to monitor physiological responses to heat stress such as heart rate, body temperature, and physical activity directly in the field using wearable devices would provide invaluable data for managing health risks in smallholder farmers and residents in SSA. Objective monitoring of physical activity has rapidly advanced in recent decades with the development of commercial and research-grade wearables. Compared to research-grade technologies, consumer-grade wearables are often lower in cost, easier to use, less obtrusive and not tied to a specific location (Dunn et al., 2018); however, these advantages often come at the expense of data accuracy.

Despite the growing use of wearables in high-income settings, there is limited research on their application in low-income, climate-vulnerable regions such as SSA (Koch et al., 2022). Recent studies have demonstrated the utility of wearable devices in low-resource settings though concerns remain about the trade-offs between affordability and accuracy (Huhn et al., 2022; Matzke et al., 2024). The present study seeks to fill this gap by comparing the accuracy of consumer-grade wearables under controlled conditions. Previous studies already dealt with comparison of different wearables measuring the same physiological parameter under controlled conditions (Nelson et al., 2016; Gillinov et al., 2017; Wahl et al., 2017; Eisenkraft et al., 2023). In this study, however, we focus on multiple physiological parameters that are relevant for assessing the environmental impact on human health and performance at individual level. For this purpose, a sample of young adults was equipped with a set of wearable devices for monitoring heart rate, body temperature, and physical acitvity (steps, energy expenditure) during rest and activity periods in a laboratory environment.

2 Methods

2.1 Study participants

We recruited young men and women for our study among medical students through advertisements on the internal Charité student's platform and social media. Those interested were eligible for inclusion if they were between the ages of 18 and 30 and had no history of competitive training. On the other hand, interested were excluded if they had any form of cardiovascular, metabolic, and neurological diseases, or any physical impairments that would prevent participation in an incremental test on a treadmill. Following explanations of the study aim and protocol, including experimental procedures and known risks, participants provided informed written consent prior to commencing study participation. Based on a sample size calculation using the results of a comparative study between commercial trackers and a portable ECG (Godino et al., 2020), the study sample was planned with 20 participants. To ensure conclusive statistical results at the end of the study, we recruited a total of 22 participants (11 women, 11 men). Their anthropometric data were as follows: age, mean 24.0 (SD 2.4) years; body weight, mean 70.2 (SD 7.7) kg; height, mean 176 (SD 9.1) cm; body mass index, mean 22.6 (SD 1.6) kg/m^2. The study was

TABLE 1 Overview of selected consumer-grade and research-grade wearables.

	Consumer-grade		Research-grade		
Weareable device	Withings Pulse HR	Tucky Thermometer	Faros Bittium 180	GENEActiv	Tcore sensor with data logger headband
Company	Withings France SA, Issy-les-Moulineaux, France	e-TakesCare, Versailles, France	Bittium Corporation, Oulu, Finland	Activinsights, Kimbolton, UK	Sensor: Drägerwerk AG and Co. KGaA, Lübeck, Germany, data logger: HealthLabFunkMaster, KORA Industrie-Elektronik GmbH, Hambühren, Germany
Dimension	18 × 10 × 44 mm	84 × 27 × 7 mm	48 × 29 × 12 mm	43 × 40 × 13 mm	Sensor: 60 × 50 × 4 mm, data logger: 48 × 30 × 5 mm
Weight	45 g	8 g	13 g	28 g	Sensor: 3 g data logger: 15 g
Wear location	wrist	under armpit	3 electrodes on thorax	wrist	forehead
Sample rate	every minute for heart rate (1 Hz in workout mode), steps, energy expenditure	every minute	up to 1,000 Hz	up to 100 Hz	0.5 Hz
Internal storage	yes	no	yes	yes	yes (data logger)
Data transfer	bluetooth low energy	bluetooth low energy	USB cable	cradle with USB cable	USB cable (data logger)
Measurement features	heart rate, distance, calories, sleep	body (shell) tempereature, sleeping position monitor	1-lead electrocardiography, tri-axial accelerometer	tri-axial accelerometer light exposure, near body temperature	body (core) temperature

approved by the Ethics Committee of Charité–Universitätsmedizin Berlin (Date: 9 April 2021, EA 4/050/21).

2.2 Data acquisition

The research and consumer-grade wearables considered for evaluation in this study were selected from a study protocol designed for the purpose of providing scientific information on their reliability for the use in the setting of population monitoring in SSA (Barteit et al., 2021). Table 1 provides details of the consumer- and research-grade wearables in the present study.

2.2.1 Consumer-grade wearables

Withings (Withings France SA, Issy-les-Moulineaux, France): We used the Withings Pulse HR device to measure heart rate (HR), steps taken, and calories burned. Data from the internal storage was wirelessly synchronized with a mobile device via the Health Mate application.

Tucky (e-TakesCare, Versailles, France): The Tucky device, a flexible thermometer patch, was used to measure axillary temperature. The recordings were transfered directly via Bluetooth to a mobile device that used the Tucky application.

2.2.2 Research-grade wearables

Faros™ (Bittium Corporation, Oulu, Finland): The Faros Bittium 180 is a gold-standard portable one-lead electrocardiography monitor. It enables long-duration beat-to-beat recordings both inside and outside hospital and healthcare facilities (Laborde et al., 2017; Hartikainen et al., 2019; Bent et al., 2020; Funston et al., 2022; Lang et al., 2022).

Tcore™ (Drägerwerk AG and Co. KGaA, Lübeck, Germany): The Tcore sensor calculates core body temperature (CBT) using a dual-sensor heat flux technology integrated into a soft sensor attached to the forehead (Werner and Gunga, 2020). Accuracy and validty of this technology is given elsewhere (Gunga et al., 2008; Mendt et al., 2017; Soehle et al., 2020; Janke et al., 2021; Engelbart et al., 2023). In this study, the sensor cable was connected to a data logger (HealthLabFunkMaster, KORA Industrie-Elektronik GmbH, Hambühren, Germany) and the data logger was integrated into a custom-made headband.

GENEActiv (Activinsights, Kimbolton, UK): We used the GENEActiv to record raw acceleration data (range ±8 g) along three orthogonal axes (x-, y- and *z*-axis). Post-processing of the tri-axial accelerometric data enables an objective assessment of physical activities (e.g., energy expenditure, step count) and sleep behavior (Scott et al., 2017; Sanders et al., 2019; Fraysse et al., 2020; Antczak et al., 2021; Jenkins et al., 2022; Hachenberger et al., 2023).

Cortex Metalyzer 3B (CORTEX Biophysik GmbH, Leipzig, Germany): The Cortex Metalyzer 3B is a spiroergometry system designed for measuring oxygen consumption and carbon dioxide production using breath-by-breath gas analysis to calculate energy expenditure (EE) via indirect calorimetry. The device was calibrated once and directly before the study for volume and gas concentrations. For gas calibration, a mixture of 15% oxygen, 5% carbon dioxide, and balance nitrogen was used.

2.3 Study procedure

The measurements were conducted in the laboratories of the Institute of Physiology, Charité–Universitätsmedizin Berlin on

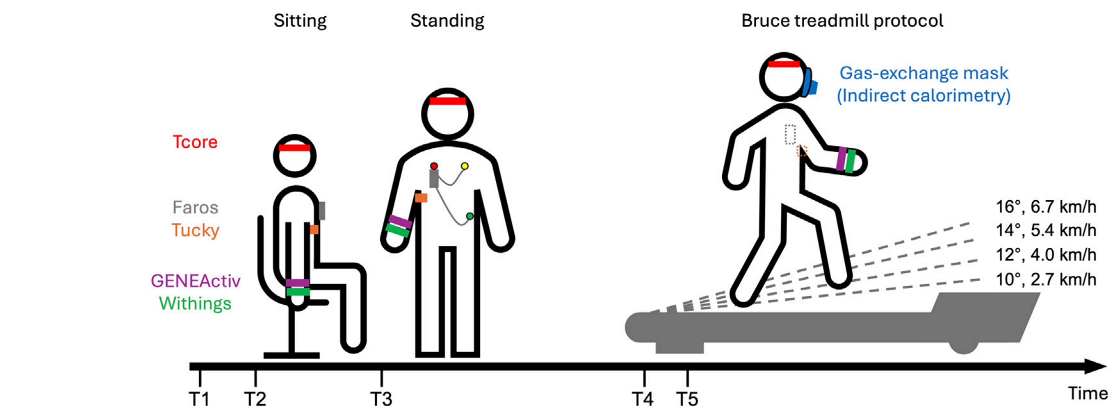


FIGURE 1
Schematic overview of the experimental protocol: Study participants were first equipped with various devices (T1). After initial setups, study participants sat for 10 min (T2) and then rested for an additioinal 10 min while standing (T3). Participants were fitted with a mask connected to the Cortex Metalyzer and rested for 3 min on the treadmill (T4). Study participants started the classic Bruce protocol (T5). The classic Bruce treadmill test consists of 3-min stages, with speed and slope increasing every 3 min without breaks. Speed and slope are displayed for the first four stages.

weekdays between 9:00 and 14:30 in September 2021. Study participants followed a structured, laboratory-based protocol that included two different resting phases followed by different locomotion phases on a motorized treadmill (Figure 1). In particular, we wanted to simulate intensities typical of the daily routines of subsistence farmers in SSA regions. For example, the metabolic equivalent of task (MET) for the classic Bruce treadmill protocol is estimated to be 4.2 MET for the first stage and 8.3 MET for the third stage according to the FRIEND equation (Kokkinos et al., 2017). MET values of 4.5 and 7.8 correspond to routine chores with small animals and shovel or pitchfork work, respectively (Pickett et al., 2015).

Participants were first equipped with various devices. The Tucky device was placed under the right armpit using Tucky double-sided adhesive Adh21. Due to an initial detachment on the first study participant, we have since positioned the device closer to the chest and additionally secured it with medical adhesive tape. The Tcore sensor and headband was fastened on the participants' forehead. The Faros was positioned on the chest and secured with medical adhesive tape to ensure signal quality. GENEActiv and Withings were placed on the wrist of the non-dominant arm, with GENEActiv positioned directly above Withings.

After these initial setups, participants sat for 10 min and then rested for an additional 10 min while standing. Following this period of rest, measurements continued on a motorized treadmill (h/p/cosmos quasar med 4.0, Nussdorf-Traunstein, Germany). Similar comparative studies also utilized treadmills as test environments (Thiebaud et al., 2018; Thomson et al., 2019). On the treadmill, participants were fitted with a mask over their mouth and nose which was connected to the Cortex Metalyzer, and were also fitted with a harness system to prevent falls on the treadmill. After a 3-min rest on the treadmill, the study participants started the Bruce protocol continuing until complete exhaustion. The classic Bruce treadmill test consists of 3-min stages, with speed and slope increasing every 3 minutes without breaks (Fletcher et al., 2013). The first four stages are as follows: Stage1: 2.7 km/h (1.7 mph), 10%; Stage2: 4.0 km/h (2.5 mph), 12%; Stage3: 5.4 km/h (3.4 mph), 14%; Stage4: 6.7 km/h (4.2 mph), 16%.

Following the treadmill test, the collected data were retrieved and stored on a study computer. Data from Withings and Tucky were downloaded from their respective platforms and spiroergometric data were exported via MetaSoft software, while Tcore, Faros, and GENEActiv data were transferred directly from their internal storage. To ensure synchronization among all considered data logs for later data analysis, timestamps were documented during the experiments. First, the times on the computers associated with the different monitors were recorded at the beginning of each measurement day to account for potential time offsets. This was necessary because GENEActiv, Faros and Tcore were initialized with the study computer, while Withings and Tucky were initialized with the same mobile device, and spiroergometry was conducted using a separate computer. Secondly, the time (on the study computer) at which the rest and activity measurements began was noted.

2.4 Analysis

2.4.1 Data processing

For the data analysis, we considered the period from minute 3 to 8 (6 min) of the 10-min rest phases in sitting (Sit) and standing (Stand) to reduce variability due to excitement or changes in posture. Only the first four stages of the Bruce protocol were analyzed, as all 22 study participants successfully completed these stages. Recordings required processing due to differing units and sampling rate. Faros' R-R intervals were transformed to HR (using the formula: HR = 60/R-R) and synchronized with the HR measurements taken every second by the Withings wearable. Both HR_{Faros} and $HR_{Withings}$ were then averaged to 1-min intervals. Tucky measures temperature under the armpit (axillary temperature). To obtain an equivalent rectal (core body, CBT) temperature and enable comparison with Tcore temperature (CBT_{Tcore}), we added 0.7°C to the recorded Tucky temperature

(CBT_{Tucky}) as suggested by the Tucky sensor description. CBT_{Tcore}, initially recorded at 0.5 Hz, was averaged to 1-min intervals. The step count estimate from Withings ($SC_{Withings}$) was compared with $SC_{GENEActiv}$, the result of a step counting function implemented in the R package "GENEAclassify" (Campbell et al., 2023). The input for this function was the vector magnitude, VM = sqrt (x^2+y^2+z^2), which we calculated from the tri-axial acceleration data recorded with the GENEActiv. Since the GENEActiv sampling rate was initially set to 10 Hz to be consistent with in field studies, $SC_{GENEActiv}$ was averaged to 1-min intervals. Energy expenditure during the Bruce test was captured using three different approaches. The first was the indirect calorimetry method, the gold standard for determining energy expenditure by measuring the volume of oxygen consumed and the volume of carbon dioxide produced (Ndahimana and Kim, 2017). Output of indirect calorimetry (EE_{IC}) was the objective measure of the metabolic equivalent of task (MET, 1MET = 3.5 mlO_2 kg^{-1} min^{-1}). The second was with Withings ($EE_{Withings}$), which however provide data values in kcal per minute. We converted this data into MET using an equation presented in ACSM's Guideline for Exercise Testing and Prescription (Riebe, 2014). In the third approach, EE was estimated with a prediction formula (EE = 5.01 + 1.000 ENMO) derived from accelerometry data ($EE_{GENEActiv}$) of free-living adults (White et al., 2016). We calculated the Euclidian norm minus one (ENMO = VM-1) again using the tri-axial acceleration data recorded with the GENEActiv.

2.4.2 Statistical analysis

For the resting (Sit, Stand) and locomotion phases (Stage1, Stage2, Stage3, and Stage4), agreement between two approaches was verified using the following indicators to facilitate comparison with related previous works.

- Pearson correlation: This coefficient *r* was determined to specify the degree of linear relationship.
- Lin's concordance correlation coefficient (LCCC): Lin's CCC includes precision in addition to Pearson's *r* (Lin, 1989), providing a more comprehensive measurement of agreement.
- Bland–Altman method (Bland and Altman, 1986): This method provided the mean difference between the methods (bias) and the limits of agreement (LoA, bias±1.96SD of the differences). Lin's CCC and Bland-Altman analysis were carried out with the R package "SimplyAgree" (Caldwell, 2022).
- Mean absolute percentage error (MAPE): MAPE was calculated according to the formula:

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{CG_t - RG_t}{CG_t}\right|$$

where CG_t represented the consumer-grade measurement and RG_t represented the research-grade measurement.

The difference between two methods was tested using the *t*-test or the Wilcoxon signed-rank test, depending on the result of the Shapiro-Wilk test for normality. The level of significance was set at 0.05 (two-sided), and *P* values were adjusted according to Holm to account for multiple testing. All statistical analyses were carried out using R (version 4.2.0; R Core Team, 2022). Scatterplots and bar charts were created with the R package "ggplot2" (Wickham, 2016).

3 Results

The final dataset for HR, CBT, SC, and EE analysis included 21 participants. To ensure data quality, we excluded HR data of one participant, as 40% of the $HR_{Withing}$ readings during Sit, Stand, and Stage1 were between 43 and 58 bpm, inconsistent with the non-athlete status of our study participants. Additionally, the associated HR_{Faros} readings were almost twice as high each time. For CBT, data from one participant were excluded because the Tucky wearable fell off during treadmill exercise. For SC and EE, one GENEActiv file was corrupted.

Figure 2 displays scatterplots comparing HR, CBT, SC, and EE across all phases. Individual differences between methods are shown in Figure 3. Table 2 provides an overview of HR, CBT, SC, and EE values during rest and locomotion phases, including statistical summaries. Table 3 summarizes the agreements between the methods for all phases.

3.1 Heart rate

In both resting states, the heart rate was similar for both methods. With increasing physical activity, $HR_{Withings}$ did not increase to the same extent as the HR_{Faros} (Figure 3A). At the 4th stage, the mean difference between the methods was −12 bpm, the largest and statistically significant (Table 2). Correlations were strong and positive for Sit and Stand ($r \geq 0.82$, LCCC ≥ 0.76). However, the agreement between $HR_{Withings}$ and HR_{Faros} decreased with increasing physical activity (Table 3). For example, MAPE was more than twice as high from Stage2 (≥10%) as during both resting phases (≤4%).

3.2 Core body temperature

CBT_{Tucky} was consistently lower than CBT_{Tcore} in all phases (Figure 2B), which was confirmed by statistical analysis (Table 2). The difference between the methods was smallest at rest (Sit, −0.8°C, $t_{20} = -5.44$, $P < 0.001$) and largest in the fourth stage of the Bruce test (−1.8°C, $t_{20} = -10.35$, $P < 0.001$). CBT_{Tucky} remained unchanged across different situations (ranged between 36.3°C and 36.5°C), while CBT_{Tcore} increased with physical effort (ranging between 37.2°C and 38.1°C). Similar to HR, the correlations between the temperature monitors declined with physical activity. In addition, LoA became wider and the MAPE increased (Table 3).

3.3 Step count

At a treadmill speed of 2.7 km/h (Stage1), step counts were similar between $SC_{Withings}$ and $SC_{GENEActiv}$ (72.2 vs. 71.5 steps/min, z = 0.54, $P = 0.61$). However, the difference between the methods increased with increasing speed (Figure 3C), while $SC_{Withings}$ increasingly exceeding $SC_{GENEActiv}$ (Table 2). For example, at a treadmill speed of 6.7 km/h (Stage4), $SC_{Withings}$ exceeded $SC_{GENEActiv}$ by about 17 steps/min. (152.2 vs. 134.9 steps/min, $t_{20} = 8.07$, $P < 0.001$). On the other hand, LoA at Stage4 was only half as wide as at Stage1 (Table 3). MAPE was highest in Stage1 (38%), but was only around 10% in the following three stages.

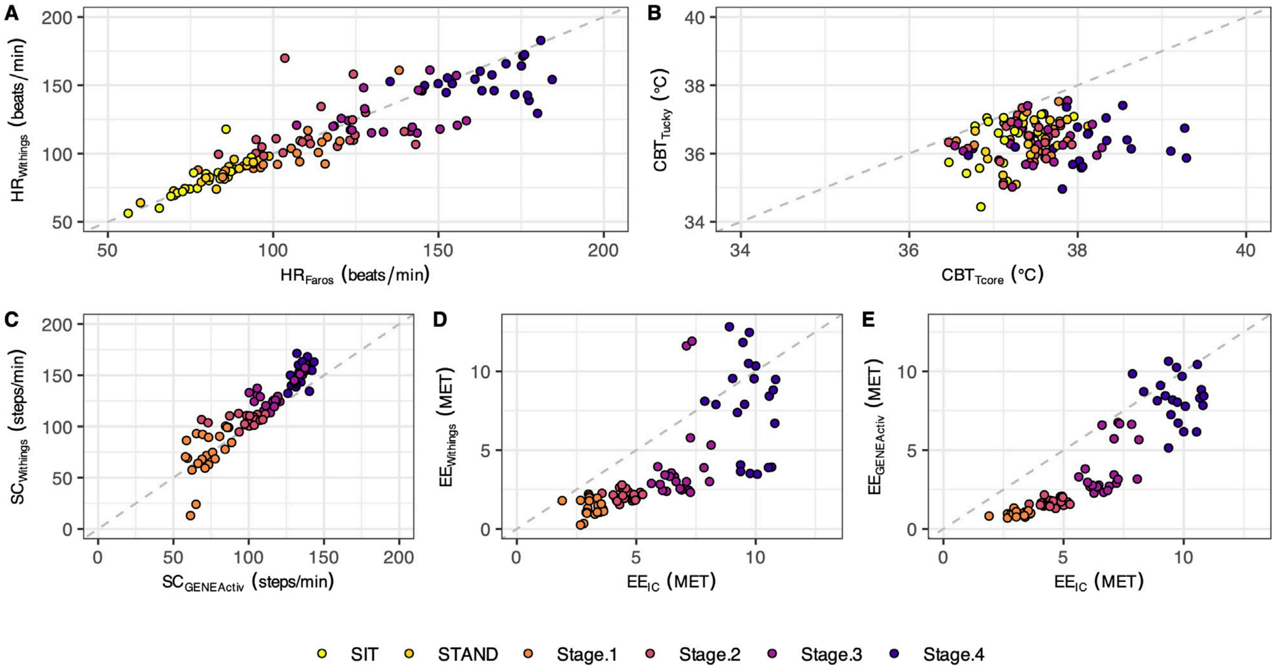


FIGURE 2
Scatterplots with the identity line. The plotted points represent individual mean values (n = 21) for the different test phases (each phase shown in a different color). The scatterplots illustrate the data for heart rate **(A)**, core body temperature **(B)**, step count **(C)**, and energy expenditure **(D, E)**.

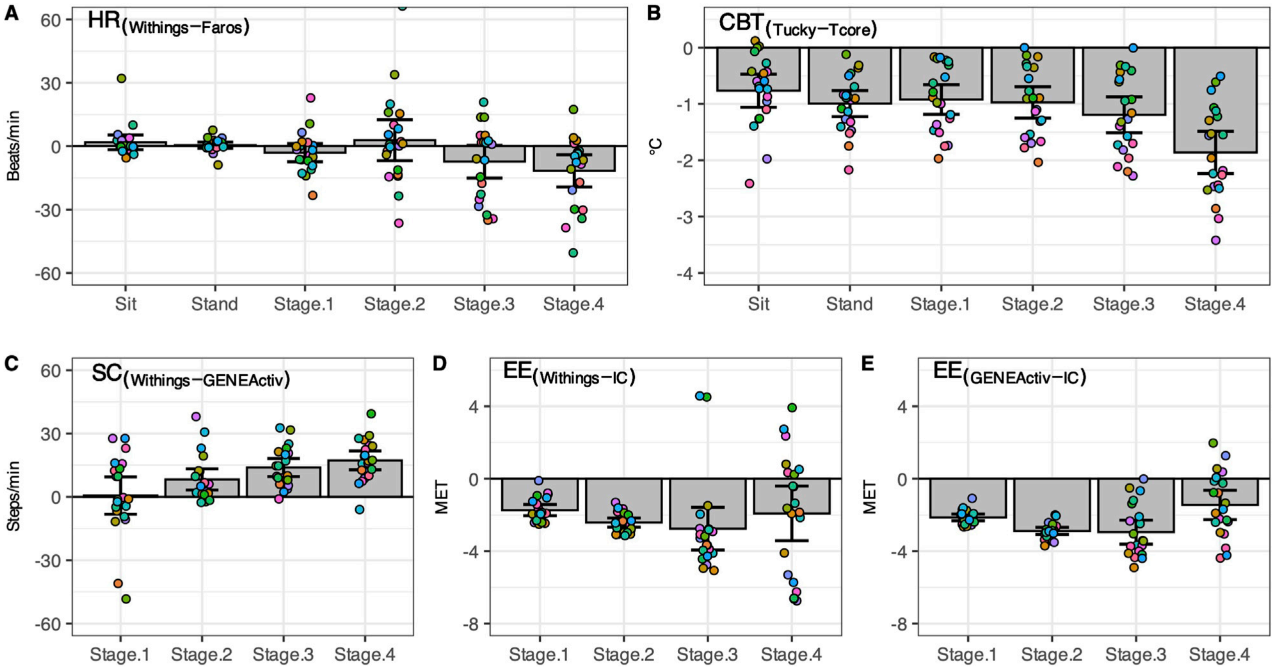


FIGURE 3
Difference between consumer-level and research-grade monitors. Individual differences (n = 21, circles) as well as mean ± 95% CI are shown for heart rate **(A)**, core body temperature **(B)**, step count **(C)**, and energy expenditure **(D, E)**. Sit, sitting position; Stand, standing position; Stage1 to Stage4, first four stages of the classic Bruce treadmill test; IC, indirect calorimetry.

TABLE 2 Summary of heart rate, core body temperature, step count, and energy expenditure during rest and locomotion phases measured using a consumer-grade and a research-grade method (n = 21).

Variable and condition	Consumer-grade[a]	Research-grade[a]	Consumer minus research		
			95% CI	*P* value	*P* value$_{adj}$
HR					
Sit	80.5 (13.2)	78.7 (9.3)	−1.7 to 5.3	0.97[b]	0.99
Stand	86.6 (10.3)	86.1 (10.5)	−1.0 to 2.0	0.49	0.99
Stage1	100.4 (17.5	103.5 (15.0)	−7.4 to 1.2	0.15	0.62
Stage2	115.7 (18.7)	112.8 (15.7)	−6.9 to 12.5	0.55	0.99
Stage3	128.2 (15.1)	136.3 (14.2)	−15.1 to 0.3	0.06	0.31
Stage4	154.0 (12.4)	165.7 (13.2)	−19.3 to −4.1	0.004	0.026
CBT					
Sit	36.4 (0.8)	37.2 (0.4)	−1.1 to −0.5	<0.001	<0.001
Stand	36.4 (0.5)	37.4 (0.3)	−1.2 to −0.8	<0.001	<0.001
Stage1	36.5 (0.6)	37.4 (0.3)	−1.2 to −0.6	<0.001	<0.001
Stage2	36.4 (0.6)	37.4 (0.4)	−1.3 to −0.7	<0.001	<0.001
Stage3	36.4 (0.6)	37.5 (0.4)	−1.5 to −0.9	<0.001	<0.001
Stage4	36.3 (0.6)	38.1 (0.7)	−2.3 to −1.5	<0.001	<0.001
SC					
Stage1	72.2 (22.1)	71.5 (9.4)	−8.2 to 9.4	0.61[b]	0.61
Stage2	106.3 (5.2)	98.1 (11.4)	3.2 to 13.2	<0.001[b]	0.001
Stage3	131.9 (13.5)	118.0 (11.3)	9.6 to 18.2	<0.001	<0.001
Stage4	152.2 (11.1)	134.9 (4.8)	1.28 to 21.8	<0.001	<0.001
EE$_1$					
Stage1	1.3 (0.5)	3.1 (0.4)	−2.0 to −1.4	<0.001[b]	<0.001
Stage2	2.1 (0.3)	4.6 (0.4)	−2.7 to −2.1	<0.001	<0.001
Stage3	4.1 (2.7)	6.8 (0.7)	−4.0 to −1.6	0.004[b]	0.008
Stage4	7.8 (3.1)	9.7 (0.8)	−3.4 to −0.4	0.015	0.015
EE$_2$					
Stage1	1.0 (0.2)	3.1 (0.4)	−2.3 to −1.9	<0.001	<0.001
Stage2	1.7 (0.2)	4.6 (0.4)	−3.1 to −2.7	<0.001	<0.001
Stage3	3.9 (1.7)	6.8 (0.7)	−3.6 to −2.2	<0.001[b]	<0.001
Stage4	8.3 (1.5)	9.7 (0.8)	−2.2 to −0.6	0.001	0.004

Variable: HR, heart rate (bpm) of Withings and Faros; CBT, core body temperature (°C) of Tucky and Tcore; SC, step count (steps/min) of Withings and GENEActiv; EE$_1$, energy expenditure (MET) of Withings and indirect calorimetry; EE$_2$, energy expenditure (MET) of GENEActiv and indirect calorimetry. Condition: Sit, sitting position; Stand, standing position; Stage1 to Stage4, first four stages of the classic Bruce treadmill test. P value$_{adj}$: P value corrected for multiple comparison.

[a] Mean (SD).

[b] Wilcoxon signed-rank test (otherwise *t*-test).

3.4 Energy expenditure

EE$_{IC}$ increased with each subsequent intensity level of the Bruce test (3.1, 4.6, 6.8 and 9.7 MET for Stage1 to Stage4). Reference EE$_{IC}$ was significantly underestimated by both alternative methods, EE$_{Withings}$ and EE$_{GENEActiv,}$ in each of the four treadmill stages (Figures 3D,E). The bias to IC increased for both methods during the first three stages (up to −2.9 MET). At Stage4, the bias was only −1.9 MET (EE$_{Withings}$) and −1.4 MET (EE$_{GENEActiv}$), but the LoA was widest at this stage. Although the agreement between EE$_{GENEActiv}$ and EE$_{IC}$ appeared to be

TABLE 3 Relationship and agreement between the methods for heart rate, core body temperature, step count, and energy expenditure during rest and locomotion phases (n = 21).

Variable and condition	*r*	LCCC	LoA[a]	MAPE (%)
HR				
Sit	0.82	0.76	1.8 (15.1)	4
Stand	0.95	0.95	0.5 (6.3)	3
Stage1	0.84	0.81	−3.1 (18.6)	7
Stage2	0.24	0.23	2.8 (41.8)	12
Stage3	0.33	0.29	−7.4 (33.3)	11
Stage4	0.16	0.11	−11.7 (32.7)	10
CBT				
Sit	0.53	0.22	−0.8 (1.3)	2
Stand	0.40	0.10	−1.0 (1.0)	3
Stage1	0.23	0.07	−0.9 (1.1)	3
Stage2	0.23	0.07	−1.0 (1.2)	3
Stage3	0.17	0.04	−1.2 (1.4)	3
Stage4	0.19	0.04	−1.8 (1.6)	5
SC				
Stage1	0.48	0.35	0.6 (38.0)	38
Stage2	0.30	0.16	8.2 (21.6)	8
Stage3	0.73	0.43	13.9 (18.4)	10
Stage4	0.48	0.11	17.3 (19.2)	11
EE_1				
Stage1	−0.02	0.00	−1.7 (1.3)	200
Stage2	−0.09	0.00	−2.4 (1.1)	118
Stage3	0.29	0.07	−2.8 (5.1)	113
Stage4	−0.19	−0.07	−1.9 (6.5)	60
EE_2				
Stage1	0.16	0.00	−2.1 (0.8)	228
Stage2	0.25	0.01	−2.9 (0.9)	176
Stage3	0.46	0.09	−2.9 (2.9)	100
Stage4	−0.16	−0.07	−1.4 (3.5)	26

Variable: HR, heart rate (bpm) of Withings and Faros; CBT, core body temperature (°C) of Tucky and Tcore; SC, step count (steps/min) of Withings and GENEActiv; EE_1, energy expenditure (MET) of Withings and indirect calorimetry; EE_2, energy expenditure (MET) of GENEActiv and indirect calorimetry. Condition: Sit, sitting position; Stand, standing position; Stage1 to Stage4, first four stages of the classic Bruce treadmill test. LCCC, Lin's concordance correlation coefficient; MAPE, mean absolute percentage error.

[a] LoA: limits of agreement, bias (1.96SD).

better than between $\text{EE}_{\text{Withings}}$ and EE_{IC}, the agreement between the methods for EE was generally low (Table 3).

4 Discussion

In this study, we measured HR, CBT, SC, and EE during both rest and treadmill phases using reference methods and consumer-grade devices (Withings Pulse HR and Tucky thermometer). We evaluated the accuracy of these parameters against established reference methods (Faros for HR, Tcore for CBT, GENEActiv for SC, and indirect calorimetry for EE). Our results showed that the wrist-worn Withings wearable demonstrated poor agreement or significant differences compared to Faros for HR, indirect calorimetry for EE, and to step-count method using tri-axial acceleration data from GENEActiv. The agreement between

Tucky's rectal equivalent and Tcore's CBT was low at rest and during the treadmill test with significant temperature differences ranging from −1.8 to −0.8°C.

4.1 Comparison with previous work

In a previous validation study of wearables for HR measurement, a LCCC>0.80 was presented as an acceptable accuracy (Gillinov et al., 2017). Accordingly, our results showed that the Withings device demonstrated acceptable agreement with Faros for low physical activities (Sit: LCCC = 0.76, Stand: LCCC = 0.91, Stage1: LCCC = 0.81). The same applies if MAPE threshold is less than 10% (Boudreaux et al., 2018). In our study, MAPE was ≤4% during both resting states and ranged between 7 and 12% during the treadmill locomotion. However, in another study, the device under test was only considered valid if several criteria were met, e.g., LCCC>0.90 and MAPE<5% (Navalta et al., 2020). Furthermore, agreement in HR with the criterion measure during physical activity seems to be lower than during rest, which is in line with previous findings (Thomson et al., 2019; Bent et al., 2020). Devices that use photoplethysmography to monitor HR tend to be inaccurate at higher intensities of physical activity due to artifacts caused by intense hand movements (Castaneda et al., 2018; Bent et al., 2020; Navalta et al., 2020). In addition to motion artefacts from physical activity, ambient light, misalignment between the skin surface, and poor tissue perfusion can also be a source of error (Alzahrani et al., 2015). Skin tone is apparently not a source of errors (Bent et al., 2020), which is an important observation for studies involving African populations, for example. Interestingly, Stahl et al. (2016) observed a decrease in MAPE at treadmill speeds of >3.2 km/h, attributing this to improved perfusion due to increased intensity. In the present study, a small decline in MAPE was observed at treadmill speeds of >4.0 km/h. Nevertheless, not only the user of wrist-worn HR monitor or the ambient conditions seem to affect measurement accuracy, but also the device itself. Müller et al. (2019) investigated the validity of HR measures of a high-cost consumer-based tracker and a low-cost tracker in a laboratory setting, showing the high-cost tracker had smaller errors and a higher agreement with the criterion measure than the low-cost tracker.

For a step counter to be considered accurate, the MAPE should be less than 1% compared to the criterion measure when walking on a treadmill at a speed of 4.8 km/h (Tudor-Locke et al., 2006). In a recent review of the validation of treadmill step-counting technologies, median MAPE values for wrist-worn monitors ranged from 6.6% to 10.7% at speeds between 3.2 and 6.4 km/h (Moore et al., 2020). In our study, the MAPE ranged from 8% to 38% at speeds between 2.7 and 6.7 km/h (Stage1 to Stage4). In addition, the bias was lowest for Stage1 at 0.6 steps/min and highest for Stage4 at 17.3 steps/min, indicating an increasing overestimation in steps by the Withings Pulse HR with increasing treadmill speed. On the other hand, one could argue that estimating steps using a step counting algorithm with tri-axial acceleration data is not a gold standard. Therefore, we compared the estimates in our study with published hand-counted steps from treadmill experiments of Ducharme et al. (2021) and Tudor-Locke et al. (2019) (Supplementary Table S1). It was shown that both the $SC_{Withings}$ and the $SC_{GENEActiv}$ estimated about 17 steps/min less at speed of 2.7 km/h, which was the largest difference compared to published data. Low accuracy of step counting at slow walking speeds is a common issue with wrist-worn wearables (Moore et al., 2020). At treadmill speeds of 5.4 and 6.7 km/h, differences between hand-count $SC_{Withings}$ were about −12 and −18 steps/min, while differences between hand-count and $SC_{GENEActiv}$ were only about 2 and -1 steps/min. These observations suggest a paradox: bias was best at slow walking speed of 2.7 km/h because both wearables were equally inaccurate. Since the use of raw acceleration data provides a flexibility in processing, selecting a better performing step count function should be considered. For example, Ducharme et al. (2021) recently published a transparent algorithm for step detection, and the open-source Verisense step count algorithm has been optimized (Maylor et al., 2022; Rowlands et al., 2022). While Withings Pulse HR utilizes changes in the acceleration caused by foot impact during walking, the exact algorithm is not disclosed.

The $EE_{Withings}$ showed low overall agreement with EE_{IC} during the treadmill test. The same applies to $EE_{GENEActiv}$, where acceleration data from GENEActiv was used to estimate EE using a prediction formula for physical activity energy expenditure (White et al., 2016). In both comparisons, the MAPE value was very high at Stage1 (≥200%), but decreased with increasing treadmill locomotion levels and was lowest in Stage4 (Withings: 60%, GENEActiv: 26%). However, Passler et al. (2019) considered a tested device valid if MAPE is less than 10%. The decrease in MAPE with increasing treadmill speed (and grade) indicates better agreement with higher physical workload. In fact, estimated HR by wrist-worn photoplethysmography devices in combination with physiological modeling tended to have lower MAPE for EE estimation during activities above the aerobic threshold (Parak et al., 2017). Moreover, in the present study both wrist-worn devices for EE estimation clearly underestimated the EE for the criterion measure (indirect calorimetry). Wearable trackers for EE estimation predominantly underestimate EE even in a controlled environment (Evenson et al., 2015; Wahl et al., 2017; Fuller et al., 2020). Wearables were typically examined while worn on the wrist (Fuller et al., 2020), though a greater accuracy can be achieved when placed on the hip or shirt collar (Woodman et al., 2017). EE estimates from devices worn on the wrist or hip generally vary in accuracy depending on physical intensity and type of activity (Howe et al., 2009; O'Driscoll et al., 2020). Recently, Ogata et al. presented an equation to improve EE estimation using accelerometer-based MET value and individual HR and showed that estimated total energy expenditure in rescue workers was one-third higher with the combined approach than with the accelerometer-based method alone (Ogata et al., 2024).

Most wearable thermometers were developed to continuously monitor skin temperature, few in order to reflect changes in CBT (Tamura et al., 2018). In the present study, we compared two sensors attached to the skin: the Tucky thermometer under the right armpit and the Tcore sensor on the forehead. Although adding 0.7°C to the measured values of Tucky improved agreement with rectal temperature, correlations between Tucky's rectal measurements and Tcore's CBT estimate decreased with increased physical activity (highest during Sit and the lowest during Stage4 of the Bruce treadmill test). In addition, the bias in each of the six activity phases was at least −0.8°C, indicating that Tucky's rectal measurements underestimated traditional rectal temperature measurement. For example, Gunga et al. (2008) validated the

Tcore precursor with rectal temperature measurement during treadmill activities (25%–55% maximum work intensity) at different ambient temperatures, demonstrating a good agreement during resting (Bias: 0.01°C, LoA: 0.74 to 0.72) and working periods (Bias: 0.08°C, LoA: 0.77 to 0.61) at ambient temperature of 25°C. Wearable thermometers are considered in agreement if they comply with the clinically meaningful recommendations of bias of ±0.5°C and LoA of ±1.0°C (Tamura et al., 2018). In our study, however, the bias was at least −0.8°C and the LoA were −2 to 0°C at Stand (narrowest) and −3.5°C to 0.2°C at Stage4 (widest). Our results suggest that the higher the intensity of physical activity, the lower the accuracy of Tucky's measurements. This inaccuracy could be attributed to the thermoregulatory processes of the skin. Increased physical activity can lead to increased perspiration, which aims to cool the skin and CBT through evaporation. In the context of varying and intensive physical activity, Tucky under the armpit did not achieve sufficient accuracy with CBT. Similar observation was reported for another adesive axillary thermomenter patch. Temperatures of adesive axillary thermomenter showed good agreement with those from the conventional axillary method (Bias: 0.15°C, LoA: 1.13 to 0.99), but failed to those of the bladder as the CBT (Bias: 1.11°C, LoA: 3.19 to 0.98) (Boyer et al., 2021).

4.2 Strength and limitations

This study has several strengths. Firstly, we investigated two devices, Withings Pulse HR and Tucky thermometer, that had not been validated in an independent lab study previously, focusing on their utility for in-field assessment of physiological variables in different situations of varying physical activity. Therefore, a structured protocol consisting of successively changing intensities of activity was implemented. A structured procedure and laboratory-based setting enabled a high precision of comparison and reproducibility of results.

This study was limited to healthy, fair-skinned adults aged 20–29 years. Future research should include a more diverse cohort and a comparison of multiple skin tones, especially when using optical heart rate monitors. Motion that largely affects positioning of wearables, such as treadmill running for a wrist-worn tracker, may impact accuracy and the significance of validation research. Potential interference between devices worn simultaneously on the same wrist might also represent a possible limitation of the study. Additionally, although treadmill-based incremental testing can represent the cardiovascular strain of physical activity during agricultural work, it does not correspond to the actual biomechanics and motions of such physical activity.

5 Conclusion

In recent years, research interest in consumer-grade wearables has surged, driven by the potential of these sensors for a broad range of applications, from on-the-field ergonomic assessments to follow-ups in rehabilitation medicine. In this study, we evaluated the Withings Pulse HR wearable for HR, SC, and EE quantification and the Tucky thermometer for CBT. The Withings device demonstrated good performance in HR monitoring at low physical activity intensities and in SC at higher activity levels. However, the agreement between the Tucky thermometer measured temperature and CBT was low at rest and gradually declined with increased physical activity. In summary, both evaluated consumer-grade wearables did not achieve adequate accuracy for research purposes in controlled environments. However, Withings Pulse HR may be useful for long-term monitoring in the field, as it can effectively detect and recognize general changes in activity and corresponding physiological variables (HR, SC, EE) despite its lack of precision.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Ethics Committee of Charité–Universitätsmedizin Berlin. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

SM: Data curation, Formal Analysis, Visualization, Writing–original draft, Writing–review and editing. GZ: Investigation, Methodology, Writing–original draft. MR: Formal Analysis, Writing–review and editing. H-CG: Funding acquisition, Writing–review and editing. AB: Writing–review and editing. SB: Conceptualization, Resources, Writing–review and editing. MAM: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The study was part of the German Research Foundation (DFG) funded research unit “Climate Change and Health in sub-Saharan Africa” (FOR 2936), specifically individual project “Climate change, heat stress and their impact on health and working capacity” (DFG Grant number 660477). Authors SM, H-CG and MAM acknowledge the support of the German Aerospace Center -Deutsches Zentrum für Luft- und Raumfahrt (DLR) through the grants 50WB2117, 50WB2030 and 50WB2330. We also acknowledge support from the Open Access Publication Fund of Charité–Universitätsmedizin Berlin.

Acknowledgments

We would like to acknowledge the Charité medical students for their voluntary participation in and dedication to this study. Author MR was supported by Ricerca Corrente, Ministero della Salute.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2025.1491401/full#supplementary-material

References

Alzahrani, A., Hu, S., Azorin-Peris, V., Barrett, L., Esliger, D., Hayes, M., et al. (2015). A multi-channel opto-electronic sensor to accurately monitor heart rate against motion artefact during exercise. *Sensors (Basel)* 15, 25681–25702. doi:10.3390/s151025681

Antczak, D., Lonsdale, C., Del Pozo Cruz, B., Parker, P., and Sanders, T. (2021). Reliability of GENEActiv accelerometers to estimate sleep, physical activity, and sedentary time in children. *Int. J. Behav. Nutr. Phys. Act.* 18, 73. doi:10.1186/s12966-021-01143-6

Asare-Nuamah, P. (2021). Climate variability, subsistence agriculture and household food security in rural Ghana. *Heliyon* 7, e06928. doi:10.1016/j.heliyon.2021.e06928

Ayal, D. Y. (2021). Climate change and human heat stress exposure in sub-Saharan Africa. *CABI. Reviews.* doi:10.1079/PAVSNNR202116049

Barteit, S., Boudo, V., Ouedraogo, A., Zabré, P., Ouremi, L., Sié, A., et al. (2021). Feasibility, acceptability and validation of wearable devices for climate change and health research in the low-resource contexts of Burkina Faso and Kenya: study protocol. *PLoS One* 16, e0257170. doi:10.1371/journal.pone.0257170

Barteit, S., Sié, A., Zabré, P., Traoré, I., Ouédraogo, W. A., Boudo, V., et al. (2023). Widening the lens of population-based health research to climate change impacts and adaptation: the climate change and health evaluation and response system (CHEERS). *Front. Public. Health.* 11, 1153559. doi:10.3389/fpubh.2023.1153559

Bent, B., Goldstein, B. A., Kibbe, W. A., and Dunn, J. P. (2020). Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ Digit. Med.* 3, 18. doi:10.1038/s41746-020-0226-6

Bland, J. M., and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1, 307–310. doi:10.1016/S0140-6736(86)90837-8

Boudreaux, B. D., Hebert, E. P., Hollander, D. B., Williams, B. M., Cormier, C. L., Naquin, M. R., et al. (2018). Validity of wearable activity monitors during cycling and resistance exercise. *Med. Sci. Sports Exerc* 50, 624–633. doi:10.1249/MSS.0000000000001471

Bouchama, A., Abuyassin, B., Lehe, C., Laitano, O., Jay, O., O'Connor, F. G., et al. (2022). Classic and exertional heatstroke. *Nat. Rev. Dis. Primers.* 8, 8. doi:10.1038/s41572-021-00334-6

Boyer, J., Eckmann, J., Strohmayer, K., Koele, W., Federspiel, M., Schenk, M., et al. (2021). Investigation of non-invasive continuous body temperature measurements in a clinical setting using an adhesive axillary thermometer (SteadyTemp®). *Front. Digit. Health* 3, 794274. doi:10.3389/FDGTH.2021.794274

Caldwell, A. R. (2022). SimplyAgree: an R package and jamovi module for simplifying agreement and reliability analyses. *J. Open Source Softw.* 7, 4148. doi:10.21105/joss.04148

Campbell, C., Gott, A., Langford, J., Sweetland, C., and Sweetland, P. (2023). GENEAclassify: segmentation and classification of accelerometer data. Available at: https://cran.r-project.org/package=GENEAclassify.

Castaneda, D., Esparza, A., Ghamari, M., Soltanpur, C., and Nazeran, H. (2018). A review on wearable photoplethysmography sensors and their potential future applications in health care. *Int. J. Biosens. Bioelectron.* 4, 195–202. doi:10.15406/ijbsbe.2018.04.00125

Chavaillaz, Y., Roy, P., Partanen, A.-I., Da Silva, L., Bresson, É., Mengis, N., et al. (2019). Exposure to excessive heat and impacts on labour productivity linked to cumulative CO2 emissions. *Sci. Rep.* 9, 13711. doi:10.1038/s41598-019-50047-w

Diboulo, E., Sié, A., Rocklöv, J., Niamba, L., Yé, M., Bagagnan, C., et al. (2012). Weather and mortality: a 10 year retrospective analysis of the Nouna Health and demographic surveillance system, burkina faso. *Glob. Health. Act.* 5 , 6 –13 . doi:10.3402/gha.v5i0.19078

Ducharme, S. W., Lim, J., Busa, M. A., Aguiar, E. J., Moore, C. C., Schuna, J. M., et al. (2021). A transparent method for step detection using an acceleration threshold. *J. Meas. Phys. Behav.* 4, 311–320. doi:10.1123/jmpb.2021-0011

Dunn, J., Runge, R., and Snyder, M. (2018). Wearables and the medical revolution. *Per. Med.* 15, 429–448. doi:10.2217/pme-2018-0044

Egondi, T., Kyobutungi, C., Kovats, S., Muindi, K., Ettarh, R., and Rocklöv, J. (2012). Time-series analysis of weather and mortality patterns in Nairobi's informal settlements. *Glob. Health. Act.* 5, 23–32. doi:10.3402/gha.v5i0.19065

Eitelwein, O., Fricker, R., Green, A., and Racloz, V. (2024) . Quantifying the impact of climate change on human health. Available at: https://www3.weforum.org/docs/WEF_Quantifying_the_Impact_of_Climate_Change_on_Human_Health_2024.pdf.

Eisenkraft, A., Goldstein, N., Fons, M., Tabi, M., Sherman, A. D., Ben Ishay, A., et al. (2023). Comparing body temperature measurements using the double sensor method within a wearable device with oral and core body temperature measurements using medical grade thermometers—a short report. *Front. Physiol.* 14. doi:10.3389/FPHYS.2023.1279314/FULL

Engelbart, G., Brandt, S., Scheeren, T., Tzabazis, A., Kimberger, O., and Kellner, P. (2023). Accuracy of non-invasive sensors measuring core body temperature in cardiac surgery ICU patients – results from a monocentric prospective observational study. *J. Clin. Monit. Comput.* 37, 1619–1626. doi:10.1007/s10877-023-01049-7

Evenson, K. R., Goto, M. M., and Furberg, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. *Int. J. Behav. Nutr. Phys. Act.* 12, 159. doi:10.1186/s12966-015-0314-1

Fletcher, G. F., Ades, P. A., Kligfield, P., Arena, R., Balady, G. J., Bittner, V. A., et al. (2013). Exercise standards for testing and training: a scientific statement from the American Heart Association. *Circulation* 128, 873–934. doi:10.1161/CIR.0b013e31829b5b44

Fraysse, F., Post, D., Eston, R., Kasai, D., Rowlands, A. V., and Parfitt, G. (2020). Physical activity intensity cut-points for wrist-worn GENEActiv in older adults. *Front. Sports Act. Living* 2, 579278. doi:10.3389/fspor.2020.579278

Frimpong, K., Odonkor, S. T., Kuranchie, F. A., and Nunfam, V. F. (2020). Evaluation of heat stress impacts and adaptations: perspectives from smallholder rural farmers in Bawku East of Northern Ghana. *Heliyon* 6, e03679. doi:10.1016/j.heliyon.2020.e03679

Fuller, D., Colwell, E., Low, J., Orychock, K., Tobin, M. A., Simango, B., et al. (2020). Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. *JMIR Mhealth Uhealth* 8, e18694. doi:10.2196/18694

Funston, R., Gibbs, A., Diven, J., Francey, J., Easlea, H., Murray, S., et al. (2022). Comparative study of a single lead ECG in a wearable device. *J. Electrocardiol.* 74, 88–93. doi:10.1016/j.jelectrocard.2022.08.004

Gillinov, S., Etiwy, M., Wang, R., Blackburn, G., Phelan, D., Gillinov, A. M., et al. (2017). Variable accuracy of wearable heart rate monitors during aerobic exercise. *Med. Sci. Sports Exerc* 49, 1697–1703. doi:10.1249/MSS.0000000000001284

Godino, J. G., Wing, D., de Zambotti, M., Baker, F. C., Bagot, K., Inkelis, S., et al. (2020). Performance of a commercial multi-sensor wearable (Fitbit Charge HR) in measuring physical activity and sleep in healthy children. *PLoS One* 15, e0237719. doi:10.1371/JOURNAL.PONE.0237719

Gunga, H. C., Sandsund, M., Reinertsen, R. E., Sattler, F., and Koch, J. (2008). A non-invasive device to continuously determine heat strain in humans. *J. Therm. Biol.* 33, 297–307. doi:10.1016/j.jtherbio.2008.03.004

Hachenberger, J., Teuber, Z., Li, Y.-M., Abkai, L., Wild, E., and Lemola, S. (2023). Investigating associations between physical activity, stress experience, and affective wellbeing during an examination period using experience sampling and accelerometry. *Sci. Rep.* 13, 8808. doi:10.1038/s41598-023-35987-8

Hartikainen, S., Lipponen, J. A., Hiltunen, P., Rissanen, T. T., Kolk, I., Tarvainen, M. P., et al. (2019). Effectiveness of the chest strap electrocardiogram to detect atrial fibrillation. *Am. J. Cardiol.* 123, 1643–1648. doi:10.1016/j.amjcard.2019.02.028

Howe, C. A., Staudenmayer, J. W., and Freedson, P. S. (2009). Accelerometer prediction of energy expenditure: vector magnitude versus vertical axis. *Med. Sci. Sports Exerc* 41, 2199–2206. doi:10.1249/MSS.0b013e3181aa3a0e

Huhn, S., Axt, M., Gunga, H.-C., Maggioni, M. A., Munga, S., Obor, D., et al. (2022). The impact of wearable technologies in health research: scoping review. *JMIR Mhealth Uhealth* 10, e34384. doi:10.2196/34384

Janke, D., Kagelmann, N., Storm, C., Maggioni, M. A., Kienast, C., Gunga, H.-C., et al. (2021). Measuring core body temperature using a non-invasive, disposable double-sensor during targeted temperature management in post-cardiac arrest patients. *Front. Med. (Lausanne)* 8, 666908. doi:10.3389/fmed.2021.666908

Jenkins, C. A., Tiley, L. C. F., Lay, I., Hartmann, J. A., Chan, J. K. M., and Nicholas, C. L. (2022). Comparing GENEActiv against actiwatch-2 over seven nights using a common sleep scoring algorithm and device-specific wake thresholds. *Behav. Sleep. Med.* 20, 369–379. doi:10.1080/15402002.2021.1924175

Katiyatiya, C. L. F., Muchenje, V., and Mushunje, A. (2014). Farmers' perceptions and knowledge of cattle adaptation to heat stress and tick resistance in the eastern cape, South Africa. *Asian-Australas. J. Anim. Sci.* 27, 1663–1670. doi:10.5713/ajas.2014.141

Kjellstrom, T., Holmer, I., and Lemke, B. (2009). Workplace heat stress, health and productivity - an increasing challenge for low and middle-income countries during climate change. *Glob. Health. Act.* 2, 2047. doi:10.3402/gha.v2i0.2047

Koch, M., Matzke, I., Huhn, S., Gunga, H.-C., Maggioni, M. A., Munga, S., et al. (2022). Wearables for measuring health effects of climate change-induced weather extremes: scoping review. *JMIR Mhealth Uhealth* 10, e39532. doi:10.2196/39532

Kokkinos, P., Kaminsky, L. A., Arena, R., Zhang, J., and Myers, J. (2017). New generalized equation for predicting maximal oxygen uptake (from the fitness registry and the importance of exercise national database). *Am. J. Cardiol.* 120, 688–692. doi:10.1016/J.AMJCARD.2017.05.037

Laborde, S., Mosley, E., and Thayer, J. F. (2017). Heart rate variability and cardiac vagal tone in psychophysiological research - recommendations for experiment planning, data analysis, and data reporting. *Front. Psychol.* 8, 213. doi:10.3389/fpsyg.2017.00213

Lang, M., Mendt, S., Paéz, V., Gunga, H.-C., Bilo, G., Merati, G., et al. (2022). Cardiac autonomic modulation and response to sub-maximal exercise in Chilean hypertensive miners. *Front. Physiol.* 13, 846891. doi:10.3389/fphys.2022.846891

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45, 255–268. doi:10.2307/2532051

Matzke, I., Huhn, S., Koch, M., Maggioni, M. A., Munga, S., Muma, J. O., et al. (2024). Assessment of heat exposure and health outcomes in rural populations of western Kenya by using wearable devices: observational Case study. *JMIR Mhealth Uhealth* 12, e54669. doi:10.2196/54669

Maylor, B. D., Edwardson, C. L., Dempsey, P. C., Patterson, M. R., Plekhanova, T., Yates, T., et al. (2022). Stepping towards more intuitive physical activity metrics with wrist-worn accelerometry: validity of an open-source step-count algorithm. *Sensors* 22, 9984. doi:10.3390/s22249984

Masson-Delmotte, V., Zhai, P., Pörtner, H.-O., Roberts, D., Skea, J., Shukla, P., et al. (2018). *Global warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change*. Cambridge University Press. doi:10.1017/9781009157940

Mendt, S., Maggioni, M. A., Nordine, M., Steinach, M., Opatz, O., Belavý, D., et al. (2017). Circadian rhythms in bed rest: monitoring core body temperature via heat-flux approach is superior to skin surface temperature. *Chronobiol Int.* 34, 666–676. doi:10.1080/07420528.2016.1224241

Moore, C. C., McCullough, A. K., Aguiar, E. J., Ducharme, S. W., and Tudor-Locke, C. (2020). Toward harmonized treadmill-based validation of step-counting wearable technologies: a scoping review. *J. Phys. Act. Health* 17, 840–852. doi:10.1123/jpah.2019-0205

Müller, A. M., Wang, N. X., Yao, J., Tan, C. S., Low, I. C. C., Lim, N., et al. (2019). Heart rate measures from wrist-worn activity trackers in a laboratory and free-living setting: validation study. *JMIR Mhealth Uhealth* 7, e14120. doi:10.2196/14120

Navalta, J. W., Montes, J., Bodell, N. G., Salatto, R. W., Manning, J. W., and DeBeliso, M. (2020). Concurrent heart rate validity of wearable technology devices during trail running. *PLoS One* 15, e0238569. doi:10.1371/journal.pone.0238569

Ndahimana, D., and Kim, E.-K. (2017). Measurement methods for physical activity and energy expenditure: a review. *Clin. Nutr. Res.* 6, 68–80. doi:10.7762/cnr.2017.6.2.68

Nelson, M. B., Kaminsky, L. A., Dickin, D. C., and Montoye, A. H. K. (2016). Validity of consumer-based physical activity monitors for specific activity types. *Med. Sci. Sports. Exerc.* 48, 1619–1628. doi:10.1249/MSS.0000000000000933

O'Driscoll, R., Turicchi, J., Beaulieu, K., Scott, S., Matu, J., Deighton, K., et al. (2020). How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies. *Br. J. Sports Med.* 54, 332–340. doi:10.1136/bjsports-2018-099643

Ogata, H., Negishi, Y., Koizumi, N., Nagayama, H., Kaneko, M., Kiyono, K., et al. (2024). Individually optimized estimation of energy expenditure in rescue workers using a tri-axial accelerometer and heart rate monitor. *Front. Physiol.* 15, 1322881. doi:10.3389/fphys.2024.1322881

Parak, J., Uuskoski, M., Machek, J., and Korhonen, I. (2017). Estimating heart rate, energy expenditure, and physical performance with a wrist photoplethysmographic device during running. *JMIR Mhealth Uhealth* 5, e97. doi:10.2196/mhealth.7437

Park, J., Bangalore, M., Hallegatte, S., and Sandhoefner, E. (2018). Households and heat stress: estimating the distributional consequences of climate change. *Environ. Dev. Econ.* 23, 349–368. doi:10.1017/S1355770X1800013X

Passler, S., Bohrer, J., Blöchinger, L., and Senner, V. (2019). Validity of wrist-worn activity trackers for estimating VO2max and energy expenditure. *Int. J. Environ. Res. Public Health* 16, 3037. doi:10.3390/ijerph16173037

Pickett, W., King, N., Lawson, J., Dosman, J., Trask, C., Brison, R. J., et al. (2015). Farmers, mechanized work, and links to obesity. *Prev. Med. Balt.* 70, 59–63. doi:10.1016/J.YPMED.2014.11.012

R Core Team (2022). R: A Language and Environment for Statistical Computing. ,Vienna, Austria. Available at: https://www.R-project.org/ (Accessed February 29, 2024).

Riebe, D. (2014). "General principles of exercise prescription," in *ACSM's guidelines for exercise testing and prescription*. Editor L. S. Pescatello (Philadelphia: Wolters Kluwer/Lippincott Williams and Wilkins Health), 161–193.

Rowlands, A. V., Maylor, B., Dawkins, N. P., Dempsey, P. C., Edwardson, C. L., Soczawa-Stronczyk, A. A., et al. (2022). Stepping up with GGIR: validity of step cadence derived from wrist-worn research-grade accelerometers using the verisense step count algorithm. *J. Sports Sci.* 40, 2182–2190. doi:10.1080/02640414.2022.2147134

Sanders, S. G., Jimenez, E. Y., Cole, N. H., Kuhlemeier, A., McCauley, G. L., Van Horn, M. L., et al. (2019). Estimated physical activity in adolescents by wrist-worn GENEActiv accelerometers. *J. Phys. Act. Health* 16, 792–798. doi:10.1123/jpah.2018-0344

Sapari, H., Selamat, M. I., Isa, M. R., Ismail, R., and Wan Mahiyuddin, W. R. (2023). The impact of heat waves on health care services in Low- or middle-income countries: protocol for a systematic review. *JMIR. Res. Protoc.* 12, e44702. doi:10.2196/44702

Scott, J. J., Rowlands, A. V., Cliff, D. P., Morgan, P. J., Plotnikoff, R. C., and Lubans, D. R. (2017). Comparability and feasibility of wrist- and hip-worn accelerometers in free-living adolescents. *J. Sci. Med. Sport* 20, 1101–1106. doi:10.1016/j.jsams.2017.04.017

Soehle, M., Dehne, H., Hoeft, A., and Zenker, S. (2020). Accuracy of the non-invasive Tcore^{TM} temperature monitoring system to measure body core temperature in abdominal surgery. *J. Clin. Monit. Comput.* 34, 1361–1367. doi:10.1007/s10877-019-00430-9

Stahl, S. E., An, H.-S., Dinkel, D. M., Noble, J. M., and Lee, J.-M. (2016). How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough? *BMJ Open Sport Exerc Med.* 2, e000106. doi:10.1136/bmjsem-2015-000106

Tamura, T., Huang, M., and Togawa, T. (2018). Current developments in wearable thermometers. *Adv. Biomed. Eng.* 7, 88–99. doi:10.14326/abe.7.88

Thiebaud, R. S., Funk, M. D., Patton, J. C., Massey, B. L., Shay, T. E., Schmidt, M. G., et al. (2018). Validity of wrist-worn consumer products to measure heart rate and energy expenditure. *Digit. Health* 4, 2055207618770322. doi:10.1177/2055207618770322

Thomson, E. A., Nuss, K., Comstock, A., Reinwald, S., Blake, S., Pimentel, R. E., et al. (2019). Heart rate measures from the Apple Watch, Fitbit Charge HR 2, and electrocardiogram across different exercise intensities. *J. Sports Sci.* 37, 1411–1419. doi:10.1080/02640414.2018.1560644

Tudor-Locke, C., Aguiar, E. J., Han, H., Ducharme, S. W., Schuna, J. M., Barreira, T. V., et al. (2019). Walking cadence (steps/min) and intensity in 21–40 year olds: CADENCE-adults. *Int. J. Behav. Nutr. Phys. Activity* 16, 8. doi:10.1186/s12966-019-0769-6

Tudor-Locke, C., Sisson, S. B., Lee, S. M., Craig, C. L., Plotnikoff, R. C., and Bauman, A. (2006). Evaluation of quality of commercial pedometers. *Can. J. Public Health* 97 (Suppl. 1), S10–S16. doi:10.1007/BF03405359

Wahl, Y., Düking, P., Droszez, A., Wahl, P., and Mester, J. (2017). Criterion-validity of commercially available physical activity tracker to estimate step count, covered distance and energy expenditure during sports conditions. *Front. Physiol.* 8, 725. doi:10.3389/fphys.2017.00725

Werner, A., and Gunga, H.-C. (2020). "Monitoring of core body temperature in humans," in *Stress challenges and immunity in space*. Editor A. Choukèr (Cham: Springer International Publishing), 477–498. doi:10.1007/978-3-030-16996-1_26

White, T., Westgate, K., Wareham, N. J., and Brage, S. (2016). Estimation of physical activity energy expenditure during free-living from wrist accelerometry in UK adults. *PLoS One* 11, e0167472. doi:10.1371/journal.pone.0167472

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Cham: Springer International Publishing. doi:10.1007/978-3-319-24277-4

Woodman, J. A., Crouter, S. E., Bassett, D. R., Fitzhugh, E. C., and Boyer, W. R. (2017). Accuracy of consumer monitors for estimating energy expenditure and activity type. *Med. Sci. Sports Exerc* 49, 371–377. doi:10.1249/MSS.0000000000001090

World Health Organization (2023). *Clim. change*. Available at: https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health (Accessed July 13, 2024).

Yengoh, G. T., and Ardö, J. (2020). Climate Change and the Future Heat Stress Challenges among Smallholder Farmers in East Africa. *Atmosphere (Basel).* 11, 753. doi:10.3390/atmos11070753

frontiers | Frontiers in Physiology

TYPE Original Research
PUBLISHED 05 March 2025
DOI 10.3389/fphys.2025.1480018

Distress detection in VR environment using Empatica E4 wristband and Bittium Faros 360

Jelena Medarević[1]*, Nadica Miljković[1,2], Kristina Stojmenova Pečečnik[1] and Jaka Sodnik[1]

[1]Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia, [2]School of Electrical Engineering, University of Belgrade, Belgrade, Serbia

Check for updates

OPEN ACCESS

EDITED BY
Colin K. Drummond,
Case Western Reserve University,
United States

REVIEWED BY
Adolphe J. Béquet,
Catholic University of Toulouse, France
Kuan Tao,
Beijing Sport University, China
Jonathan Giron,
Reichman University, Israel

*CORRESPONDENCE
Jelena Medarević,
✉ jm98847@student.uni-lj.si

RECEIVED 13 August 2024
ACCEPTED 10 February 2025
PUBLISHED 05 March 2025

CITATION
Medarević J, Miljković N, Stojmenova Pečečnik K and Sodnik J (2025) Distress detection in VR environment using Empatica E4 wristband and Bittium Faros 360.
Front. Physiol. 16:1480018.
doi: 10.3389/fphys.2025.1480018

Introduction: Distress detection in virtual reality systems offers a wealth of opportunities to improve user experiences and enhance therapeutic practices by catering to individual physiological and emotional states.

Methods: This study evaluates the performance of two wearable devices, the Empatica E4 wristband and the Faros 360, in detecting distress in a motion-controlled interactive virtual reality environment. Subjects were exposed to a baseline measurement and two VR scenes, one non-interactive and one interactive, involving problem-solving and distractors. Heart rate measurements from both devices, including mean heart rate, root mean square of successive differences, and subject-specific thresholds, were utilized to explore distress intensity and frequency.

Results: Both the Faros and E4 sensors adequately captured physiological signals, with Faros demonstrating a higher signal-to-noise ratio and consistency. While correlation coefficients were moderately positive between Faros and E4 data, indicating a linear relationship, small mean absolute error and root mean square error values suggested good agreement in measuring heart rate. Analysis of distress occurrence during the interactive scene revealed that both devices detect more high- and medium-level distress occurrences compared to the non-interactive scene.

Discussion: Device-specific factors in distress detection were emphasized due to differences in detected distress events between devices.

KEYWORDS

virtual reality, user experience, wearables, Empatica E4, Faros 360, distress detection, mean heart rate, RMSSD

1 Introduction

Virtual Reality (VR) environments have gained significant popularity in recent years, offering immersive and interactive experiences that can simulate realistic scenarios. Alongside the visual and auditory components, the measurement within VR environments can provide a deeper understanding of human responses and experiences. By capturing physiological signals such as heart rate (HR), electrodermal activity (EDA), and motion data (MD) like acceleration, researchers can explore the correlations of user engagement, emotional states, cognitive processes, and user experiences during VR interactions (Egan et al., 2016). VR's ability to create a strong sensation of being physically present in the virtual environment and the perception that virtual events are genuinely occurring ensures that users react to virtual scenarios as they would in real-life and collectively contribute to overall sense of presence in virtual environment (Slater et al., 2022).

The information collected with distress detection in VR systems has the potential to enhance user experience and holds promising implications across various fields. In VR therapy, it can be used to monitor and regulate patients' emotional states during exposure sessions (Rahman et al., 2022), such as in the treatment of phobias (Raghav et al., 2016) or Post-Traumatic Stress Disorder (PTSD) (Wout et al., 2017). Physiological monitoring systems, detecting indicators like increased heart rate (Rahman et al., 2022) and skin conductance (Wout et al., 2017), allow therapists to dynamically adjust virtual environments in real-time, optimizing therapy based on individual needs. In training simulations, particularly in high-pressure scenarios such as medical (Rahman et al., 2022) or military (Wout et al., 2017) training, distress detection becomes a valuable tool. By evaluating trainees' stress levels, VR systems can identify areas requiring additional support or practice (Parsons and Reinebold, 2012). Beyond therapy and training, distress detection contributes significantly to human-computer interaction, enabling VR systems to adapt the presented content based on the users' emotional states for more natural and intuitive interactions (Duric et al., 2002). Additionally, VR systems equipped with physiological sensors can offer stress management experiences, providing guided meditation or calming environments that respond to users' distress levels, creating a feedback loop to enhance relaxation (Gromala et al., 2015). The field of Neuroergonomics, which examines brain function in real-world environments, offers another potential application of distress detection in VR, particularly in optimizing user-performance in safety-critical professions (Parasuraman, 2003).

One of the primary physiological signals used for distress detection is heart rate, which measures the number of heart beats per minute. Heightened distress or emotional responses can lead to changes in HR (Mack et al., 2006), making it a fundamental parameter in assessing distress levels during VR experiences (Robitaille and McGuffin, 2019). Recent research suggests that there might be subtle differences in HR patterns between males and females. Studies have indicated that females tend to exhibit slightly higher average resting HRs compared to males, which could be attributed to hormonal and physiological variations between the sexes (Altini and Plews, 2021; Quer et al., 2020). Beyond sex-related distinctions, HR is influenced by a variety of factors, including physical activity levels, stressors, emotional states (Wu et al., 2019), fatigue (Tran et al., 2009), and even caffeine consumption (Koenig et al., 2013) and environmental conditions (Tiwari et al., 2021).

For reliable distress detection, the non-intrusivity of measurement devices is paramount. In VR, user immersion and experience are crucial, and intrusive devices can compromise data accuracy and user comfort. Wearable sensors and cameras provide non-intrusive data acquisition (Heikenfeld et al., 2018), preserving the naturalness of the VR experience and encouraging user compliance. In this context, we decided to utilize Empatica E4 (E4) wristband and Faros 360 chest strap (Faros), particularly as in the previous research we already validated E4 against Faros (Gruden et al., 2019). This study focused on evaluating E4 and Faros 360 devices in assessing drivers' physiological responses during various driving conditions, emphasizing their effectiveness in measuring heart rate variability (HRV) and EDA, but noting challenges with motion artifacts affecting data quality, particularly in distinguishing different driving demands. However, it showed that the user-friendly nature of E4 sets it apart in experimental settings, offering easy mounting and usage—crucial factors when subjects are engaged in multitasking scenarios requiring sustained focus. Such non-intrusive nature of E4 wristband ensures seamless data collection (Heikenfeld et al., 2018) contributing to the authenticity and reliability of physiological responses in virtual reality settings.

However, it is important to note that results of (McCarthy et al., 2016) point out the low data quality of physiological signals obtained using E4 due to motion artefacts, especially the Blood Volume Pulse (BVP) signal, often used to estimate the HR signal. Even though Empatica released a new device – (Empatica Embrace Plus, 2024), we chose E4 device as its data is easily accessible (Looff et al., 2022), since the new device does not provide the API or access to the raw data in real time anymore. With the new device the data collection should be performed through a proprietary Empatica app and web server which is not ideal for research purposes, but it is worth noting that, for offline purposes, researchers can access raw.*avro* files from the server, and if needed, convert it to *.csv* format (Béquet, 2023). Furthermore, E4 device is still present on the market and majority of researchers still use it. One of Faros' main limitations is direct-skin placement, which may be uncomfortable for some subjects, especially those with skin sensitivities or those requiring prolonged wear. Adhesive reactions, pressure from the chest strap, and sweating can impact user comfort and compliance. Its placement makes it less practical in applied settings like workplace monitoring, and its requirement to have precise sensor placement adds to setup complexity and potentially affects data quality.

Through the E4 measurement evaluation and comparison with Faros, this paper explores the importance of reliable data acquisition in VR environments for distress detection. Our proposed method for distress detection involves a straightforward thresholding approach and a rule-based system, contributing to the precision and efficiency of the analysis. The method uses distress detection thresholds that are subject-specific in order to tailor the method to each subject's unique physiological profile.

Subjects were exposed to a baseline measurement and two VR scenes–a non-interactive scene (NIS) in which the subjects observed nature, and an interactive scene (IS) with distress induction in which the subjects were required to solve the Hanoi tower problem using a VR controller while being surprised with various distractors.

The main research questions addressed in this study are as follows:

1. Can both the Faros and E4 devices effectively detect distress in individuals in IS?
2. What is the level of data quality (determined through level of noise contamination) achieved by the E4 device when measuring heart rate used for distress detection?

The first research question explores firstly the possibility of using Faros/E4 in distress intensity and frequency detection based on the HR parameters, and secondly also the E4 performance compared to

Faros. Only responses collected in the IS are considered since this scene is created for the purpose of eliciting distress in test subjects.

The second question is primarily focused on validating the performance and data quality of the E4 device assessed by evaluating the level of noise present in the heart rate measurements used for distress detection. The goal is to confirm if E4 can emerge as a user-friendly option for future studies, by exploring whether its heart rate measurements exhibit sufficiently high data quality, ensuring that distress detection remains unaffected. Baseline measurements, NIS and IS data is considered, since it is expected that the data quality is high for each measurement.

In summary, these research questions form the core inquiries guiding the investigation, aiming to assess the accuracy and potential of the E4 device and the comparative performance of the Faros and E4 devices in distress detection.

2 Materials and methods

In this section, the experimental design and procedures used to investigate physiological responses in a motion-controlled virtual environment are outlined. Two commercially available devices, Empatica E4 wristband and Faros, which were utilized to measure and record various physiological parameters during the study, are introduced. Before the experiment description, an overview of each device's capabilities and functionalities is provided, setting the context for their use in this research. Subsequently, the experiment design is presented, along with subject details, and the VR scenes employed to capture physiological data.

2.1 Empatica E4 wristband device

Empatica E4 wristband (shown in Figure 1) is a commercially available physiological monitoring device (Empatica, 2024). It is equipped with several sensors that enable the measurement of multiple physiological parameters. Using photoplethysmography (PPG) it can capture the Blood Volume Pulse, which provides information on changes in blood volume in the microvascular bed, allowing for the estimation of HR and inter-beat interval (IBI) (Allen, 2007). Additionally, the device includes an EDA sensor, which measures changes in the electrical conductance of the skin, reflecting the user's sympathetic nervous system activity and emotional responses (Boucsein, 2012). Moreover, the E4 wristband incorporates a temperature sensor, enabling the monitoring of skin temperature variations, a 3-axis accelerometer that measures acceleration in three directions, enabling the detection of motion and physical activity that can help researchers understand the subjects' movements and activity levels during data collection.

Utilizing E4 wristband within a motion-controlled VR environment offers several advantages, particularly in the context of capturing physiological responses. The devices' less obtrusive and user-friendly nature allows seamless data collection without disrupting subjects' experiences in the VR scenarios. Being worn on the wrist, it enables continuous monitoring and real-time data transmission, making it well-suited for prolonged data collection during immersive VR sessions.

However, it is crucial to address potential limitations associated with the E4 wristband. Subjects' activities during VR sessions may introduce motion artifacts and affect data quality. To mitigate this issue, careful measures were taken to control subjects' motion during data collection, ensuring more accurate and reliable physiological measurements (Böttcher et al., 2022). Considering the context of our study, the E4 wristband ease of use, portability, and compatibility with VR scenarios make it an appropriate choice for HR.

2.2 Faros 360 device

Faros 360 (shown in Figure 2) is a commercially available physiological monitoring device designed for electrocardiographic (ECG) signal recording. It allows the measurement of the electrical activity of the heart, providing information on heart rate and cardiac rhythm. The device is equipped with high-quality ECG sensors that enable accurate and reliable data collection, and enables 3-channels ECG measurement and data streaming via Bluetooth (Bittium Faros, 2024).

This device focuses on ECG measurements and high-quality ECG signals which makes it well-suited for providing precise HR data, crucial in understanding subjects' cardiovascular responses during VR scenes in a motion-controlled VR environment.

However, it is essential to consider the limitations associated with Faros application in the VR context. The device is not capable of capturing other physiological parameters, such as EDA or skin temperature, which can also provide important information about subjects' emotional and physiological states during VR experiences. Additionally, the placement of ECG electrodes on the chest may introduce potential challenges, as subjects may have to wear additional equipment that could affect their comfort and immersion during the VR sessions. There are three different ways to mount Faros to a participant's chest, including Fast-Fix (Bittium's proprietary electrode), cable sets, and using a textile belt with two electrodes and a mounting pad for Faros. For our study, we chose the third option, using a textile belt with two electrodes and a mounting pad, to balance signal quality and participant comfort. Faros 360 was chosen for this study due to its specialization in ECG measurements, allowing the acquisition of precise HR data during the motion-controlled VR scenes. By leveraging the capabilities of Faros 360, subjects' cardiac responses can be understood, enhancing insights into their physiological reactions. Moreover, Faros 360 serves as a valuable reference for evaluating the performance of E4 wristband. Through a comparison of the data obtained from both devices, the consistency and reliability of the E4 wristband physiological measurements in the VR environment can be assessed. This comparative analysis has the potential to provide a comprehensive understanding of the strengths and limitations of each device, enabling informed decisions about their applications in future physiological research within VR settings.

2.3 Virtual reality

The study was conducted with the HTC Vive Pro Eye (2024) VR Headset (HTC Corporation, Vive, 2024). The system consists of a headset with integrated glasses with stereoscopic screens for

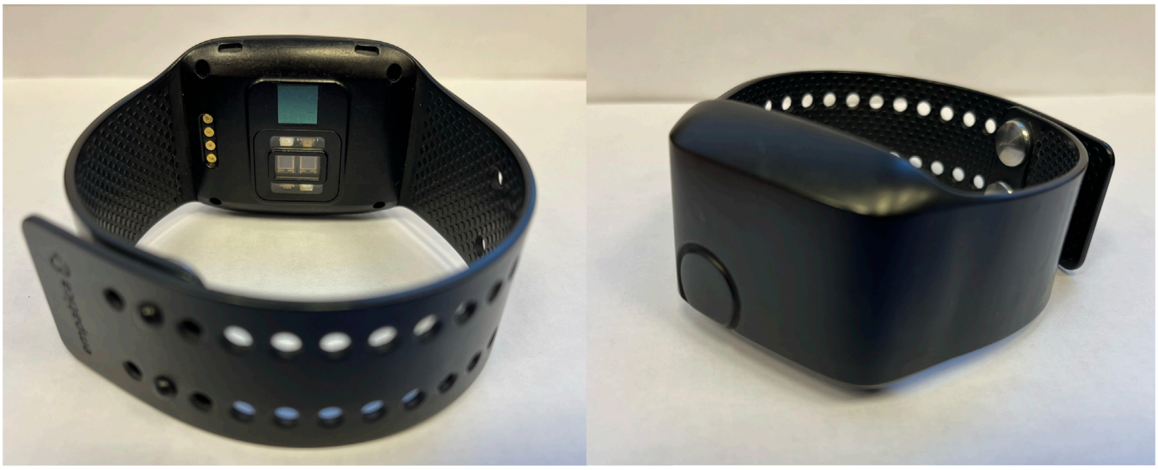


FIGURE 1
Empatica E4 device. This photography was taken at the Faculty of Electrical Engineering, University of Ljubljana.

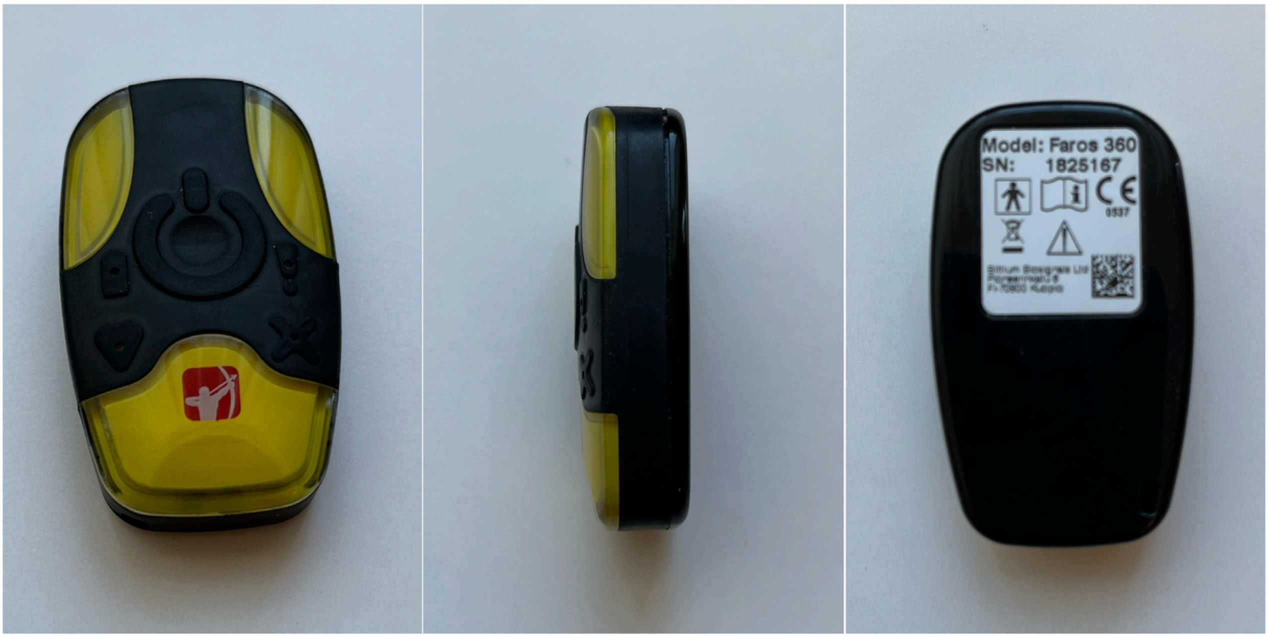


FIGURE 2
Bittium Faros 360 device. This photography was taken at the Faculty of Electrical Engineering, University of Ljubljana.

displaying content in virtual reality, and two hand-held controllers that are used to manipulate and interact with the environment and the displayed objects in it. The two screens (one for each eye) of the glasses are high-definition OLED screens with a diagonal of 8.89 cm (3.5 inches). Each screen has a resolution of 1440 × 1600 pixels, which means that the headset displays content with a total resolution of 2880 × 1600 pixels or 615 pixels per inch. The refresh rate is up to 90 Hz and offers 110-degree field of view. The headset straps and the distance between the screens are adjustable, which allows for adaptations that best conform to the subjects' needs (head size, pupillary distance, etc.). The hand-held controllers have a touch-sensitive surface, which the subject uses to input controls in a similar way as they would when using a touchpad on a laptop computer. The headset is equipped with speakers for playing sound.

For the baseline measurement subjects were equipped with Faros and E4 wristband for measuring the HR and BVP (respectively) and seated quietly with no significant movement, on a chair in the middle of the cabinet for 4 minutes. The length of the baseline data capture was consistent with the second and third parts of the experiment to ensure uniformity across all phases. This initial phase provided

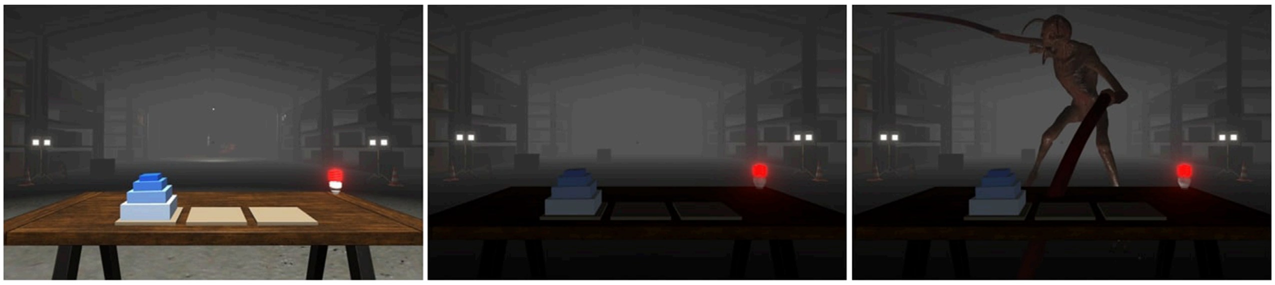


FIGURE 3
Left) Fall of the first box, center) fall of the second box, right) monster attack.

data on participants' resting state and physiological responses in the absence of virtual reality stimuli.

In the second part of the experiment, participants remained seated with both Faros and E4, but this time they also wore HTC Vive Pro Eye VR headset. The basic scene of Steam VR Home was used as the non-interactive scene (NIS), which is a virtual environment consisting of a house with a terrace on top of a mountain. The house is surrounded by trees, and in the distance the subject has a beautiful view of the surrounding mountains. In the background, the subject can hear the wind blowing and birds chirping. During the experiment, the subjects were placed on the mentioned terrace, where they could observe the view in the distance and the birds flying around them. This environment was chosen with a goal to keep the subjects calm during this scene and to not induce distress.

The interactive VR scene (IS) was created using Blender (Stichting Blender Foundation, Amsterdam, Netherlands) for the visual elements, and Unity (Unity Software Inc., San Francisco, United States) for the creation of the actual scene and implementation of the Hanoi Tower problem game. The scene was set in a poorly lit and slightly dimmed warehouse. A forklift is driving in the background and ambient sound of a warehouse and the forklift moving is played through the speakers. Shelves with cardboard boxes are placed left and right from the subject.

The subject is (virtually) seated in front of a table in the middle of the room (Figure 3), so they cannot see any of the walls of the space. A table lamp is lit red at the beginning of the test, which turns green upon successful completion of the Hanoi Tower task.

The scene test starts when the subject is satisfied with the placement of the VR and the view (and distance) they have of the table in front, which enables easy and comfortable moving of the cubes to complete the task. The subject is instructed to only move the dominant hand, keeping the non-dominant hand on the armrest. On the table, bases and cubes with different sizes are presented for solving the Hanoi Tower problem. The subject is first presented with three cubes and asked to arrange them in a predefined pattern. Upon successful completion, the scene ends. The scene is then reset, and the subject is presented with an additional cube, resulting in four cubes. Again, the subject is asked to arrange them in a predefined pattern. Upon successful completion, the scene ends. After that, the last scene is presented, where the subject is presented with five cubes and again asked to arrange them in a predefined pattern. Two minutes (120 s) after the start of the scene, one of the boxes falls from the left shelf to the floor accompanied by a loud bang. A few seconds before the end of the 3 minutes (180 s), a tension sound effect like typical sounds used in horror movies is played. This effect adds an extra level of suspense by telling the test taker that something is about to happen. As soon as the tension sound effect ends, another box falls from the right shelving unit, this time with a louder bang and further into the room. Unlike the first event, the second fall causes the lights on the ceiling to flicker, which can also be heard, for 1.5 s. Boxes falling are very short events, lasting less than 5 s.

After 4 minutes (240 s), a monster, making loud noises, jumps from the ceiling in front of the test subject and starts attacking them. As the monster lands on the ground, the lights in the background go out and the warehouse becomes very dark. A light placed under the table and aimed at the monster begins to flicker, illuminating the monster's face. This lasts less than 5 s, and at this point, the scene and the whole trial ends. The subject is at that moment instructed to take off the VR glasses.

2.4 Experiment design

The study involved subjects aged over 18 and under 40, due to the difference in physiological signals after a certain age (Quer et al., 2020; Zhang, 2007; Acharya et al., 2004), with no known cardiovascular diseases. Each subject participated on a voluntary basis, and they could withdraw or stop the experiment at any point. The study was conducted in accordance with Declaration of Helsinki (General Assembly of the World Medical Association, 2014) and the study design as well as the study execution strictly followed the Code of ethics for researchers and Guidelines for ethical conduct in research involving people issued by (University of Ljubljana, 2024). Before the study, the participants were informed about the study goals and asked to sign the Informed consent provided by the ethical committee of University of Ljubljana. We acquired data on the subjects' sex referring to the biological features related to both physical and physiological characteristics (Coen and Banister, 2012). The information on participant's sex was self-reported on a voluntary basis.

We have followed Cohen's guidelines for interpreting effect sizes (Cohen, 1988), with a slight modification for effect size distribution analysis for HRV studies as suggested by (Quintana, 2016; Laborde et al., 2017). The data for three subjects was not recorded appropriately in total, so the final sample size resulted in 8 female and 10 male subjects participated in the study, with a mean age of 22.3 ± 1.3 years (minimum age: 20, maximum age: 25).

The experiment was conducted in a motion-controlled virtual reality environment. Prior to the experiment, both the E4 and Faros devices were attached to the subject. Subjects were seated, with the E4 device positioned on their non-dominant hand, which they were instructed to keep on an armrest throughout the experiment to minimize movement. They were also instructed on the importance of remaining still to ensure data quality. Interaction with the VR scene was conducted using their dominant hand. Faros was placed with the textile belt right below the chest muscle. Once the VR headset was turned on, physiological responses were measured using both devices simultaneously. Each subject was exposed to a:

1. Baseline measurement
2. Non-interactive scene (NIS)
3. Interactive scene (IS),

and for each condition a 4-min recording was obtained.

2.5 Variables

In order to perform data quality assessment and distress detection, several key metrics had to be calculated.

2.5.1 Data quality metrics

The Signal-to-Noise Ratio (SNR) was estimated for each heart rate signal, measuring the strength of the desired signal relative to background noise or interference (Box, 1988). *Psignal* was calculated as the average of squared HR values recorded by each device, and *Pnoise* was estimated as the average squared difference between the HR values and their mean–essentially the variance of the HR signal. This calculation was performed separately for both the Empatica and Faros devices.

$$SNR = 10\,log_{10}\frac{P_{signal}}{P_{noise}}$$

Higher SNR values indicate stronger and more reliable heart rate signals, providing insights into signal fidelity and measurement accuracy.

Correlation is a statistical method used to evaluate the potential linear relationship between two continuous variables. The correlation coefficient, a dimensionless quantity ranging from −1 to +1, quantifies the strength of this presumed linear association. A coefficient closer to +1 indicates a strong positive correlation, while a value closer to −1 suggests a strong negative correlation. A coefficient near 0 indicates a weak or no linear relationship between the variables (Swinscow and Campbell, 2002; Witte R.S. and Witte J.S., 2017). In this study, both Pearson (Kirch, 2008) and Spearman (Dodge, 2008) correlation coefficients were employed to assess the linear and monotonic relationships between the Faros and E4 HR signals, respectively. While Pearson correlation measures linear associations, Spearman correlation evaluates monotonic relationships, making it less sensitive to nonlinear associations or outliers (Siegel and Castellan, 1981).

Furthermore, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) (Willmott et al., 1985) were computed for both Faros and E4 HR signals for each subject. The MAE represents the average absolute difference between corresponding heart rate values from Faros and E4, with Faros considered the reference or "ground truth" value. A smaller MAE indicates a lower overall error, reflecting a higher level of agreement between the two devices. On the other hand, RMSE represents the square root of the averaged squared differences between heart rate values, placing greater emphasis on larger errors compared to MAE.

Each of these metrics was calculated for 16 out of 18 subjects' baseline, NIS and IS scene, excluding two subjects with corrupted signals in IS. Out of the remaining 16 subjects, six were female and 10 were male.

2.5.2 Distress metrics

To compare the distress detection capabilities of E4 wristband and Faros based on HR features, two parameters were used:

1. Mean Heart Rate (Mean HR) provides an average HR value over a specific time period and serves as a measure of central tendency for HR data. It can be useful for understanding the overall level of cardiac activity during a specific time interval. It is commonly used to compare HR values between individuals or different conditions, such as rest and exercise (Karvonen and Vuorimaa, 1988). Mean HR is typically expressed in beats per minute [bpm].
2. Root Mean Square of Successive Differences (RMSSD) on the other hand, quantifies the variability in time intervals between consecutive heartbeats. It is calculated by taking the square root of the average of the squared differences between adjacent HR values. RMSSD reflects the high-frequency components of HRV, which are primarily influenced by parasympathetic (vagal) activity (Shaffer and Ginsberg, 2017). Higher RMSSD values indicate greater variability in HR, suggesting a more flexible autonomic nervous system and better cardiac health. RMSSD is often used as a marker of parasympathetic activity and can be used to assess the balance between sympathetic and parasympathetic regulation of the heart.

As outlined in (Stauss, 2014), it is crucial to avoid using HRV parameters in isolation without considering the mean level of HR, as this approach can lead to serious misinterpretation of experimental data. In this study, we chose Mean HR and RMSSD since their accuracy is preserved even when short-term recordings are used (Baek et al., 2015). Also, these two features are both important parameters in the analysis of HRV and are commonly used in research and clinical settings to assess autonomic function, cardiac health, and the impact of interventions or conditions on the cardiovascular system (Kamath et al., 2012).

Ultra-short-term (UST) recordings for HRV estimation have shown promise due to their efficiency in clinical and research settings. While UST measurements exhibit strong correlations with longer recordings for certain HRV parameters, such as mean HR and RMSSD, their accuracy may vary for other parameters like standard deviation of NN intervals (SDNN). Contextual factors, such as recording method (e.g., ECG vs PPG), age, health and artifact removal procedures, and the choice of HRV parameters can influence the reliability of UST measurements (Shaffer and Ginsberg, 2017). In this study, a 10 s window for segmentation and calculating mean HR and RMSSD was employed, since the correlations with the longer short-term recordings was reported as high enough–for mean HR in (Baek et al., 2015; Shaffer et al., 2016) and for RMSSD in (Baek et al., 2015; Nussinovitch et al., 2011; Salahuddin et al., 2007). Also, using a time window of 10 s helps capture short-term fluctuations in heart rate. It provides a balance between capturing immediate changes and avoiding noise or transient spikes that may occur within shorter time intervals. Munoz et al. (2015) conducted a comprehensive investigation into HRV measurements across a large adult sample (N = 3,387) and found that averaging over multiple 10-s segments, regardless of whether they are continuous, can provide reliable estimates of RMSSD. This approach aligns with contemporary practices in HRV analysis, where shorter recording periods are deemed sufficient for capturing meaningful physiological variability in quasi real-time. However, further research is needed to standardize protocols, establish normative values, and ensure consistent application of UST HRV measurements in place of conventional 5-min and 24-h recordings (Electrophysiology, 1996).

2.6 Data analysis

The data analysis section focuses on the processing and evaluation of the physiological data collected from both the E4 wristband and the Faros device during the motion-controlled VR experiment. The data undergoes thorough preprocessing to ensure data quality, followed by an assessment of E4 data quality. The section further explores distress detection using data from both devices and examines distress intensity and frequency. Additionally, a detailed analysis of distress during the interactive scene is conducted to gain insights into subjects' physiological responses.

2.6.1 Preprocessing

The collected physiological signals underwent comprehensive preprocessing to improve data quality and reliability. For this step and the analysis, Python version 3.9.7 (Python Software Foundation, Delaware, United States) was used. Synchronization between physiological recordings and VR events was achieved using timestamps from both data sources, since the distress events timestamps were known. The Faros ECG signal, sampled at 500 Hz, was subjected to various preprocessing steps using the *biosppy* Python package (Carreiras et al., 2018) and its *ecg.py* script. The first step was the application of a bandpass Finite Impulse Response (FIR) filter to eliminate artifacts outside of ECG frequency range. The order of this filter was calculated as 0.3*sampling rate, and the cutoff frequencies were set to 3 and 45 Hz, with a goal to eliminate baseline drift, remove low-frequency noise such as muscle artifacts and electrode motion artifacts, preserve ECG waveform and exclude higher-frequency noise. Hamilton segmentation (Hamilton, 2002) was used to accurately detect and isolate the QRS complexes and correction of R-peaks was done using template matching. Heart rate was calculated based on the array of R-peaks, and it was smoothed using moving average filter of type boxcar and window size equal to three samples (6 ms).

Similarly, the E4 BVP signal, sampled at 64 Hz, underwent preprocessing using the *ppg.py* script of the *biosspy* package. The first preprocessing step for the BVP signal involved filtering using 4th Butterworth bandpass filter with cutoff frequencies set at 1 and 8 Hz, applied with a goal to remove respiration influence (0.2–0.33 Hz), high frequency noise, and preserve heart rate range (from 1 Hz to 3 Hz, or 60–180 bpm). Both filters from *ppg.py* and *ecg.py* scripts use *filtfilt* function from the *scipy* package to perform zero-phase filtering, meaning that the filter is applied in both the forward and reverse directions, effectively eliminating any phase distortion introduced by the filtering process.

Onset detection was performed utilizing Elgendi's method (Elgendi et al., 2013), and heartbeat extraction was done using detected peaks. The final HR signals were obtained using moving average smoothing of type *boxcar* and window size set to three samples (46.88 ms).

Both HR signals were upsampled to 4 Hz, to ensure accurate alignment between the two datasets, facilitating meaningful comparative analysis of the physiological responses captured by the Faros and E4 devices.

2.6.2 Data quality assessment

To ensure the quality and reliability of the HR signals, key metrics described in Section 2.5.1. were calculated. SNR was used to evaluate each signal, with higher SNR values being preferable.

The correlation coefficient was used to quantify the similarity between the Faros and E4 HR signals, providing insights into their relationship. A higher correlation indicates a stronger connection and similar patterns, validating the accuracy and reliability of E4 HR measurements compared to Faros as the reference. A strong correlation signifies good E4 signal quality and reliability, while lower correlation may suggest potential measurement errors or artifacts.

MAE and RMSE were compared in order to determine the level of agreement between E4 and Faros for each of the conditions, with lower values indicating a higher degree of agreement.

In addition to the calculated metrics, Bland-Altman plot (Altman and Bland, 1983), a statistical method for assessing the agreement between two measurement techniques, was generated, and analyzed to further assess the agreement between Faros and E4 HR signals. This plot provides a comprehensive visualization of the mean differences and limits of agreement, offering valuable insights into the overall consistency and potential bias between the two measurement methods, aiding in the identification of any systematic bias or trends that may not be apparent in individual metrics.

2.6.3 Distress detection using Faros and E4: intensity and frequency

In order to detect distress during both scenes, three additional preprocessing steps needed to be performed: HR signal segmentation, HR feature calculation and feature threshold calculation, after which the detection analysis was performed.

HR signal was segmented using a 10 s window, and features were calculated as described in Section 2.5.2. The calculation of feature thresholds was guided by the recognition that heart rate can vary significantly among individuals due to factors such as sex, age, fitness level, health conditions, and other physiological differences (Alexandre et al., 2012). To account for these individual variations (Whitehead et al., 1977), a personalized threshold approach was implemented, aiming to establish distinct thresholds for each subject based on their unique baseline HR.

The baseline measurement served as the reference for deriving feature thresholds to compare with signal segments from NIS and IS. For the Mean HR feature, the threshold was determined by setting it to the minimum value of each subject's baseline HR, which is considered as the resting or normal HR. On the other hand, the threshold for the RMSSD feature was calculated as the median of the RMSSD values computed during the entire baseline period, in order to make it more robust to outliers and work with non-normal distribution (Altman and Bland, 1983; Rossi, 2022).

The utilization of individualized thresholds allows for a relative comparison of the features, as it considers the baseline HR specific to each subject. Specifically, when comparing the Mean HR feature with its threshold, a 30% increase above the personalized threshold was employed. Likewise, for the RMSSD feature, a 50% decrease below the subject's personalized threshold was used. This relative difference approach offers a more meaningful indication of significant changes in physiological responses, as it considers the unique baseline characteristics of each individual (Alexandre et al., 2012).

These two features are used in the analysis that compares the distress detection capabilities of E4 and Faros. The following steps were performed to compare the HR signals and their performance in detecting distress during each scene, so each step was performed on both E4 and Faros data, for each subject and each scene:

1. Threshold calculation: based on the Baseline measurement, calculated for each subject and its features.
2. Segmentation: HR signal is divided into non-overlapping segments of 10 s.
3. Feature calculation: Mean HR and RMSSD were calculated for each segment of the HR signal.
4. Threshold comparison: For each HR signal segment, both Mean HR and RMSSD are compared to their respective thresholds.
5. Distress detection: If the Mean HR of a segment exceeded its baseline threshold by 30%, it was considered elevated and labeled as one in the output vector. Similarly, if the RMSSD of a window was 50% below its baseline threshold, it was considered low and labeled as one in the output vector. Otherwise, a value of 0 was assigned, in both cases.
6. Threshold comparison vectors: for both Mean HR and RMSSD a threshold comparison vector was obtained.
7. Interpretation: A value of one in either the Mean HR and RMSSD threshold comparison vector pair indicated "medium distress level", a value of one in both the Mean HR and RMSSD threshold comparison vector pair indicated "high distress level" and a value of 0 in both the Mean HR and RMSSD threshold comparison vector pair indicated "low distress level" or absence of it. This step was performed to granulate the data and present a more detailed distress state of a subject.

For example, if a Mean HR threshold comparison vector is equal to [0 1 0 1], and its RMSSD threshold comparison vector is equal to [1 1 0 0] the resulting distress vector would be equal to [1 2 0 1] which would translate to:

- 1: medium distress level
- 2: high distress level
- 0: low distress level

8. Distress level cases occurrence: number of occurrences for each level of distress was counted.

A statistical analysis was done to compare the occurrence of distress of a certain level (low, medium, or high) for IS, between Faros and E4. For this analysis, the Wilcoxon signed-rank test (Wilcoxon, 1992) was used, since it is suitable for comparing paired data from the same group if the data is not normally distributed. We have used it along with Cliff's *delta* (C_{delta}) as the effect size measure. In this case, the Faros and E4 data came from the same subjects, and it was measured during IS. The hypothesis related to this test is that there should be no statistically significant difference between the distress level occurrences detected by E4 and Faros during IS.

The test was conducted with a confidence level of 95%, ensuring a reliable measure of statistical significance. The obtained p-values were then compared to the predetermined alpha level of 0.05, allowing us to assess whether the observed differences were statistically significant.

2.6.4 Interactive scene distress analysis

As it was already mentioned, during the interactive scene (IS) subjects are required to solve the Hanoi tower problem using a VR controller while being surprised by various distractors:

1. A box falls from the shelf in the 2nd min (120 s) after the scene starts.
2. The second box falls from the shelf in the 3rd minute (180 s) after the scene starts and the suspense sound effect is played.
3. The monster appears 4 min (240 s) into the scene, and it marks the end of the scene.

The idea of this analysis was to try to detect distress (if there is any) during the events at 120/180/240 s, for both Faros and E4. Since each HR signal is segmented into 10 s segments, the following steps are performed:

1. Segment extraction: For each event, three segments were extracted: (1) the segment immediately before the event, (2) the segment starting at the onset of the event with 10 s duration, and (3) the segment capturing the 10 s following the event onset. This approach accounted for potential delays in physiological responses while ensuring comprehensive coverage of distress reactions.
2. Distress detection was conducted by checking if distress was detected in at least one or two segments out of the three for each event situation. The resulting distress vectors explained in Section 2.4.4., step 7 was used to perform these checks.
3. Calculate distress occurrence in percents: based on the previous step values, the percentage of situations in which detected distress coincided with the VR events for each subject was calculated. This was done using two criteria: one-third (1/3) of the segments detecting distress and two-thirds (2/3) of the segments detecting distress. The resulting percentages ranged from 0% (no detected distress coinciding with VR events) to 100% (all detected distress coinciding with VR events).

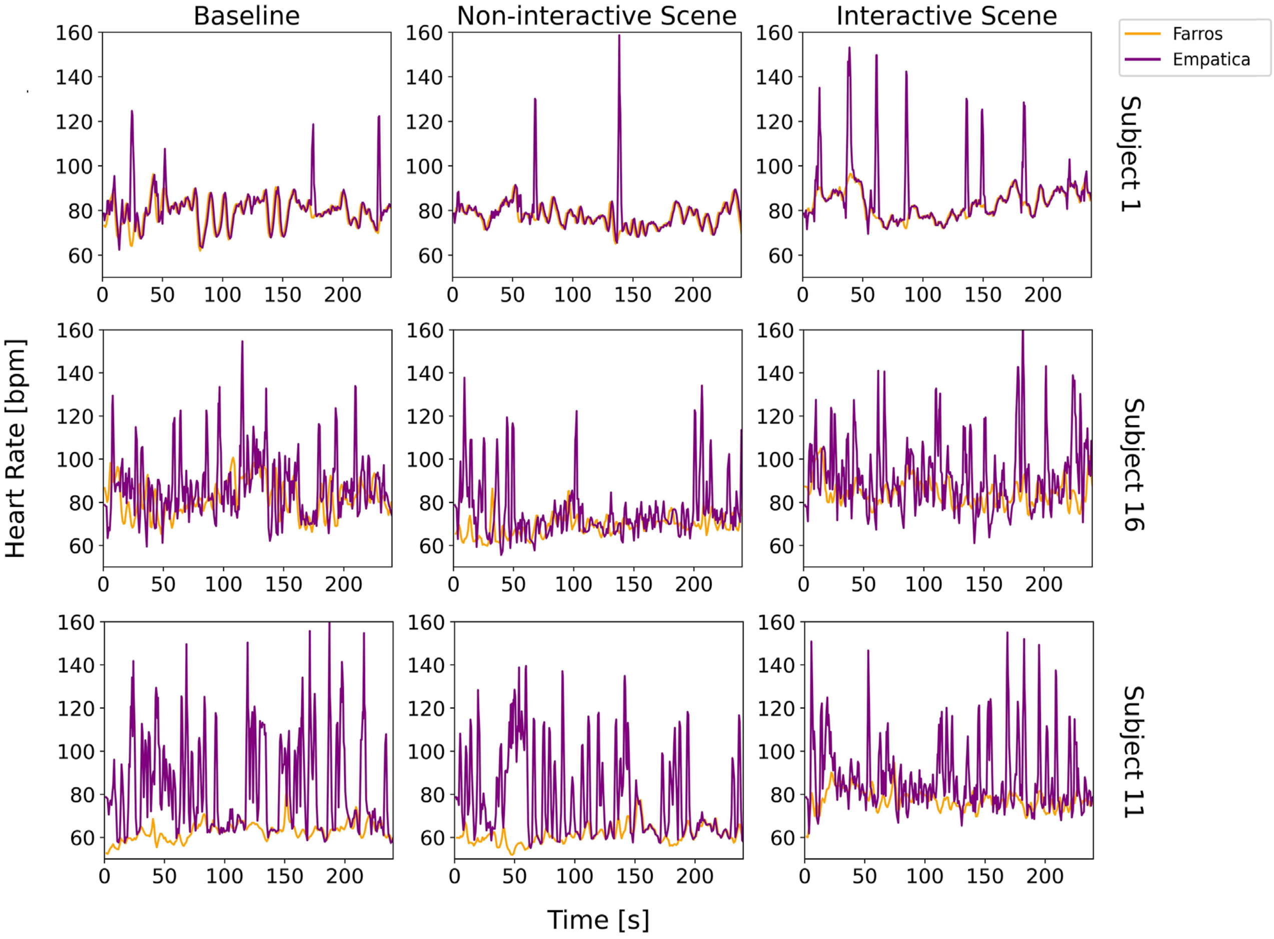


FIGURE 4
Visual comparison of Faros and E4 heart rate signals for three distinct cases (Subject 1, Subject 16, and Subject 11), for each scene.

3 Results

3.1 Data quality assessment results

To provide a comprehensive understanding of the data quality assessment, we start with a visual inspection of the HR signals from both devices. We specifically highlight three distinct cases, represented by Subjects 1, 11, and 16. These cases show case varying degrees of overlap between Faros and E4 signals: high overlap, medium overlap with E4 signal contamination, and low overlap with substantial E4 signal contamination, respectively (Figure 4). This visual representation serves as a foundational step in our analysis, allowing us to closely examine the specific differences in noise-related data quality assessment between the two devices.

In the case of Subject 1, there is little difference between Faros and E4 signals, for both scenes. However, occasional peaks in the E4 HR signal indicated the possibility of bad contact or unfiltered movement artifacts. Subject 16 exhibited a similar trend between the Faros and E4 HR signals. However, E4 signal was heavily contaminated with artifacts, likely caused by movement or other sources of unfiltered interference. For Subject 11, there was a clear lack of overlap between the Faros and E4 HR signals most of the time. The discrepancy could again be attributed to poor contact between the E4 device and the subject's skin and/or data loss issues.

The Bland-Altman plot is visualized in Figure 5, and it implies similar conclusions as the ones obtained by observing visual comparison of Faros and E4 HR signals in Figure 4. This plot is often used to assess how similar a new instrument or technique is at measuring something compared to the instrument or technique used as the reference (Giavarina, 2015). The abscissa of the plot displays the average measurement of the two devices, and the ordinate displays the difference in HR measurements between E4 and Faros. The further the value of the average difference (orange horizontal line) is from zero, the larger the difference in measurements between the instruments.

Figure 5 shows that that value is further away from zero if we observe Subjects from top row to the bottom row, coinciding with Figure 4 and confirming that the measurement noise increases the average difference in measurements between E4 and Faros. Besides that, for Subject 1, most of the data points are inside the 95% confidence interval (purple horizontal lines), while for

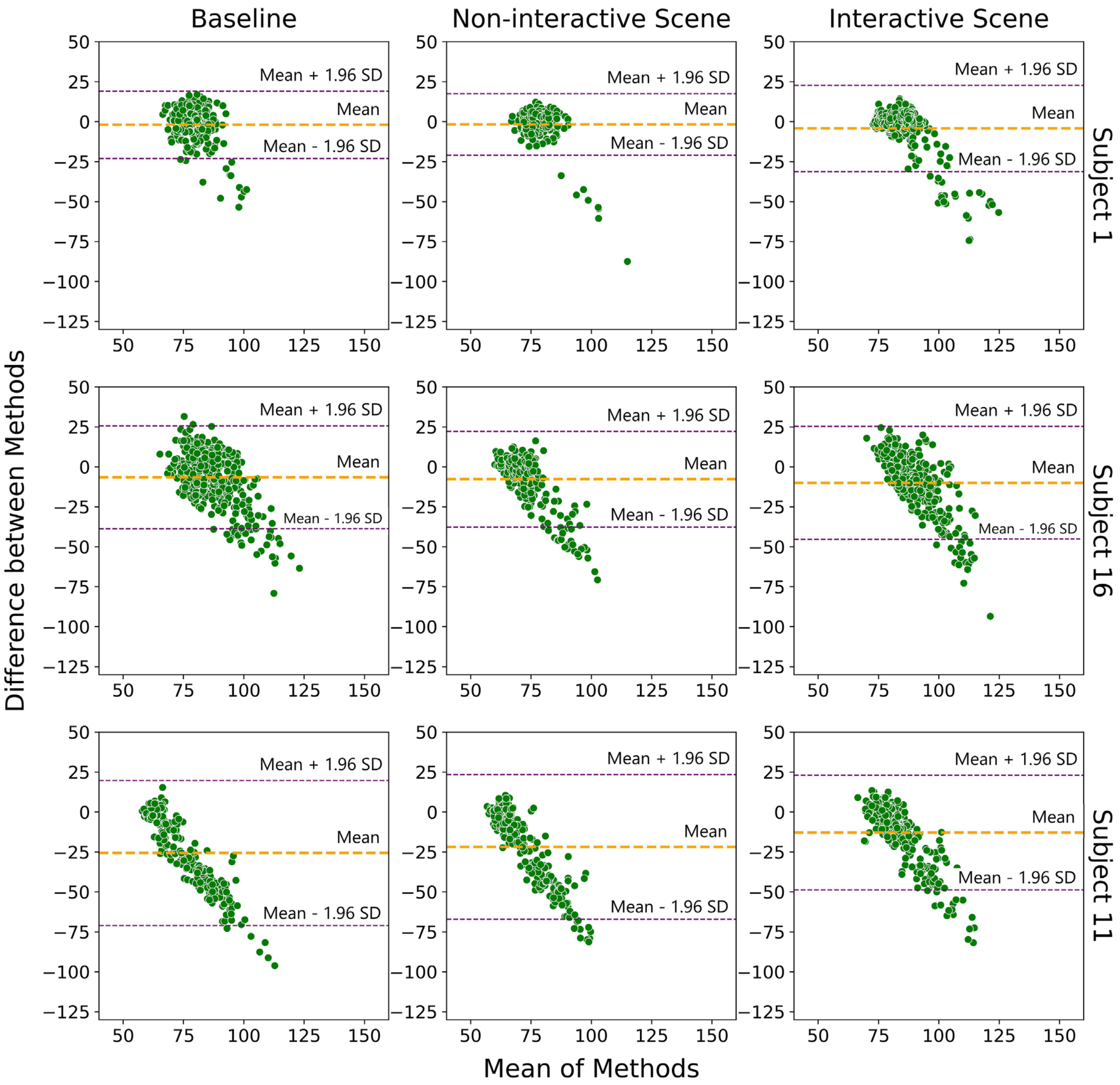


FIGURE 5
Bland-Altman plot showing the agreement between Faros and E4 heart rate signals for Subjects 1, 16, and 11. Orange line represents the mean difference in measurements between the two devices, and the two purple lines represent the upper/lower limit of the 95% confidence interval for the mean difference.

Subjects 16 and 11 we can observe that the data points are following a diagonal spread, indicating the discrepancy between measurements of E4 and Faros is biased proportionally to the magnitude of measurements. Also, in the upper right graph it can be seen that the measurements above the average line indicate biased comparison and are a consequence of subjects' movements and more pronounced artifacts, as revealed in Figure 4.

Table 1 presents the mean and standard deviations of SNR values for each sensor and scene utilized in this study, calculated across all subjects. The results indicate that Faros consistently exhibited a higher mean SNR than E4 for all measurements.

Furthermore, the standard deviations provide insights into the variability of the signal quality among the subjects for each sensor. Faros demonstrated smaller standard deviation values compared to E4, implying a greater degree of consistency in the SNR values across measurements.

Pearson and Spearman correlation coefficients mean values and standard deviation values, calculated between Faros and E4 across all subjects and for each scene, are displayed in Table 2. The correlation coefficients provide insights into the degree and direction of association between the data collected by the two devices.

TABLE 1 Signal-to-Noise Ratio (SNR) mean value and standard deviation calculated across all subjects, for both Faros and E4 and each scene.

Scene	Baseline		NIS		IS	
Sensor	E4	Faros	E4	Faros	E4	Faros
Mean ± SD [dB]	17.5 ± 3.2	22.4 ± 2.0	18.1 ± 3.9	24.2 ± 1.6	18.6 ± 3.3	22.9 ± 2.0

TABLE 2 Pearson and Spearman correlation coefficients mean values and standard deviation calculated between Faros and E4 across all subjects, for each scene.

Scene	Baseline		NIS		IS	
Sensor	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
Mean ± SD	0.24 ± 0.32	0.30 ± 0.34	0.19 ± 0.24	0.27 ± 0.27	0.31 ± 0.27	0.36 ± 0.29

TABLE 3 Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) mean values and standard deviation calculated between Faros and E4, across all subjects, for each scene.

Scene	Baseline		NIS		IS	
Sensor	RMSE	MAE	RMSE	MAE	RMSE	MAE
Mean ± SD [bpm]	14.4 ± 8.2	9.9 ± 6.5	13.7 ± 8.0	8.9 ± 6.0	13.2 ± 7.1	9.2 ± 5.9

The results show that both Pearson and Spearman correlation coefficients consistently showed positive values for all scenes, indicating a positive linear relationship between the data captured by Faros and E4 sensors. The mean Pearson correlation coefficients ranged from 0.19 to 0.31, while the mean Spearman correlation coefficients ranged from 0.27 to 0.36. These positive correlation values suggest that as the physiological measurements from Faros increase, the measurements from E4 also tend to increase, and *vice versa*. However, these values do not indicate a strong correlation between the measurements, and that may be caused by the noise present in E4 measurements. Standard deviation values ranged from 0.24 to 0.34 for Pearson and from 0.27 to 0.29 for Spearman. These standard deviations represent the variability in the correlation values among the scenes.

The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were computed to assess the agreement between the heart rate signals obtained from Faros and E4 devices for each subject and scene. The results presented in Table 3 indicate that for all three measurements (Baseline, NIS, and IS), the RMSE and MAE values were relatively small. For example, in NIS, the mean RMSE was 13.73 bpm, indicating an average difference of approximately 13.73 bpm between the HR measurements obtained from the two devices. The mean MAE in the same scenario was 8.87 bpm, representing an average absolute difference of around 8.87 bpm between the two measurements.

Furthermore, the standard deviations of RMSE and MAE values were also reported for each scene, providing information about the variability in accuracy across different subjects. In general, smaller standard deviation values suggest a higher degree of consistency and reliability in the accuracy of heart rate measurements between Faros and E4 for individual subjects.

3.2 Distress detection using Faros and E4: intensity and frequency results

The results in Table 4 provide information on the frequency of distress events at different levels (low, medium, and high) during NIS and IS as captured by both Faros and E4 devices.

Table 4 shows that, on average, there were more high- and medium- and less low-level distress occurrences detected in IS than in NIS both when using Faros and E4 HR signal. Standard deviation was higher for Faros-detected distress occurrences compared to E4-detected distress occurrences for all cases except for medium-level distress in NIS where the trend was reversed.

Comparing the two devices, it can be observed that there are differences in the number of detected distress events in each category. For example, in NIS, E4 detected more medium-level distress events compared to Faros, while Faros recorded more low-level distress events. In IS, E4 detected noticeably more medium-level distress events compared to Faros, but a similar amount of high-level distress events.

The Wilcoxon signed-rank test comparing distress detection capabilities of both devices at all distress levels, is visually represented using boxplots in Figure 6 for IS, and the p-values for are displayed in Table 5.

The $p < 0.01$ indicates that the difference in distress intensity occurrences between the devices is statistically significant for the Low and Medium distress levels. However, for the High distress level, the *p*-value of 0.57 suggests that there is no statistically significant difference in distress intensity occurrences between the devices. Cliff's *delta* was calculated for comparison of distress occurrences of each level between Faros and Empatica, and it was equal to −0.652, 0.667 and −0.018 for Low, Medium and High distress occurrences,

TABLE 4 Faros and E4 detected distress level occurrences for NIS and IS mean value and standard deviation.

Faros	NIS			IS		
	Low	Medium	High	Low	Medium	High
Mean ± SD [counts]	19 ± 8	3 ± 4	2 ± 6	9 ± 9	11 ± 7	7 ± 8
E4	**NIS**			**IS**		
	Low	**Medium**	**High**	**Low**	**Medium**	**High**
Mean ± SD [counts]	6 ± 8	14 ± 7	4 ± 5	1 ± 2	20 ± 6	6 ± 5

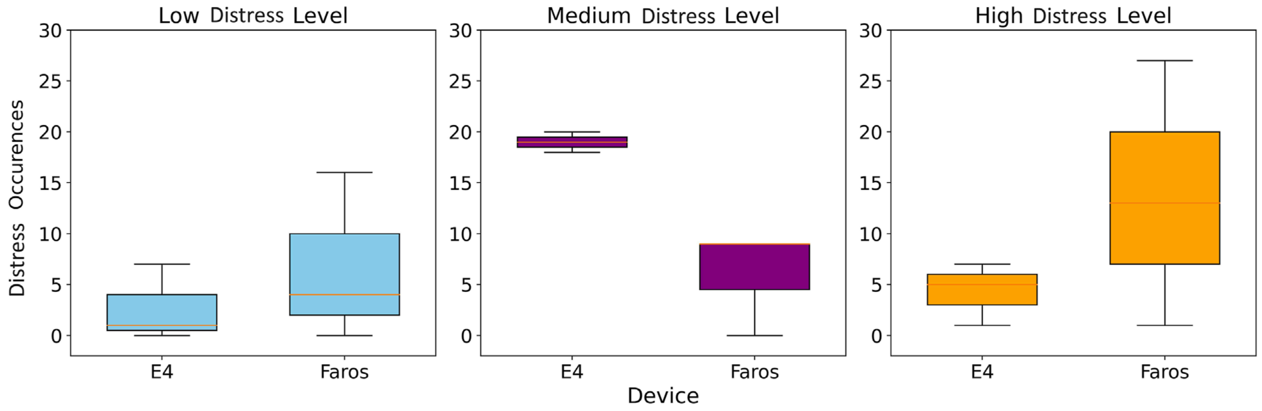


FIGURE 6
Boxplot comparisons of E4 and Faros for different distress level counts (left, middle, right plot) in IS.

TABLE 5 Distress intensity occurrences comparison between Faros and E4 device for different distress levels in IS. P-values were obtained using the Wilcoxon signed-rank test with 95% confidence interval.

Scene	Distress Intensity
Low	$p < 0.01$
Medium	$p < 0.01$
dHigh	$p = 0.57$

respectively. These values showed a significant difference between devices for Low and Medium distress occurrences, since any value above |0.474| is considered a large effect.

3.3 Interactive scene distress analysis results

The analysis of distress occurrence during the interactive IS was conducted using two criteria: detection in one-third (1/3) or two-thirds (2/3) of the segments (before, during, and after the triggering stimulus/event). The stricter criterion required distress presence for at least 20 s around the triggering event. The mean and standard deviation of the percentage of distress occurrences detected for both criteria (1/3 and 2/3) are shown in Table 6 for both Faros and E4 devices.

TABLE 6 Mean and Standard Deviation (SD) of percentage of distress occurrences detected coinciding with VR triggering situations during IS.

Device	Faros		E4	
Criteria	1/3	2/3	1/3	2/3
Mean ± SD [%]	88.9 ± 24.1	90.9± 21.6	98.2 ± 7.9	96.3 ± 15.7

The results indicate that both Faros and E4 devices detected a relatively high percentage of distress occurrences coinciding with triggering events during IS, with E4 showing slightly higher mean percentages compared to Faros. The standard deviation reflects the variability in event-related distress detection across participants for each criterion and device, showing that the results obtained using E4 HR signal were in a narrower range than the ones obtained using Faros HR signal (7.86% and 15.71% compared to 24.1% and 21.6%, respectively).

Figure 7 shows the periods of medium (yellow) and high (red) level distress occurrences for both E4 and Faros HR signals during NIS and IS. Although in NIS E4 HR signal is contaminated with more noise than Faros HR signal, there is no distress detected, which is expected since NIS is the scene without emotional events or triggers. The impact of noise on false positive distress occurrence detection can be seen in IS E4 HR signal, where Medium-level distress occurrence was detected from 0 to 80 s due to multiple high-intensity HR peaks that are not present in Faros HR signal.

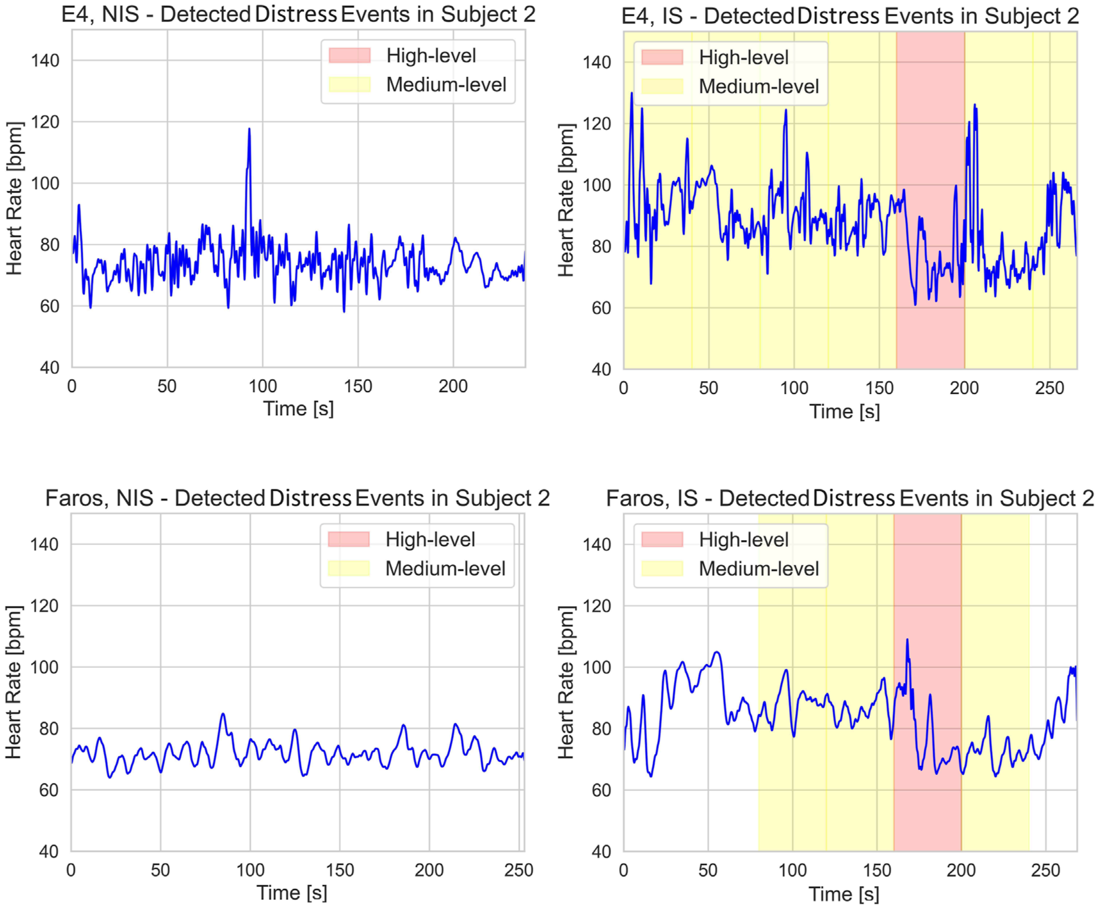


FIGURE 7
E4 (upper row) and Faros (lower row) heart rate signal with annotated medium- (yellow) and high- (red) level distress occurrences during NIS (left) and IS (right).

4 Discussion

The presented study compared the distress detection capabilities of two wearable devices, Faros and E4. The results shed light on the strengths and limitations of each device and their potential applications in assessing emotional responses during interactive scenarios. Moreover, the analysis of distress occurrence during the interactive scene provided valuable insights into the physiological responses of participants during the dynamic and immersive virtual reality experience.

To explore if both devices used in this study can effectively detect distress during IS and answer the first research question, the analysis of distress occurrence was conducted, and it revealed interesting patterns during IS. Faros and E4 detected more high- and medium-level distress occurrences in IS than in NIS, demonstrating the impact of surprising elements on participants' emotional responses. While E4 showed slightly higher mean percentages of distress occurrences compared to Faros, the devices exhibited similar overall performance in detecting distress occurrences during interactive scene-induced distress events. However, differences in detected distress events between two devices highlight the importance of considering device-specific factors in distress detection studies.

The Wilcoxon signed-rank test indicated that Faros and E4 exhibited statistically significant differences in distress intensity occurrences for low and medium levels, but not for high level in IS. Since the high distress level occurrences are the ones, we aimed to detect in the interactive scene, this is a significant result that indicates that both devices detected strong subject responses to distress inducing events, which implies that E4 could be used for high distress detection in motion-controlled environment. The significant differences between distress intensity occurrences for low and medium levels could be attributed to the noise contamination problem characteristic of E4 causing false positive distress detection. Noisy HR signal results in increased RMSSD values which prevents successful distress detection as with RMSSD we are looking for a decrease below the threshold. The significant difference between the two devices for low and medium distress levels is primarily the consequence of multiple false positives of E4 (cases where E4 detects distress even if it is not present). Since with E4 the RMSSD was less frequently below the designated threshold, therefore correctly interpreted as no distress, the occurrences of these false positive detections can mainly be attributed to the Mean HR signal of E4. This is not completely in line with the results from (Menghini et al., 2019) that clearly state that Mean HR measures for E4 show the best accuracy over various conditions. We believe the false positives observed in our study could also be attributed to the selected duration of the segments used for calculation as a 10 s segment is more prone to erroneous distress detection in case of a noisy HR signal.

Ragot et al. (2018) compared a laboratory based Biopac sensor to wearable E4 device for detecting emotion valence and intensity (distress) using selected features, for which the correlation coefficients ranged from 0.13 to 0.99, indicating non-consistency among different parameters. Our approach showed that higher distress levels are consistently detected with both devices when using appropriate feature engineering, comparable to the results of the Machine Learning approach used in (Ragot et al., 2018) on a similar-sized study sample of 19 volunteers. While valuable comparison presented in (Ragot et al., 2018) confirms that Empatica can be used for emotion valence and intensity classification in a non-noisy, static environment, our study shows that E4 device can be used for high-level distress detection in motion controlled, interactive VR environment.

With regards to our second research question, we addressed specifically the problem of data quality of E4 as this device is known to be more prone to motion artifacts and results in poorer SNRs. This was done mostly through noise contamination analysis and direct comparison of both devices. While Faros consistently exhibited higher mean SNR and smaller standard deviation, E4 signals occasionally showed false peaks indicating possible bad contact or unfiltered movement artifacts.

Both Pearson and Spearman correlation coefficients showed positive linear relationships between Faros and E4 data across all scenes. However, the correlation coefficients are relatively moderate, with large standard deviations, indicating that the strength of the association between the data from these two devices is not exceptionally strong. The moderate correlation can in our opinion again be explained by the motion (and other) artefacts, common for E4 measurements.

Despite relatively low correlation coefficients, small Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values indicate good agreement in measuring heart rate. The smaller standard deviation values for Faros further suggest higher consistency and reliability in accuracy compared to E4. It is important to consider that the Faros HR, obtained from the ECG signal with a higher sampling frequency of 500 Hz, provides better temporal resolution compared to E4 HR. This difference in resolution could also contribute to potential errors and discrepancies in the E4 HR signal.

It becomes evident that the presence of artifacts, poor contacts, data loss and different sampling frequency impact the reliability and alignment between the Faros and E4 HR signals. This corresponds to findings reported in (Schuurmans et al., 2020) where E4 also proved to be a practical and valid tool for research on HR and HRV, but only in movement-controlled conditions (in this study, subjects were meditating). With that, our third research question can be answered by stating that E4 device exhibits good performance in measuring heart rate, but with lower reliability and accuracy compared to Faros, due to the limitations based on lower sampling frequency and presence of artifacts and poor contacts that introduce noise to the measurement.

The results of our study and their interpretation should be considered with several limitations:

1. The noise present in the E4 HR signal after preprocessing still impacts the accuracy of distress detection, as observed in the example of Subject 2, IS with false positives. Future work should include the employment of advanced signal processing techniques to recognize noisy segments and to minimize the impact of noise on distress analysis.
2. We did not assess Heart Rate Variability (HRV) from Faros, as it may differ from Pulse Rate Variability (PRV) from E4 and thus could introduce even more discrepancies in results (Park et al., 2022).
3. The study focused on specific triggering events during IS to assess distress occurrences. While these events were carefully designed, they may not fully represent the complexity and variability of emotional responses in real-world scenarios. Further investigations incorporating a wider range of emotional stimuli and experiences would provide a more comprehensive understanding of distress detection in dynamic environments.
4. The main limitation of this study was the sample size, which was relatively small and age cohort as it included only university students and consisted of 8 females (6, since for two females the measurements were corrupted) and 10 males. This may limit the generalizability of the findings. Future studies with larger and more diverse samples are needed to validate the observed trends.
5. The study focused on distress intensity occurrences, which can be influenced by sex, age, individual differences and contextual factors. Incorporating these factors in future research would contribute to a more nuanced understanding of the complex interplay between physiological responses, individual differences, and emotional experiences. This includes intra-individual factors such as caffeine intake and fatigue levels, which may influence physiological responses and heart rate variability. We should consider controlling these variables in the future, for their potential impact on distress detection in VR settings.
6. We have mostly focused on the participant comfort and non-intrusivity of the setup, which is why we didn't include measurement of signals like EEG, which require additional equipment, adding bulk and pressure and reduces participant comfort, but, for example, has been shown useful in classifying distress and no distress situations (Eldeeb et al., 2021). We should aim to assess additional physiological measurements that could provide more insights into the physiology behind the distress assessment using VR in healthy subjects.
7. No self-reported measures of distress were included in our study. While validated questionnaires could provide valuable ground truth data, they are inherently limited by biases such as social desirability, recall errors, acquiescence, and participant fatigue. Future studies should incorporate these measures, such as the Generalized Anxiety Disorder-7 (GAD-7) questionnaire (Spitzer et al., 2006), while carefully considering their limitations when interpreting self-reported data alongside physiological measurements

In this paper we presented a comprehensive approach to measuring and understanding subjects' physiological responses within the motion-controlled VR environment by using two commercially available wearable devices. The study highlights primarily the importance of considering device-specific factors and data quality when using such wearable devices for distress detection. Faros demonstrated superior signal quality and consistency compared to E4 by retaining higher mean signal-to-noise ratios (from 4.3 dB to 6.1 dB) for all scenes, making it a more reliable choice for studies requiring high-quality HR data. Although correlation coefficients between data measured by Faros and E4 were consistently positive, they revealed relatively weak correlations with correlation coefficients below 0.4. Both devices, however, showed good agreement in measuring heart rate with average absolute difference less than 9 bpm, indicating their potential utility in assessing emotional responses during motion controlled interactive VR scenarios. Moreover, both devices performed well in detecting distress occurrences related to the triggering events and to the high distress levels. We found no statistically significant difference between Faros and E4 data for comparing high distress intensity occurrences (p-value = 0.57), while this is not true for low and medium distress intensities (p-value <0.01).

In addition to device comparison, we have also proposed a simple rule- and subject-specific threshold-based distress detection method that showed promising results and performance, especially when detecting distress coinciding with the distress-inducing events which were included in the interactive scene by design. Threshold fine-tuning and exploring different options and threshold values is out of scope of this paper, but it is one of the directions we would consider in our future work.

While E4 device shows promising potential as a practical alternative to Faros for distress detection, especially in scenarios where wrist-worn monitoring is preferred, researchers must be mindful of the specific research objectives and the level of data accuracy and consistency required. For studies that demand the highest level of data reliability and signal stability, Faros remains the preferred choice. Nonetheless, these findings open the door for further investigations and advancements in wearable physiological monitoring technologies. Our future research could include adding more distress-inducing scenarios and improving existing ones, considering more physiological features for distress detection, testing multiple commercially available devices, and trying to minimize the movement artifacts through device placement or different movement artifact removal methods. Future research could be directed towards examination of different distress inducing scenarios, comparison of other relevant physiological features for distress detection, testing multiple wearable devices, minimization of the movement artifacts with appropriate processing methods, and fine-tuning the feature thresholds for distress detection.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Institutional Review Board of Department of ICT. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

JM: Formal Analysis, Investigation, Methodology, Software, Writing–original draft, Writing–review and editing. NM: Data curation, Formal Analysis, Methodology, Validation, Writing–review and editing. KP: Investigation, Methodology, Writing–review and editing. JS: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Slovenian Research Agency within the research program ICT4QoL–Information and Communications Technologies for Quality of Life, grant number P2-0246. NM was partly supported by the Ministry of Education, Science, and Technological Development, Republic of Serbia, grant number 451-03-68/2022-14/200103.

Acknowledgments

The authors would like to acknowledge Nejc Prezelj for developing the VR scenes used in the experiment as well as Ana Farič and Katarina Pikec for their assistance with the data collection.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2025.1480018/full#supplementary-material

References

Acharya, U. R., N, K., Sing, O., Ping, L., and Chua, T. (2004). Heart rate analysis in normal subjects of various age groups. *Biomed. Eng. OnLine* 3 (1), 24. doi:10.1186/1475-925X-3-24

Alexandre, D., Da Silva, C. D., Hill-Haas, S., Wong, D. P., Natali, A. J., De Lima, J. R. P., et al. (2012). Heart rate monitoring in soccer: interest and limits during competitive match play and training, practical application. *J. Strength and Cond. Res.* 26 (10), 2890–2906. doi:10.1519/JSC.0b013e3182429ac7

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiol. Meas.* 28 (3), R1–R39. doi:10.1088/0967-3334/28/3/R01

Altini, M., and Plews, D. (2021). What is behind changes in resting heart rate and heart rate variability? A large-scale analysis of longitudinal measurements acquired in free-living. *Sensors* 21, 7932. doi:10.3390/s21237932

Altman, D. G., and Bland, J. M. (1983). Measurement in medicine: the analysis of method comparison studies. *J. R. Stat. Soc. Ser. D Statistician* 32 (3), 307–317. doi:10.2307/2987937

Baek, H. J., Cho, C. H., Cho, J., Woo, J. M., Kim, S. J., and Lee, K. M. (2015). Reliability of ultra-short-term analysis as a surrogate of standard 5-min analysis of heart rate variability. *Telemedicine e-Health* 21 (5), 404–414. doi:10.1089/tmj.2014.0104

Béquet, A. J. (2023). *EmbracePlus_AVRO2CSV.py* Source code. *GitHub*. Available at: https://github.com/abequet/Physio_processing/blob/main/EmbracePlus_AVRO2CSV.py.

Bittium Faros (2024). Bittium Faros. Available at: https://www.bittium.com/medical/bittium-faros (Accessed February 24, 2024).

Böttcher, S., Vieluf, S., Bruno, E., Joseph, B., Epitashvili, N., Biondi, A., et al. (2022). Data quality evaluation in wearable monitoring. *Sci. Rep.* 12, 21412. doi:10.1038/s41598-022-25949-x

Boucsein, W. (2012). *Electrodermal activity*. Springer Science and Business Media.

Box, G. (1988). Signal-to-noise ratios, performance criteria, and transformations. *Technometrics* 30 (1), 1–17. doi:10.1080/00401706.1988.10488313

Carreiras, C., Alves, A. P., Lourenço, A., Canento, F., Silva, H., and Periquito, P. (2018). Biosppy: biosignal processing in python. *Computers* 3 (28), 2018.

Coen, S., and Banister, E. (2012). *What a difference sex and gender make: a gender, sex, and health research casebook*. doi:10.14288/1.0132684

Cohen, D. (1988). *Statistical power analysis for behavioral sciences*. Hillsdale, MI: Erlbaum.

Dodge, Y. (2008). *The concise encyclopedia of statistics*. Springer Science and Business Media.

Duric, Z., Gray, W. D., Heishman, R., Li, F., Rosenfeld, A., Schoelles, M. J., et al. (2002). Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proc. IEEE* 90, 1272–1289. doi:10.1109/jproc.2002.801449

Egan, D., Brennan, S., Barrett, J., Qiao, Y., Timmerer, C., and Murray, N. (2016). "An evaluation of Heart Rate and ElectroDermal Activity as an objective QoE evaluation method for immersive virtual reality environments," in *2016 eighth international conference on quality of multimedia experience (QoMEX)* (IEEE), 1–6.

Eldeeb, S., Susam, B. T., Akcakaya, M., Conner, C. M., White, S. W., and Mazefsky, C. A. (2021). Trial by trial EEG based BCI for distress versus non distress classification in individuals with ASD. *Sci. Rep.* 11 (1), 6000. doi:10.1038/s41598-021-85362-8

Electrophysiology, T. (1996). Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation* 93 (5), 1043–1065. doi:10.1161/01.cir.93.5.1043

Elgendi, M., Norton, I., Brearley, M., Abbott, D., and Schuurmans, D. (2013). Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions. *PloS one* 8 (10), e76585. doi:10.1371/journal.pone.0076585

Empatica (2024). Empatica. Available at: https://www.empatica.com/research/e4/(Accessed February 21, 2024).

Empatica Embrace Plus (2024). Empatica EmbracePlus. Available at: https://www.empatica.com/embraceplus/(Accessed February 21, 2024).

General Assembly of the World Medical Association (2014). World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. *J. Am. Coll. Dent.* 81 (3), 14–18.

Giavarina, D. (2015). Understanding bland altman analysis. *Biochem. medica* 25 (2), 141–151. doi:10.11613/BM.2015.015

Gromala, D., Tong, X., Choo, A., Karamnejad, M., and Shaw, C. D. (2015). "The virtual meditative walk: virtual reality therapy for chronic pain management," in *Proceedings of the 33rd Annual ACM conference on human factors in computing systems*, 521–524.

Gruden, T., Stojmenova, K., Sodnik, J., and Jakus, G. (2019). Assessing drivers' physiological responses using consumer grade devices. *Appl. Sci.* 9, 5353. doi:10.3390/app9245353

Hamilton, P. (2002). "Open-source ECG analysis," in *Computers in cardiology* (IEEE), 101–104.

Heikenfeld, J., Jajack, A., Rogers, J., Gutruf, P., Tian, L., Pan, T., et al. (2018). Wearable sensors: modalities, challenges, and prospects. *Lab. Chip* 18, 217–248. doi:10.1039/C7LC00914C

Kamath, M. V., Watanabe, M., and Upton, A. (2012). *Heart rate variability (HRV) signal analysis: clinical applications*.

Karvonen, J., and Vuorimaa, T. (1988). Heart rate and exercise intensity during sports activities: practical application. *Sports Med.* 5, 303–311. doi:10.2165/00007256-198805050-00002

Kirch, W. (2008). "Pearson's correlation coefficient," in *Encyclopedia of public health* (Dordrecht: Springer).

Koenig, J., Jarczok, M. N., Kuhn, W., Morsch, K., Schäfer, A., Hillecke, T. K., et al. (2013). Impact of caffeine on heart rate variability: a systematic review. *J. Caffeine Res.* 3 (1), 22–37. doi:10.1089/jcr.2013.0009

Laborde, S., Mosley, E., and Thayer, J. F. (2017). Heart rate variability and cardiac vagal tone in psychophysiological research – recommendations for experiment planning, data analysis, and data reporting. *Front. Psychol.* 8, 213. doi:10.3389/fpsyg.2017.00213

Looff, P., Duursma, R., Noordzij, M., Taylor, S., Jaques, N., Scheepers, F., et al. (2022). Wearables: an R package with accompanying shiny application for signal analysis of a wearable device targeted at clinicians and researchers. *Front. Behav. Neurosci.* 16, 856544. doi:10.3389/fnbeh.2022.856544

Mack, D. C., Alwan, M., Turner, B., Suratt, P., and Felder, R. A. (2006). "A passive and portable system for monitoring heart rate and detecting sleep apnea and arousals: preliminary validation," in *1st transdisciplinary conference on distributed diagnosis and Home healthcare, 2006* (IEEE), 51–54.

McCarthy, C., Pradhan, N., Redpath, C., and Adler, A. (2016). "Validation of the Empatica E4 wristband," in *Proceedings of the 2016 IEEE EMBS international student conference (ISC)*, 1–4.

Menghini, L., Gianfranchi, E., Cellini, N., Patron, E., Tagliabue, M., and Sarlo, M. (2019). Stressing the accuracy: wrist-worn wearable sensor validation over different conditions. *Psychophysiology* 56 (11), e13441. doi:10.1111/psyp.13441

Munoz, M. L., Van Roon, A., Riese, H., Thio, C., Oostenbroek, E., and Westrik, I. (2015). Validity of (ultra-) short recordings for heart rate variability measurements. *PloS one.* 10 (9), e0138921. doi:10.1371/journal.pone.0138921

Nussinovitch, U., Elishkevitz, K. P. A., Katz, K., Nussinovitch, M., Segev, S., Volovitz, B., et al. (2011). Reliability of ultra-short ECG indices for heart rate variability. *Ann. Noninvasive Electrocardiol.* 16 (2), 117–122. doi:10.1111/j.1542-474X.2011.00417.x

Parasuraman, R. (2003). Neuroergonomics: research and practice. *Theor. Issue. Ergon. Sci.* 4 (1-2), 5–20. doi:10.1080/14639220210199753

Park, J., Seok, H. S., Kim, S. S., and Shin, H. (2022). Photoplethysmogram analysis and applications: an integrative review. *Front. Physiology* 12, 808451. doi:10.3389/fphys.2021.808451

Parsons, T. D., and Reinebold, J. L. (2012). Adaptive virtual environments for neuropsychological assessment in serious games. *IEEE Trans. Consumer Electron.* 58, 197–204. doi:10.1109/tce.2012.6227413

Quer, G., Gouda, P., Galarnyk, M., Topol, E. J., and Steinhubl, S. R. (2020). Inter- and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: retrospective, longitudinal cohort study of 92,457 adults. *PLOS ONE* 15 (2), e0227709. doi:10.1371/journal.pone.0227709

Quintana, D. S. (2016). Statistical considerations for reporting and planning heart rate variability case-control studies. *Psychophysiology* 54, 344–349. doi:10.1111/psyp.12798

Raghav, K., Van Wijk, A. J., Abdullah, F., Islam, M. N., Bernatchez, M., and De Jongh, A. (2016). Efficacy of virtual reality exposure therapy for treatment of dental phobia: a randomized control trial. *BMC oral health* 16 (1), 25–11. doi:10.1186/s12903-016-0186-z

Ragot, M., Martin, N., Em, S., Pallamin, N., and Diverrez, J. M. (2018). "Emotion recognition using physiological signals: laboratory vs. wearable sensors," in *Advances in human factors in wearable Technologies and game design: proceedings of the AHFE 2017 international conference on advances in human factors and wearable Technologies, july 17-21, 2017, the westin bonaventure hotel, Los Angeles, California, USA 8* (Springer International Publishing), 15–22.

Rahman, M. A., Brown, D. J., Shopland, N., Harris, M. C., Turabee, Z. B., Heym, N., et al. (2022). "Towards machine learning driven self-guided virtual reality exposure therapy based on arousal state detection from multimodal data," in *International conference on brain informatics* (Springer International Publishing), 195–209.

Robitaille, P., and McGuffin, M. J. (2019). "Increased affect-arousal in VR can be detected from faster body motion with increased heart rate," in *Proceedings of the ACM SIGGRAPH symposium on interactive 3D graphics and games*, 1–6.

Rossi, R. J. (2022). *Applied biostatistics for the health sciences*. John Wiley and Sons.

Salahuddin, L., Cho, J., Jeong, M. G., and Kim, D. (2007). "Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings," in *2007 29th annual international conference of the ieee engineering in medicine and biology society* (IEEE), 4656–4659.

Schuurmans, A. A. T., de Looff, P., Nijhof, K. S., Rosada, C., Scholte, R. H. J., Popma, A., et al. (2020). Validity of the Empatica E4 wristband to measure heart rate variability (HRV) parameters: a comparison to electrocardiography (ECG). *J. Med. Syst.* 44, 190–211. doi:10.1007/s10916-020-01648-w

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

Shaffer, F., Shearman, S., and Meehan, Z. M. (2016). The promise of ultra-short-term (UST) heart rate variability measurements. *Biofeedback* 44 (3), 229–233. doi:10.5298/1081-5937-44.3.09

Siegel, S., and Castellan, N. J. (1981). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-HiU Book Company.

Slater, M., Banakou, D., Beacco, A., Gallego, J., Macia-Varela, F., and Oliva, R. (2022). A separate reality: an update on place illusion and plausibility in virtual reality. *Front. virtual Real.* 3, 914392. doi:10.3389/frvir.2022.914392

Spitzer, R. L., Kroenke, K., Williams, J. B., and Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives Intern. Med.* 166 (10), 1092–1097. doi:10.1001/archinte.166.10.1092

Stauss, H. M. (2014). Heart rate variability: just a surrogate for mean heart rate? *Hypertension* 64 (6), 1184–1186. doi:10.1161/HYPERTENSIONAHA.114.03949

Swinscow, T. D. V., and Campbell, M. J. (2002). *Statistics at square one*. London: Bmj.

Tiwari, R., Kumar, R., Malik, S., Raj, T., and Kumar, P. (2021). Analysis of heart rate variability and implication of different factors on heart rate variability. *Curr. Cardiol. Rev.* 17, 74–83. doi:10.2174/1573403x16999201231203854

Tran, Y., Wijesuriya, N., Tarvainen, M., Karjalainen, P., and Craig, A. (2009). The relationship between spectral changes in heart rate variability and fatigue. *J. Psychophysiol.* 23 (3), 143–151. doi:10.1027/0269-8803.23.3.143

University of Ljubljana (2024). Etika in integriteta v raziskovanju. Available at: https://www.uni-lj.si/raziskovalno_in_razvojno_delo/etika_in_integriteta_v_raziskovanju/(Accessed February 21, 2024).

Vive Pro Eye (2024). Vive Pro eye. Available at: https://web.archive.org/web/20201111190618/https://www.vive.com/eu/product/vive-pro-eye/overview/(Accessed February 21, 2024).

Whitehead, W. E., Drescher, V. M., Heiman, P., and Blackwell, B. (1977). *Relation of heart rate control to heartbeat perception*. Biofeedback and Self-regulation 2, 371–392. doi:10.1007/BF00998623

Wilcoxon, F. (1992). "Individual comparisons by ranking methods," in *Breakthroughs in statistics: methodology and distribution* (New York, NY: Springer New York), 196–202.

Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R., et al. (1985). Statistics for the evaluation and comparison of models. *J. Geophys. Res. Oceans* 90 (C5), 8995–9005. doi:10.1029/jc090ic05p08995

Witte, R. S., and Witte, J. S. (2017). *Statistics*. John Wiley and Sons.

Wout, M., Spofford, C. M., Unger, W. S., Sevin, E. B., and Shea, M. T. (2017). Skin conductance reactivity to standardized virtual reality combat scenes in veterans with PTSD. *Appl. Psychophysiol. Biofeedback* 42, 209–221. doi:10.1007/s10484-017-9366-0

Wu, Y., Gu, R., Yang, Q., and Luo, Y. J. (2019). How do amusement, anger and fear influence heart rate and heart rate variability? *Front. Neurosci.* 13, 1131. doi:10.3389/fnins.2019.01131

Zhang, J. (2007). Effect of age and sex on heart rate variability in healthy subjects. *J. Manip. Physiological Ther.* 30 (5), 374–379. doi:10.1016/j.jmpt.2007.04.001

TYPE Original Research
PUBLISHED 27 March 2025
DOI 10.3389/fdgth.2025.1553565

Investigating the accuracy of Garmin PPG sensors on differing skin types based on the Fitzpatrick scale: cross-sectional comparison study

Annie Icenhower[1†], Claire Murphy[1†], Amber K. Brooks[2], Megan Irby[1], Kindia N'dah[1], Justin Robison[1] and Jason Fanning[1*]

[1]Department of Health and Exercise Science, Wake Forest University, Winston-Salem, NC, United States, [2]Department of Anesthesiology, Wake Forest University School of Medicine, Winston-Salem, NC, United States

OPEN ACCESS

EDITED BY
Mohammad Yavarimanesh,
University of San Diego, United States

REVIEWED BY
Lars Masanneck,
University Hospital of Düsseldorf, Germany
Pegah Rezaei,
University of Maryland, United States

*CORRESPONDENCE
Jason Fanning
✉ fanninjt@wfu.edu

[†]These authors have contributed equally to this work and share first authorship

RECEIVED 30 December 2024
ACCEPTED 11 March 2025
PUBLISHED 27 March 2025

CITATION
Icenhower A, Murphy C, Brooks AK, Irby M, N'dah K, Robison J and Fanning J (2025) Investigating the accuracy of Garmin PPG sensors on differing skin types based on the Fitzpatrick scale: cross-sectional comparison study.
Front. Digit. Health 7:1553565.
doi: 10.3389/fdgth.2025.1553565

Background: Commercial wearable devices, which are often capable of estimating heart rate via photoplethysmography (PPG), are increasingly used in health promotion. In recent years, researchers have investigated whether the accuracy of PPG-measured heart rate varies based on skin pigmentation, focusing particularly on the accuracy of such devices among users with darker skin tones. As such, manufacturers of wearable devices have implemented strategies to improve accuracy. Given the ever-changing nature of the wearable device industry and the important health implications of providing accurate heart rate estimates for all individuals no matter their skin color, studies exploring the impact of pigmentation on PPG accuracy must be regularly replicated.

Objective: We aimed to contrast heart rate readings collected via PPG using the Garmin Forerunner 45 in comparison with an electrocardiogram (ECG) during various levels of physical activity across a diverse group of participants representing a range of skin tones.

Methods: Heart rate data were collected from adult participants (18–64 years of age) at a single study session using the Garmin Forerunner 45 PPG-equipped smartwatch and the Polar H10 ECG chest strap. Skin tone was self-reported via the Fitzpatrick scale. Each participant completed two 10 min bouts of moderate-intensity walking or jogging separated by a 10 min bout of light walking.

Results: A series of mixed ANOVAs indicated no significant interaction between Fitzpatrick score and phase of the activity bout (i.e., rest at the start, first intensity ramp-up phase, first steady-state phase, active rest, second ramp-up phase, and second steady-state phase). Similarly, there was no significant main effect for the Fitzpatrick score, although there was a significant main effect for phase, which was driven by greater ECG-recorded heart rate relative to PPG during the first ramp-up phase.

Conclusion: Our findings support prior research demonstrating no significant impact of skin tone on PPG-measured heart rate, with significant differences between PPG- and ECG-measured heart rate emerging during dynamic changes in activity intensity. As commercial heart rate monitoring technology and software continue to evolve, it will be vital to replicate studies investigating the impact of skin tone due to the rapidity with which widely used wearable technologies advance.

KEYWORDS

physical activity, wearables, heart rate, equity, exercise

1 Introduction

Regular physical activity promotes overall health and well-being across the lifespan and accompanies improvements in cognition, learning and judgment skills, and symptoms of depression and anxiety (1, 2). According to a 2022 report from the World Health Organization, more than 80% of the world's adolescent population is insufficiently physically active, and people who do not engage in sufficient physical activity are at 20%–30% greater risk of death compared to sufficiently active people (3). Wearable activity monitors have risen in popularity both for consumers and researchers given their ability to objectively monitor key health behaviors including physical activity, sleep, and stress, which are each facilitated based on heart rate monitoring (4, 5). These are useful technologies for those interested in activity promotion, as a core component of successful activity behavior change is self-regulation, which is characterized by the motivation, control, and modification of behavior to achieve a desired goal (6). Successfully changing behavior through self-regulation depends on one's ability to accurately and consistently self-monitor their behaviors, as developing an accurate awareness of one's behaviors is a pre-requisite for supporting behavior change.

Consumer physical activity self-monitoring technologies represent an important and growing industry (7–9) and are increasingly common in clinical physical activity trials (10–12). Contemporary monitoring devices integrate various sensing technologies such as accelerometry, global positioning, and optical sensing to measure the intensity and duration of activity. These data are of immense value to those interested in developing novel and highly tailored activity programs. However, concern over whether sensors work similarly across individuals—and the potential to introduce a systematic bias when using monitoring technologies—has risen in popular consciousness in recent years. Wrist-worn consumer devices leverage light (photoplethysmography; PPG) to monitor peripheral blood flow, and there is concern that darker skin tones may affect the accuracy of PPG sensors (13, 14). A systematic review by Koerber and colleagues indicated that heart rate-sensing smartwatches were significantly less accurate when used on darker skin tones in comparison to lighter skin tones (13). In contrast, Bent and colleagues failed to identify significant differences in accuracy resulting from differing skin tones, though they did find that accuracy varied by device manufacturer and type of activity, regardless of skin tone (14).

One potential cause of heterogeneity in findings in the relationship between skin tone and PPG accuracy is the continuous evolution of heart rate monitoring hardware and software in the consumer market. This presents both challenges (e.g., changing study endpoints) and benefits (e.g., enhanced accuracy) in the research context. For instance, companies such as Garmin and Apple have implemented technologies to increase the intensity of the PPG light if a strong signal is not detected by the device, such as for individuals with darker skin (15). Because consumer technologies are continuously refined and improved upon, replication studies are critical—a notion increasingly recognized in mHealth research (16). To this end, the purpose of this study was to revisit the investigation of differences in electrocardiogram (ECG) and optical heart rate (PPG) data collected by Garmin Forerunner 45 and Polar H10 devices, respectively, across self-reported skin tone scores.

2 Methods

Participants with varying skin tones were recruited between Spring 2022 and Spring 2023 to compare heart rate recordings during rest, exercise, and recovery between wrist-worn PPG and chest ECG. Participants were recruited via word of mouth, flyers, and listserv emails within a college community in the southeast United States. Participants were eligible if they were between 18 and 64 years of age with no forearm tattoos that would interfere with the PPG sensors on both wrists. Participants were intentionally recruited such that no more than half of the sample self-identified as White. In addition, eligible participants had to be fluent in English, capable of communicating with research staff over the phone, willing to wear two Garmin wristwatches and a chest strap for approximately 35 min, and able to engage in aerobic exercise for at least 20 min. Eligible participants completed a pre-screen interview including a physical activity readiness questionnaire to confirm eligibility and subsequently were scheduled for a single testing session. At the session, and prior to all study procedures, research staff conducted the informed consent process, obtained written consent from the participants, and collected participant demographic characteristics and the Fitzpatrick scale as a proxy for skin tone.

PPG data were collected via the Garmin Forerunner 45, which leverages the same Garmin Elevate PPG technology used across all modern Garmin devices (17). We selected this device as Garmin devices are well-represented in both physical activity research (18) and in the commercial sectors (19, 20). Additionally, because Garmin devices all leverage the same PPG technology, selecting a single device equipped with this sensor offers an efficient means of investing in PPG accuracy in a large segment of the consumer wearable market. We want to emphasize the importance of expanding the work presented in this study to other widely used consumer devices. Participants were fitted with one Garmin Forerunner 45 on their left wrist and another on their right wrist. One Garmin collected data via the PPG sensor, and the second was connected to a Polar H10 ECG chest strap and therefore did not collect data via PPG. Electrode gel or water was placed on the sensors of the chest strap before being fastened to the participant. Participants remained seated for 5 min to collect their resting heart rate. Resting heart rate was then used to calculate heart rate reserve (HRR), which was computed by subtracting resting heart rate from the participant's estimated maximum heart rate, which was calculated using the formula: estimated maximum heart rate (BPM) = 220-age (21). Participants were then instructed to walk or

Abbreviations
PPG, photoplethysmography; ECG, electrocardiogram; HRR, heart rate reserve; BPM, beats per minute.

jog on an outdoor track at 60% of their estimated HRR for 10 min during an "exercise bout." Participants then walked at a self-selected light intensity for 10 min (i.e., "rest") and then initiated a second "exercise bout" wherein participants again were instructed to walk or jog at 60% of their estimated HRR for 10 min. We selected this protocol to give insight into both steady state and changing intensity given prior research demonstrating differential performance of PPG vs. ECG during intensity changes (14). Upon completion of the participant visit, data were downloaded from each Garmin Forerunner 45 and extracted via a custom Python script (22).

3 Measures

3.1 Heart rate

As noted above, Garmin Forerunner 45 devices were used to collect data during each session, with one paired to a Polar H10 ECG chest strap and the other using the on-device PPG sensor. The Polar H10 is among the most widely used chest ECG devices with excellent validity (23). Heart rate data are provided approximately every 5 s, and a custom Python script was utilized to time-match data collected via both devices. Specifically, datasets were merged based on closest matching timestamps, which were allowed to differ by up to 5 s. These data were subsequently plotted, and periods of rest, the first exercise bout, the rest bout, and the second exercise bout were identified with a timestamp and visual inspection of ECG data. As it has been reported that PPG data may be delayed relative to ECG data during changes in activity intensity (14), we investigated exercise bouts as a whole as well as subdivided into a "ramp" and "steady-state" period. Specifically, we visualized ECG-based heart rate data and classified the rapid increase in heart rate during the initial period of the exercise bout as the "ramp" period and the plateau in heart rate as "steady state." If a participant did not have a clear delineation between these stages (e.g., a consistent rise in heart rate across the bout), all activity was classified as exercise. In sum, PPG and ECG differences in a total of eight "phases" were investigated: rest, the first ramp, the first steady-state exercise bout, the first full exercise bout (comprising both the ramp and steady-state period), rest, the second ramp, the second steady-state exercise bout, and the second full exercise bout (comprising both the ramp and steady-state period). Average readings and differences for the ECG and PPG data were computed for each period. Differences were computed as ECG minus PPG, such that positive values indicated higher ECG-measured heart rate whereas negative values indicated higher PPG-measured heart rate. Notably, the Association for the Advancement of Medical Instrumentation recommends a maximum error of ±5 BPM for heart rate monitoring (24).

3.2 Skin tone

The Fitzpatrick scale was originally designed to classify how different skin types may react to ultraviolet light, though as Fine and colleagues note, the Fitzpatrick scale "is often used within the biophotonics community due to the effect eumelanin has on how light travels through skin. This is due to the high absorbance of eumelanin with a peak in the ultraviolet wavelength (220 nm) and a steady decay through the visible wavelength region" (25). Participants responded to 10 items related to physical traits (e.g., eye color and color of the skin in unexposed areas), sensitivity to sun exposure, and how often one typically engages in intentional sun exposure. Responses are provided on a 0–4 scale, with final scores ranging from 0 to 40. Six total categories are derived from these scores, ranging from pale white skin to deeply pigmented dark brown skin to black skin (26). In the present study, we investigated Fitzpatrick scores continuously (14, 27) as well as in three categories containing two Fitzpatrick types each (i.e., 0 = types I/II, corresponding to scores of 0–13; 1 = types III/IV, corresponding to scores of 14–27; 2 = types V/VI, corresponding to scores of 28–36). This scale is commonly used in research and clinical settings to classify one's skin tone (28), largely due to its availability, historic use, and ease of administration. However, while it is frequently used by healthcare providers as a means of describing skin color (29), it is notable that it was originally designed to measure the propensity of the skin to burn during phototherapy (29). Important limitations to this approach include that Fitzpatrick is often conflated with a measure of race or ethnicity and that there is a large degree of within-group variability in skin tone (29). We deem it important to note early that the use of the Fitzpatrick scale is a limitation driven by a lack of widely available tools and will explore opportunities for future research within the discussion.

We leveraged a series of descriptive analyses to contrast heart rate collected via chest ECG and wrist PPG among individuals with varying skin tones assessed via the Fitzpatrick scale. First, we present descriptive statistics, including mean (SD) for continuous variables and count (%), for the whole sample. Similarly, we computed descriptive statistics for the difference in heart rate within each phase of the exercise bout (start, first ramp, first steady-state exercise, first full exercise bout, rest, second ramp, second steady-state exercise, second full exercise bout), with descriptives presented for Fitzpatrick subgroups and the sample as a whole. Note that we observed two extreme outliers in the difference between ECG and PPG-recorded heart rate, and as such, we also present median and interquartile range in supplemental materials. Both individuals fell into the Fitzpatrick type V/VI category. Bland–Altman plots were produced for each phase to investigate differences in ECG and PPG heart rate by average heart rate. To investigate whether heart rate varied by Fitzpatrick score, phase of the exercise bout, or the interaction of the two, we next conducted a mixed ANOVA including phase as a within-subject factor and Fitzpatrick subgroup as a between-subject factor, confirming the assumption of homogeneity of variances. Finally, to investigate relationships between continuous Fitzpatrick scores and differences in heart rate by device during each phase, we computed a series of Pearson correlations that were interpreted as recommended by Evens and colleagues such that 0–0.2 was considered very weak, 0.2–0.4 was considered weak, 0.4–0.6 was considered moderate, 0.6–0.8 was considered strong, and 0.8+

was considered very strong (30); significant associations were visualized via scatterplot. Given the outlying values described above, we also present Spearman rho correlations in the supplemental materials. All analyses were completed in SPSS version 29 (IBM Corp., Armonk, NY, USA).

4 Results

Participant characteristics are displayed in Table 1. A total of 33 participants agreed to participate in the study, and 29 completed the study protocol and had complete data from both devices during the 1-year pilot period. Characteristics of these participants are displayed in Table 1. Briefly, 8 identified as Black, 4 as Asian, 15 as White, and 2 self-identified as more than one race. With regard to sex, 15 participants (52%) were male, and 14 (48%) were female. The average age of the participants was 22.24 ± 4.54 years, and the average Fitzpatrick score from the research sample was 21.45 ± 6.48.

Table 2 depicts the descriptive statistics during each phase by Fitzpatrick category. Supplementary Table S1 contains median and interquartile range values during each phase by Fitzpatrick category. Figure 1 depicts the Bland–Altman plots for each phase. Note that there was a violation of the sphericity assumption for the mixed ANOVA, and therefore we corrected degrees of freedom using the Greenhouse-Geisser $\varepsilon = 0.509$. The mixed ANOVA did not reveal a significant phase–Fitzpatrick category interaction ($P = 0.27$), nor was there a significant main effect for the Fitzpatrick category ($P = 0.68$). There was, however, a significant main effect for phase [$F_{(2.55,63.64)} = 19.84$, $P < .001$, $\eta^2 = .44$], and a series of *post hoc* contrasts revealed this was driven by significantly higher ECG-recorded heart rate relative to PPG during the first ramp phase, which was significantly larger than differences during any other phase ($P < .001$). Differences in ECG and PPG between other phases were not statistically significant. A second model wherein ramp and steady-state phases were not separated demonstrated similar results. Namely, the interaction between the Fitzpatrick category and phase was not significant ($P = 0.21$) nor was the main effect for the Fitzpatrick category ($P = 0.57$). The main effect for phase was significant [$F_{(2.16,56.13)} = 8.36$, $P < .001$, $\eta^2 = .24$], and this was driven by significantly greater differences in the first exercise phase ($Ps \leq 0.006$). Notably, several individuals had at least one extreme outlying value during at least one phase. As sensitivity analyses, we conducted the mixed ANOVAs without the outlying values, and the interpretation did not differ.

TABLE 1 Participant demographics ($N = 29$).

Characteristic	M (SD)
Age, M (SD)	22.24 (4.54)
Male, *n* (%)	15 (52)
Female, *n* (%)	14 (48)
Race, *n* (%)	
White	15 (52)
Black	8 (28)
Asian	4 (14)
More than 1	2 (7)
Fitzpatrick score, M (SD)	21.45 (6.48)
BMI M (SD)	23.19 (3.60)

M, mean; SD, standard deviation; BMI, body mass index (kg/m^2).

Regarding Bland–Altman plots (depicted in Figure 1), limits of agreement were >±5 BPM during all tasks. Table 3 depicts the number of cases during each phase with differences that exceeded ±5 BPM. The mean bias was 0.67 BPM during rest (a positive number indicating higher heart rate recorded via ECG relative to PPG), 10.82 BPM during the first ramp phase, 0.20 BPM during the first steady-state exercise phase, 3.49 BPM during the first overall exercise phase, −1.15 BPM during rest, 0.34 BPM during the second ramp phase, −1.11 during the second steady-state exercise phase, and −0.70 during the second exercise bout on the whole. Table 4 displays the Pearson correlation coefficients for the relationships between continuous Fitzpatrick scores and differences in heart rate. There was a significant, moderate negative correlation between Fitzpatrick score and heart rate differences during the second ramp phase, indicating those with darker skin tones demonstrated a relatively higher PPG score during this period. A scatterplot depicting this relationship can be observed in Figure 2. Spearman rho correlations are depicted in Supplementary Table S2, and, notably, interpretation does not meaningfully differ.

5 Discussion

This study aimed to examine the agreement between ECG and wrist-based PPG-measured heart rate, and whether agreement was affected by self-reported skin tone using the Fitzpatrick skin typing scale. Our results generally indicate that ECG- and PPG-measured heart rates did not differ by skin tone, except when individuals were increasing the intensity of activity after an active rest period. Here, those with lighter skin as reported via the Fitzpatrick skin typing scale demonstrated a similar response as to the first ramp phase (i.e., ECG detected higher heart rate) whereas those with darker skin reported relatively higher PPG-recorded heart rate.

TABLE 2 Mean and standard deviations for the difference of heart rate in each task overall and by Fitzpatrick category.

Phase	Fitzpatrick score			Total ($N = 29$)
	0–13 ($N = 5$)	14–27 ($N = 17$)	28–36 ($N = 7$)	
Start	−1.51 (6.49)	1.58 (3)	0.04 (1.97)	0.67 (3.67)
First ramp	14.35 (7.00)	8.48 (9.06)	13.99 (12.66)	10.82 (9.82)
First steady-state exercise	−0.08 (1.97)	−1 (6.48)	3.32 (8.43)	0.2 (6.57)
First full exercise bout	5.01 (3.59)	1.84 (3.87)	6.42 (9.32)	3.49 (5.76)
Rest	−1.00 (1.28)	−1.25 (2.78)	−1.01 (1.99)	−1.15 (2.35)
Second ramp	2.51 (3.77)	0.89 (3.45)[a]	−2.49 (4.62)	0.34 (4.07)[a]
Second steady-state exercise	−0.23 (2.33)	−1.56 (5.72)	−0.64 (2.69)	−1.11 (4.62)
Second full exercise bout	0.42 (2.35)	−0.92 (3.97)	−0.97 (2.89)	−0.7 (3.44)

Differences were computed as ECG – PPG such that more positive scores indicate higher ECG-recorded heart rate. Values are in average beats per minute.

[a] One observation missing due to a lack of clear delineation between the ramp and steady-state exercise and therefore time was all categorized as exercise.

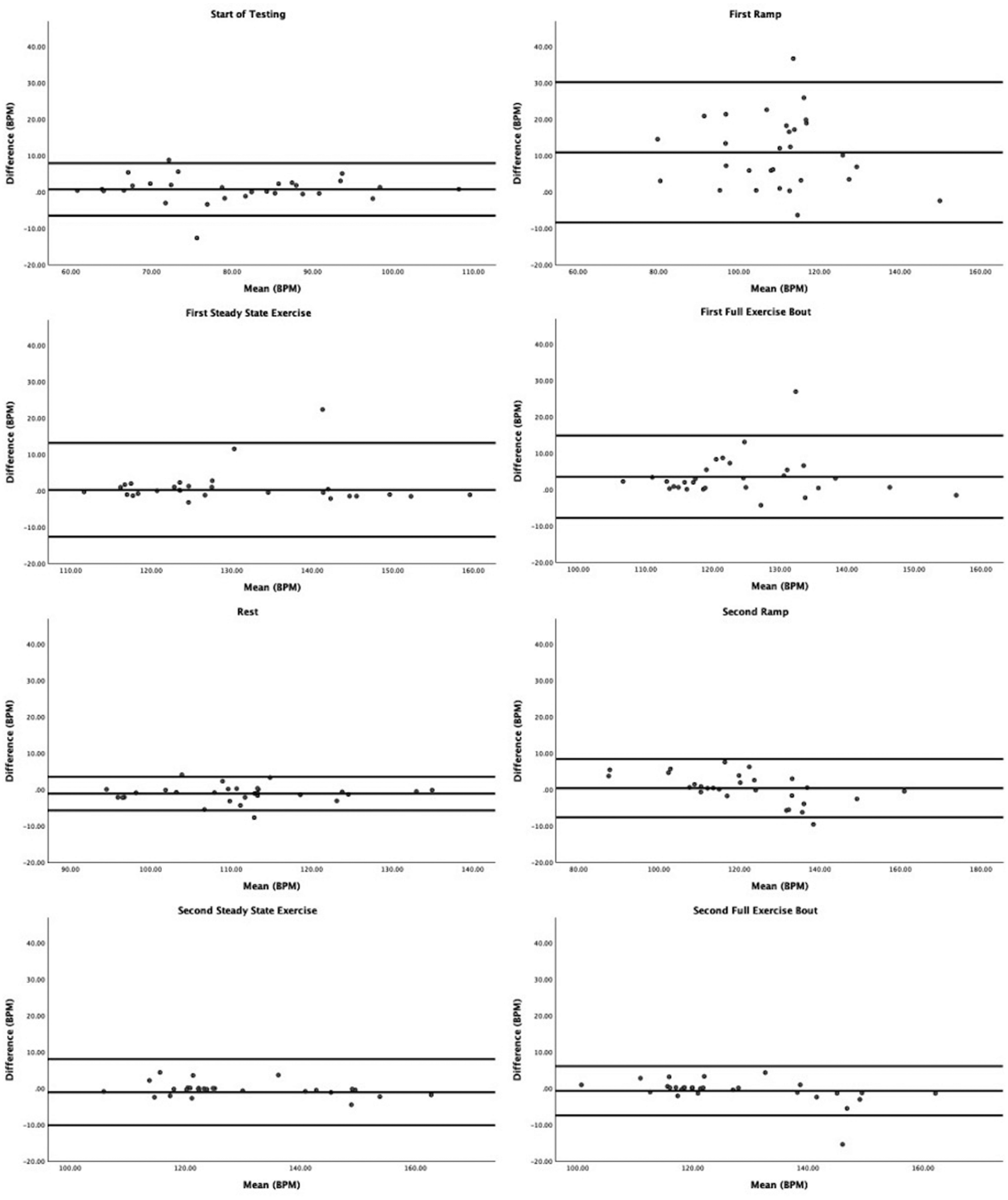


FIGURE 1
Bland–Altman plots contrasting differences in ECG vs. PPG by average across both devices. The lines are plotted for average difference and 95% confidence limits around the mean difference.

Importantly, most of these differences were small in magnitude (i.e., 71% had differences <5 BPM, which falls within an acceptable range of accuracy). More generally, we observed a range of bias (1.15–10.82 BPM) between devices, with the greatest bias observed during the initial increase in activity as individuals began their first bout of exercise from complete rest. The tendency for PPG signals to lag behind ECG during changes in heart rate is widely reported and may be attributable to several physiological causes, including a delay between changing heart rate and changes in blood volume at the wrist (14). As others have reported, this suggests utility for wrist PPG-based heart rate measurement for monitoring steady-state activity above, and less so for activities like high-intensity interval training where heart rate might rapidly accelerate and decelerate. This may be especially true for those with darker skin, as we observe a significant correlation between Fitzpatrick score and differences between ECG and PPG only when individuals

TABLE 3 Count of participants exhibiting differences between ECG- and PPG-recorded heart rate exceeding 5 BPM.

Phase	*N* (%)
Start	5 (17)
First ramp	21 (72)
First steady-state exercise	3 (10)
First full exercise bout	8 (28)
Rest	2 (7)
Second ramp	8 (28)
Second steady-state exercise	1 (3)
Second full exercise bout	2 (7)

TABLE 4 Pearson correlations between Fitzpatrick score and differences in heart rate collected via chest and wrist monitor.

Stage	Fitzpatrick score
Start	0.31
First ramp	0.10
First steady-state exercise	0.10
First full exercise bout	0.14
Rest	−0.12
Second ramp	−0.49[a]
Second steady-state exercise	−0.16
Second full exercise bout	−0.28

0–0.2, very weak; 0.2–0.4, weak; 0.4–0.6, moderate ; 0.6–0.8, strong; 0.8+, very strong.
[a]Correlation is significant at the 0.05 level (two-tailed).

increased intensity following an active recovery period. Indeed, we observed two outlying individuals who had much greater differences between ECG and PPG ratings during the ramp and/or exercise phases, and both individuals fell into the highest Fitzpatrick category.

In sum, our data, in combination with the wider body of evidence (15, 31–36), give helpful guidance to clinicians and those interested in promoting activity behavior from the perspective of selecting a heart rate monitoring device. Specifically, the use of PPG-based monitoring should be cautioned for those interested in promoting or engaging in behaviors where rapid changes in intensity are expected to be frequent. Moreover, we would note the critical importance of additional validation testing. As raised in recent reports, there remain a number of factors that could interact with skin tone to produce bias, including factors such as the presence of arm hair, ambient temperature and humidity, level of motion, skin thickness, and body mass (31). As researchers gain access to better tools to measure these factors in the field and as consumer wearable devices continue to enter the market, it will be critical to revisit this topic.

5.1 Strengths and limitations

There are several important strengths to the present study. First, participants completed study procedures outdoors while

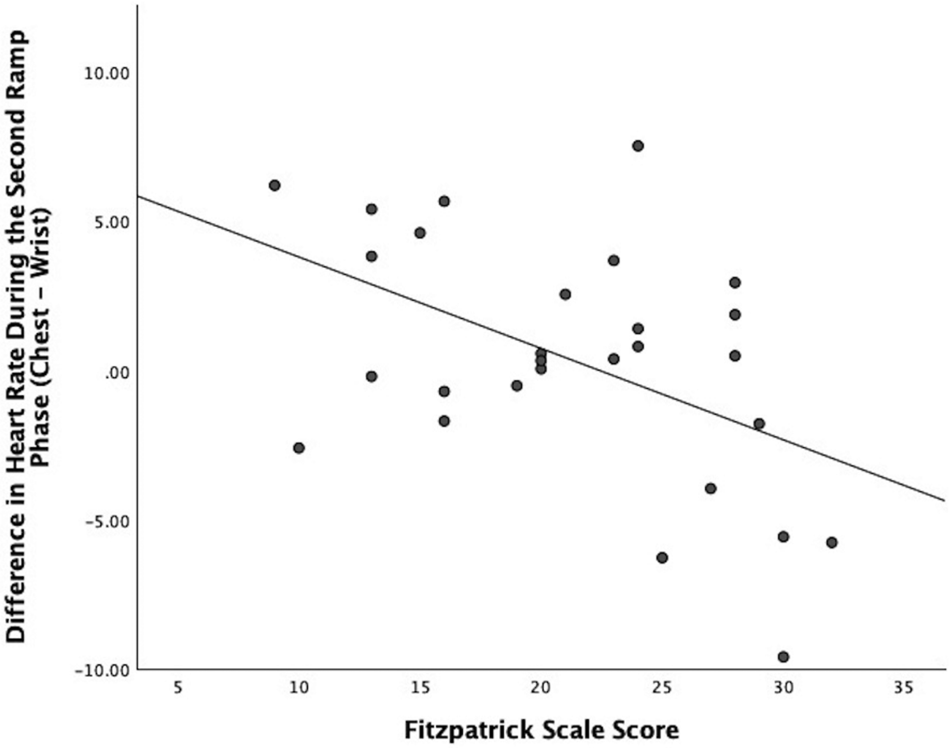


FIGURE 2
Scatter plot showing the relationship between participant Fitzpatrick scores and the difference between ECG and PPG heart rate values during the second ramp phase.

walking overground, capturing the potential negative impact of uneven motion and light on the accuracy of wrist data, which may be missed in the laboratory. Additionally, our protocol allowed for the investigation of both steady state and changes in activity intensity. This provides the ability to challenge the validity of these devices over varying conditions and various phases of activity to properly capture every phase of activity that the PPG sensor would record. Of course, there are several important limitations to consider. As described earlier, there are well-documented criticisms of the Fitzpatrick skin typing scale related to how values are interpreted and variability in tone within categories. Given the potential harm produced by biases in widely used health technologies, it is promising that researchers are actively working to create more representative cost-efficient scales, such as the newly developed 10-shade Monk skin tone scale, which became available for use following the completion of data collection for the study presented herein (37), and relatively cost-efficient and portable technologies such as the Delfin Skin Color Catch that researchers have successfully used to observe skin pigmentation (38–40). In combination, these tools may facilitate still further replication work to address several of the research gaps (31). These include investigating interactions between skin tone, arm hair, perspiration, and body mass among other potentially confounding variables. Second, as our research occurred on the campus of a small liberal arts campus, participants were college-aged adults, limiting age diversity in our sample. Extending this work to older adults may cause other discrepancies, as older adults tend to have stiffer arteries, weaker blood flow, and thinner skin (41). Third, our sample was relatively small, which may influence the stability of our findings and the width of our limits of agreement. We acknowledge that a larger, balanced sample size would yield stronger conclusions and also better consider the individual variability in PPG accuracy (25). However, we would note our findings are in line with other recent studies on the topic and are encouraged by the consistent results we observed within our diverse sample (34, 42). Fourth, the discrepancy between PPG and ECG may be subject to motion factors, although this is outside the scope of our study (43). This minor limitation was not assessed in the data processing because all our participants did the same activities, and we are focusing on skin tone. Finally, we did not quantify weather conditions during testing, and evidence suggests that both temperature and humidity may affect the quality of a PPG signal (44). It may be valuable for future research to examine whether there are any interactions between temperature, humidity, and skin pigmentation on PPG accuracy.

6 Conclusion

Wearable devices have become a mainstay in clinical trials research and in the consumer sector, and as such, understanding whether and to what extent important characteristics such as skin tone may introduce a bias into heart rate measurement is critical. Herein, we present further support that PPG and ECG-measured heart rates generally exhibit low bias but wide limits of agreement, with differences being exaggerated as activity intensity changes, but generally not varying by the skin tones represented in our sample. These findings are encouraging, supporting the utility of accessible and inexpensive PPG heart rate measurement in health research, especially when one is interested in heart rate at rest or during steady-state activity. Given the rapidity with which widely used wearable technologies advance, it will be critical that researchers routinely replicate research meant to capture any potential bases introduced by the use of these devices.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Ethics statement

This study involving humans was approved by the Institutional Review Board of Wake Forest University. This study was conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

AI: Writing – original draft, Writing – review & editing. CM: Writing – original draft, Writing – review & editing. AB: Writing – review & editing. MI: Writing – review & editing. KN: Writing – review & editing. JR: Writing – review & editing. JF: Writing – review & editing.

Funding

The authors declare that financial support was received for the research and/or publication of this article. This work was (partially) supported by the Wake Forest University Claude D. Pepper Older Americans Independence Center (P30-AG21332).

Acknowledgments

This research study was approved by the Institutional Review Board of Wake Forest University. Equipment was provided by the Behavioral Medicine Laboratory through the Health and Exercise Department. Student researchers Emma Costello and Rafa Capelate De Oliviera partially assisted in data collection.

Conflict of interest

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

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2025.1553565/full#supplementary-material

References

1. Mandolesi L, Polverino A, Montuori S, Foti F, Ferraioli G, Sorrentino P, et al. Effects of physical exercise on cognitive functioning and wellbeing: biological and psychological benefits. *Front Psychol*. (2018) 9:509. doi: 10.3389/fpsyg.2018.00509

2. Blair SN, Jacobs DR, Powell KE. Relationships between exercise or physical activity and other health behaviors. *Public Health Rep*. (1985) 100(2):172–80.

3. World Health Organization. Physical activity. (2024). Available at: https://www.who.int/news-room/fact-sheets/detail/physical-activity (accessed February 16, 2025)

4. Hao T, Walter KN, Ball MJ, Chang HY, Sun S, Zhu X. Stresshacker: towards practical stress monitoring in the wild with smartwatches. *AMIA Annu Symp Proc*. (2017) 2017:830–8.

5. Hirten RP, Danieletto M, Tomalin L, Choi KH, Zweig M, Golden E, et al. Factors associated with longitudinal psychological and physiological stress in health care workers during the COVID-19 pandemic: observational study using apple watch data. *J Med Internet Res*. (2021) 23(9):e31295. doi: 10.2196/31295

6. Bandura A. Social cognitive theory of self-regulation. *Organ Behav Hum Decis Process*. (1991) 50(2):248–87. doi: 10.1016/0749-5978(91)90022-L

7. Reeder B, David A. Health at hand: a systematic review of smart watch uses for health and wellness. *J Biomed Inform*. (2016) 63:269–76. doi: 10.1016/j.jbi.2016.09.001

8. Henriksen A, Haugen Mikalsen M, Woldaregay AZ, Muzny M, Hartvigsen G, Hopstock LA, et al. Using fitness trackers and smartwatches to measure physical activity in research: analysis of consumer wrist-worn wearables. *J Med Internet Res*. (2018) 20(3):e110. doi: 10.2196/jmir.9157

9. Jung Y, Kim S, Choi B. Consumer valuation of the wearables: the case of smartwatches. *Comput Human Behav*. (2016) 63:899–905. doi: 10.1016/j.chb.2016.06.040

10. Wang R, Blackburn G, Desai M, Phelan D, Gillinov L, Houghtaling P, et al. Accuracy of wrist-worn heart rate monitors. *JAMA Cardiol*. (2017) 2(1):104. doi: 10.1001/jamacardio.2016.3340

11. Masanneck L, Gieseler P, Gordon WJ, Meuth SG, Stern AD. Evidence from ClinicalTrials.gov on the growth of digital health technologies in neurology trials. *NPJ Digit Med*. (2023) 6(1):23. doi: 10.1038/s41746-023-00767-1

12. Marra C, Chen JL, Coravos A, Stern AD. Quantifying the use of connected digital products in clinical research. *NPJ Digit Med*. (2020) 3(1):50. doi: 10.1038/s41746-020-0259-x

13. Koerber D, Khan S, Shamsheri T, Kirubarajan A, Mehta S. Accuracy of heart rate measurement with wrist-worn wearable devices in various skin tones: a systematic review. *J Racial Ethn Health Disparities*. (2023) 10(6):2676–84. doi: 10.1007/s40615-022-01446-9

14. Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ Digit Med*. (2020) 3(1):18. doi: 10.1038/s41746-020-0226-6

15. Boxall A. Digital Trends. (2021). Skin Tone, Heart Rate Sensors, and a Push for Accuracy. Available at: https://www.digitaltrends.com/mobile/does-skin-tone-affect-ppg-heart-rate-sensor-accuracy/#dt-heading-the-heart-rate-sensor-on-your-wrist (Accessed January 20, 2024).

16. Bonnechère B, Kossi O, Mapinduzi J, Panda J, Rintala A, Guidetti S, et al. Mobile health solutions: an opportunity for rehabilitation in low- and middle income countries? *Front Public Health*. (2022) 10:1072322. doi: 10.3389/fpubh.2022.1072322

17. Heart Rate Monitoring | Garmin Technology. (cited 2025 February 17). Available online at: https://www.garmin.com/en-US/garmin-technology/health-science/heart-rate-monitoring/

18. Evenson KR, Spade CL. Review of validity and reliability of Garmin activity trackers. *J Meas Phys Behav*. (2020) 3(2):170–85. doi: 10.1123/jmpb.2019-0035

19. Statista. *Garmin Revenue Share Worldwide from 2009 to 2022, by Segment*. Hamburg: Statista GmbH (2023).

20. Statista. *Wearables: Garmin Users in the United States*. Hamburg: Statista (2024). Available online at: https://www.statista.com/study/74035/wearables-garmin-users-in-the-united-states/

21. Fletcher GF, Ades PA, Kligfield P, Arena R, Balady GJ, Bittner VA, et al. Exercise standards for testing and training: a scientific statement from the American Heart Association. *Circulation*. (2013) 128(8):873–934. doi: 10.1161/CIR.0b013e31829b5b44

22. Fanning J. GitHub. Garmin PPG. (2024). Available online at: https://github.com/jsnfng/GarminPPG.git (Accessed July 9, 2024).

23. Gilgen-Ammann R, Schweizer T, Wyss T. RR interval signal quality of a heart rate monitor and an ECG Holter at rest and during exercise. *Eur J Appl Physiol*. (2019) 119(7):1525–32. doi: 10.1007/s00421-019-04142-5

24. Blok S, Piek MA, Tulevski II, Somsen GA, Winter MM. The accuracy of heartbeat detection using photoplethysmography technology in cardiac patients. *J Electrocardiol*. (2021) 67:148–57. doi: 10.1016/j.jelectrocard.2021.06.009

25. Fine J, Branan KL, Rodriguez AJ, Boonya-ananta T, Ajmal A, Ramella-Roman JC, et al. Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring. *Biosensors*. (2021) 11(4):126. doi: 10.3390/bios11040126

26. Fors M, González P, Viada C, Falcon K, Palacios S. Validity of the Fitzpatrick skin phototype classification in Ecuador. *Adv Skin Wound Care*. (2020) 33(12):1–5. doi: 10.1097/01.ASW.0000721168.40561.a3

27. Ray I, Liaqat D, Gabel M, De Lara E. Skin tone, confidence, and data quality of heart rate sensing in WearOS smartwatches. *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops)*. Kassel, Germany: IEEE (2021). p. 213–9. Available online at: https://ieeexplore.ieee.org/document/9431120/

28. Santiago S, Brown R, Shao K, Hooper J, Perez M. Modified Fitzpatrick scale-skin color and reactivity. *J Drugs Dermatol*. (2023) 22(7):641–6. doi: 10.36849/JDD.6859

29. Ware OR, Dawson JE, Shinohara MM, Taylor SC. Racial limitations of Fitzpatrick skin type. *Cutis*. (2020) 105(2):77–80.

30. Evans J. *Straightforward Statistics for the Behavioral Sciences*. Pacific Grove, Calif: Brooks/Cole (1996).

31. Colvonen PJ. Response to: investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ Digit Med*. (2021) 4(1):38. doi: 10.1038/s41746-021-00408-5

32. Williams GJ, Al-Baraikan A, Rademakers FE, Ciravegna F, van de Vosse FN, Lawrie A, et al. Wearable technology and the cardiovascular system: the future of patient assessment. *Lancet Digit Health*. (2023) 5(7):e467–76. doi: 10.1016/S2589-7500(23)00087-0

33. Marzano-Felisatti JM, De Lucca L, Priego-Quesada JI, Pino-Ortega J. Heart rate measurement accuracy during intermittent efforts under laboratory conditions: a comparative analysis between chest straps and armband. *Appl Sci*. (2024) 14(24):11872. doi: 10.3390/app142411872

34. Etiwy M, Akhrass Z, Gillinov L, Alashi A, Wang R, Blackburn G, et al. Accuracy of wearable heart rate monitors in cardiac rehabilitation. *Cardiovasc Diagn Ther*. (2019) 9(3):26271. doi: 10.21037/cdt.2019.04.08

35. Claes J, Buys R, Avila A, Finlay D, Kennedy A, Guldenring D, et al. Validity of heart rate measurements by the Garmin Forerunner 225 at different walking intensities. *J Med Eng Technol*. (2017) 41(6):480–5. doi: 10.1080/03091902.2017.1333166

36. Pasadyn SR, Soudan M, Gillinov M, Houghtaling P, Phelan D, Gillinov N, et al. Accuracy of commercially available heart rate monitors in athletes: a prospective study. *Cardiovasc Diagn Ther*. (2019) 9(4):379–85. doi: 10.21037/cdt.2019.06.05

37. Monk E. *The Monk Skin Tone Scale*. Charlottesville: Open Science Framework (OSF) (2024).

38. Parvizi MM, Saki N, Samimi S, Radanfer R, Shahrizi MM, Zarshenas MM. Efficacy of castor oil cream in treating infraorbital hyperpigmentation: an exploratory single-arm clinical trial. *J Cosmet Dermatol*. (2024) 23(3):911–7. doi: 10.1111/jocd.16056

39. Afzal N, Nguyen N, Min M, Egli C, Afzal S, Chaudhuri RK, et al. Prospective randomized double-blind comparative study of topical acetyl zingerone with tetrahexyldecyl ascorbate versus tetrahexyldecyl ascorbate alone on facial photoaging. *J Cosmet Dermatol*. (2024) 23(7):2467–77. doi: 10.1111/jocd.16292

40. Xu Z, Wang C, Xing X, Zhang C, Xiang LF. Efficacy and safety of the combination of oral tranexamic acid and intense pulsed light versus oral tranexamic acid alone in the treatment of refractory Riehl's melanosis: a prospective, comparative study. *J Cosmet Dermatol*. (2024) 23(6):2049–57. doi: 10.1111/jocd.16257

41. Pi I, Pi I, Wu W. External factors that affect the photoplethysmography waveforms. *SN Appl Sci*. (2021) 4(1):21. doi: 10.1007/s42452-021-04906-9

42. Sañudo B, De Hoyo M, Muñoz-López A, Perry J, Abt G. Pilot study assessing the influence of skin type on the heart rate measurements obtained by photoplethysmography with the Apple Watch. *J Med Syst*. (2019) 43(7):195. doi: 10.1007/s10916-019-1325-2

43. Han H, Kim J. Artifacts in wearable photoplethysmographs during daily life motions and their reduction with least mean square based active noise cancellation method. *Comput Biol Med*. (2012) 42(4):387–93. doi: 10.1016/j.compbiomed.2011.12.005

44. Khan M, Pretty CG, Amies AC, Elliott R, Shaw GM, Chase JG. Investigating the effects of temperature on photoplethysmography. *IFAC-PapersOnLine*. (2015) 48(20):360–5. doi: 10.1016/j.ifacol.2015.10.166

frontiers | Frontiers in Digital Health

TYPE Brief Research Report
PUBLISHED 29 April 2025
DOI 10.3389/fdgth.2025.1520971

Feasibility, adherence and usability of an observational digital health study built using Apple's ResearchKit among adults aged 18–84 years

B. Brady[1,2,3]*, S. Zhou[1,2,3], D. Ashworth[1,3], L. Zheng[1,2,3], R. Eramudugolla[1,2,3] and K. J. Anstey[1,2,3]

[1]School of Psychology, The University of New South Wales Sydney, Kensington, NSW, Australia, [2]Neuroscience Research Australia, Randwick, NSW, Australia, [3]UNSW Ageing Futures Institute, UNSW Sydney, Kensington, NSW, Australia

Check for updates

OPEN ACCESS

EDITED BY
Cederick Landry,
Université de Sherbrooke, Canada

REVIEWED BY
Colin K. Drummond,
Case Western Reserve University,
United States
Verlumun Celestine Gever,
Nile University of Nigeria, Nigeria

*CORRESPONDENCE
B. Brady
✉ b.brady@unsw.edu.au

RECEIVED 01 November 2024
ACCEPTED 14 April 2025
PUBLISHED 29 April 2025

CITATION
Brady B, Zhou S, Ashworth D, Zheng L, Eramudugolla R and Anstey KJ (2025) Feasibility, adherence and usability of an observational digital health study built using Apple's ResearchKit among adults aged 18–84 years.
Front. Digit. Health 7:1520971.
doi: 10.3389/fdgth.2025.1520971

Objective: This study evaluated the Labs Without Walls app and paired Apple Watch devices for remote research among Australian adults aged 18–84.
Methods: The study app, built using Apple's open-source ResearchKit frameworks, uses a multi-timescale measurement burst design over 8-weeks. Participants downloaded the app, completed tasks over 8 weeks, and wore Apple Watch devices. Feasibility was assessed by recruitment, remote consent, and data collection without training. Adherence was measured by task completion rates. Usability was assessed by response times, a post-study survey, and qualitative feedback.
Results: 228 participants (mean age 53, age range 18–84; 62.7% female) were recruited nationwide, consented remotely, and provided data. 201 (88.16%) completed the 8-week protocol. Task adherence ranged from 100% to 70.61%. Health, environmental, and sleep data were collected passively. Usability feedback was excellent, with 84% rating the app as "extremely" or "a lot" user-friendly, 88% finding alert frequency "just right," and 95.7% finding the schedule manageable. Few age or sex differences were found.
Conclusions: The Labs Without Walls app and paired Apple Watch devices are user-friendly and enable adults aged 18–84 to complete surveys, cognitive and sensory tasks, and provide passive health and environmental data. The app can be used without formal training by males and females living in Australia, including older adults. Future iterations should consider gamification and strategies to improve daily-diary survey user experience.

KEYWORDS
life-course, digital health, mHealth, mobile app, usability

Introduction

ResearchKit, a software framework developed by Apple, allows researchers to create research apps for iOS devices. It provides a set of tools for building apps that can collect data from participants, such as survey templates, and pre-built cognitive and sensory tasks. ResearchKit easily integrates with HealthKit, allowing researchers to access health data from participants' iPhones and paired Apple Watch devices. Being open-source, researchers can also customise and extend ResearchKit to fit their specific needs. The rise of digital health has fundamentally transformed health promotion, offering innovative avenues for data collection and intervention delivery. Mobile technologies, like those supported by ResearchKit, are pivotal in this transformation, enabling researchers to reach diverse populations and gather rich, real-time data. This context highlights the growing importance of understanding the feasibility and acceptability of digital health tools in research. In this study, we used ResearchKit to create a research app, Labs Without Walls (1), to collect novel data on micro-longitudinal ageing processes among Australian adults aged 18–84.

Micro-longitudinal studies, which involve repeated measurements over various time scales (2–4), are essential for understanding human development across the lifespan. Lifespan developmental theories (5, 6) suggest that human development is a continuous process throughout life, with individual variations in developmental patterns. Mobile technologies, such as smartphones and watches, offer several advantages for conducting these studies, including improved accessibility, engagement, and temporal granularity (7–9).

Evaluating digital health interventions often begins with assessing feasibility, task adherence, and usability (10). These evaluations can identify methodological elements that are acceptable for different participants and contribute to data quality over time. Benchmarks for success can vary widely, influenced in part by wide variation in the nature, intensity and duration of digital health studies. For example, a review of participant engagement in mobile app interventions found an average overall study retention rate of 67.83% from 54 included studies (11). Study retention ranged from 14% in a mental health study among 348 participants across 12 weeks (12) to 100% retention in a weight loss study among 12 participants across four weeks (13).

The acceptability of research apps and wearables, including task adherence and usability, might vary by participant age or sex. While some studies suggest potential differences, the literature lacks evidence on age or sex differences in multi-timescale measurement burst designs among life-course samples over extended periods.

Demonstrating the acceptability of research apps built with ResearchKit is crucial. Despite age-related differences in digital literacy (14), research has shown that older adults can effectively use digital technologies. For example, a review by Wrzus and Neubauer (15) found no clear age-related trend in compliance rates in ecological momentary assessment (EMA) studies, though women were generally more compliant than men. This aligns with research on gender differences in conscientiousness (16). By demonstrating the acceptability of research apps and wearables across the lifespan, researchers can challenge stereotypes and expand the potential reach of research to hard-to-reach populations, including older adults and others who may not usually be included in research.

This study aims to evaluate the feasibility, adherence, and usability of the Labs Without Walls research app (1) and paired Apple Watch devices (Apple Inc) for studying micro-longitudinal processes among Australian adults aged 18–84 over an 8-week period.

We pre-registered the following hypotheses:

- **H1 (Feasibility, Adherence):** Participants aged 18–84 years will be able to be successfully e-consented, able to input survey data through the Labs Without Walls research app, and have passive data collected using an Apple Watch (Apple Inc) over 8 weeks.
- **H2 (Usability):** The user experience of the Labs Without Walls research app and Apple Watch (Apple Inc) will be rated as acceptable by research participants aged 18–years.

Materials and methods

This study was approved by the University of New South Wales Human Research Ethics Committee (approval number HC200792). The study design and hypotheses were preregistered on May 4, 2022, using Open Science Framework, before completing data collection. The study protocol is published elsewhere (1).

Participants

228 Australian adults (18–84) participated in the 8-week study using the Labs Without Walls app. Sample size was estimated based on thresholds of.05 (two-tailed) probability of rejecting the null hypothesis and power of.80, and a previous meta-analysis which estimated the odds ratio of subjective age (one of the primary interests of this broader project) impacting overall health to be 1.57 (17). G*Power determined a minimum sample size of 129. We over-recruited to account for covariates and potential attrition. Participants were recruited through social media, mailing lists, and volunteer databases. Eligible participants (aged 18–85, residing in Australia, owning an iPhone, not requiring text-to-speech to use iPhone) were invited to download the app. Non-responders were followed up with three attempts. Informed e-consent was obtained, and explicit permissions were required for passive data collection on health and environmental measures.

Design

As described in the study protocol (1), the research app was built for iOS using customised templates provided by Apple ResearchKit (Apple Inc). Amazon Web Services was used to host secure back-end data collection. All participants were provided with an Apple Watch Series 5 (Apple Inc) and Apple wired EarPods (Apple Inc) to use for the duration of the study. Participants returned the Apple Watch (but not the EarPods) at the end of the study, with postage paid for by the study team. Over eight weeks, participants completed a multi-timescale measurement burst protocol, including a baseline survey, repeated surveys on COVID-19 experiences, week-long daily survey sprints which explored daily subjective aging and gender expression, repeated game-like cognitive and sensory tasks, and an end of study usability survey. Participants also provided passively collected health and environmental data from the iPhone and Apple Watch (Apple Inc). Following the baseline survey, study tasks were intended to take no more than a few minutes per day to complete. Further details regarding the technical architecture of the app, study tasks and schedule are reported elsewhere (1).

Outcomes

Feasibility

Feasibility was assessed by the ability to remotely recruit a life-course sample, the location of participants indicating the ability to recruit from a wide geographic area, and overall completion rates. Our benchmarks were successful enrolment and retention of adults aged between 18 and 85 years, recruitment from a wider geographical area than traditional lab-based methods, and a completion rate of 68% or higher (11).

Adherence

Adherence was assessed by task completion rates and data completeness. Lower completion rates might indicate difficulty completing tasks in the context of their daily lives (18) or declining adherence over time (19). Incomplete health, environmental, or sleep data might indicate less compliance with the study protocol. Our benchmark for task adherence was 60% completion. We did not set specific benchmarks for passively collected data but anticipated higher completeness for daily behaviours and non-optional tasks.

Usability

Usability was assessed by task completion times, an end-of-study survey, and qualitative feedback. Our benchmark for task completion times was alignment with estimated times. For the survey, we aimed for 80% positive ratings for study schedule manageability, watch wearability, usability, alert frequency, and setup/charging ease. Qualitative feedback was sought for future iterations.

Statistics

Descriptive statistics (frequencies and percentages or means and standard deviations) were used to describe the characteristics of the sample, geographic spread of participants, study completion, task adherence, and the amount of health and environmental data collected from participants across study days. Survey and task completion times were presented as a median for the full sample, to avoid skew due to outliers (e.g., where a study survey remained open and incomplete for several hours). Due to a higher number of females than males in the study sample, Independent-samples Mann–Whitney *U*-tests were used to compare males and females on study outcomes. Linear regression analyses explored the relationship between age-in-years and continuous outcomes. Logistic regression explored the relationship between age and binary outcomes. To allow for possible non-linear effects of age, a quadratic age term was entered into each regression model. Pairwise deletion was used to account for missing data. All statistical analyses were completed using SPSS version 27 (20). Qualitative data provided by participants was reviewed and coded according to the topic(s) raised in each comment. Codes were then used to quantify the frequency of mentions of each topic, and illustrative comments were reported verbatim.

Results

Participants

As shown in Supplementary Figure S1, 500 participants expressed interest in joining the study between May 2021 and February 2023. Of those, 342 met our inclusion criteria and were invited to download the Labs Without Walls app and join the study. 228 participants provided study data. Sociodemographic characteristics are summarised in Table 1. The sample was more highly educated and included slightly lower rates of White adults than the general Australian population (21).

Location of participants

Participants were recruited from all but one of Australia's States and Territories, spanning the breadth of the continent. 2.65% joined the study from the Australian Capital Territory, 63.27% from New South Wales, 12.39% from Queensland, 4.87% from South Australia, 1.17% from Tasmania, 8.85% from Victoria, and 4.87% from Western Australia. Participants were mostly located in urban centres or regional coastal areas, reflecting Australia's population density.

General study completion

201 participants completed the Day 56 sprint survey, suggesting an overall study completion rate of 88.16%. A logistic regression analysis was conducted to examine the effect of age and its squared term on the likelihood of completing the final study day. Note, the Wald statistic reported below, calculated as the square of the ratio of the regression coefficient to its standard error, is used in association with p values to assess the statistical significance of the predictor variable (in this case, age). Neither age in years (B = 0.116, SE = 0.070, Wald = 2.754, $p = .097$, OR = 1.123), nor the age-squared term (B = -.001, SE = 0.001, Wald = .779, $p = .377$, OR = .999) were significant, suggesting that there was no linear or non-linear effect of age on likelihood of completing the day 56 survey. Males and females did not differ in the proportion who completed the Day 56 survey, $\chi 2 = 2.44$, $p = .118$.

Adherence

Surveys, cognitive and sensory tasks, and sprints

Figure 1 shows the percentage of the sample who completed each survey, cognitive and sensory task and sprint day.

TABLE 1 Sociodemographic characteristics of the Labs Without Walls sample (*N* = 228).

Sociodemographic characteristic	*n* (%) or mean (SD)
Age in years	53.09 (18.43), range 18 to 84 years
Years of education[a]	17.84 (3.90) range 8 to 32 years
Sex-at-birth	
Female	143 (62.7%)
Male	85 (37.3%)
Gender Identity	
Man	84 (36.8%)
Woman	139 (61.0%)
Non-binary or gender fluid	4 (1.8%)
Another term (no self-description provided)	1 (0.4%)
Race/Ethnicity	
Arab/West Asian	4 (1.75%)
Black	1 (0.44%)
East Asian	15 (6.58%)
Hispanic/Latin American	2 (0.88%)
South Asian	13 (5.7%)
South-East Asian	8 (3.51%)
White/Caucasian	177 (77.63%)
Other Identity: Arab/West Asian and White/Caucasian	1 (0.44%)
Other Identity: Black and White/Caucasian	1 (0.44%)
Other Identity: East Asian and Arab/West Asian	1 (0.44%)
Other Identity: East Asian and South-East Asian	1 (0.44%)
Other Identity: East Slavic and White/Caucasian	1 (0.44%)
Other Identity: Hispanic/Latin American and White/Caucasian	1 (0.44%)
Other Identity: Jewish	1 (0.44%)
Other Identity: Mediterranean/Southern European	1 (0.44%)
Relationship Status	
Married	111 (48.7%)
In a relationship	45 (19.7%)
Single	51 (22.4%)
Widowed	10 (4.4%)
Divorced	11 (4.8%)
Household Income	
<$300 per week	5 (2.2%)
$300 - $575 per week	16 (7.0%)
$576 - $1,075 per week	45 (19.7%)
$1,076 - $1,700 per week	43 (18.9%)
$1,701 - $2,400 per week	34 (14.9%)
>$2,400 per week	67 (29.4%)
Don't know	18 (7.9%)
Employment	
Employed	142 (62.3%)
Unemployed	17 (7.5%)
Retired	69 (30.3%)

[a]Years of education indicates a sum of the number of years participants reported attending primary school, secondary school, TAFE and/or University.

Adherence ranged from 100% for the baseline survey, to 70.61% for each of the Tone Audiometry tests.

Apple watch, health, and environmental data

The percentage of the sample who wore an Apple Watch (Apple Inc) and provided health and environmental data is presented in Figure 2. Watches were worn for a median of 55/56 days (range 0 to 56 days) and for an average 16.76 h (SD = 5.38) on the days worn. Over the course of the 8-week study, the percentage of the sample who wore the watch each day fluctuated from 96.05% ($n = 219$) on Day 1, to 79.39% ($n = 181$) on Day 56.

Independent-samples Mann–Whitney *U*-tests revealed that males and females did not differ in the median number of days that the watch was worn, standardised U = 1.844, $p = .260$, or the median number of hours that the watch was worn per day, standardised U = −.013, $p = .989$.

Separate linear regression analyses were conducted to explore the relationship between a) total number of days the watch was worn, and b) total number of hours worn per day and age, including a quadratic term to account for potential non-linear effects of age. Only the regression model predicting total number of days the watch was worn was significant, F(2, 227) = 15.714, $p < .001$, and accounted for approximately 12.3% of the variance in user satisfaction ($R^2 = .123$). Age in years was a significant predictor, $\beta = 1.097$, t(227) = 3.936, $p < .001$. The positive coefficient suggests that higher age was associated with higher number of days wearing the Apple Watch. The quadratic age term was also significant, $\beta = -0.009$, t(227) = −3.200, $p = .002$, indicating a non-linear relationship between age and number of days wearing the watch showing that the positive relationship between age and days of wear diminishes at higher ages (see Supplementary Figure S2). Age did not predict the average number of hours that the watch was worn each day.

Sleep tracking

132 participants provided sleep tracking data. Excluding those who did not provide sleep data, sleep was tracked for a median of 7.00 nights (M = 8.59, SD = 7.13, range 1–36 nights). Independent-samples Mann–Whitney *U*-tests showed that males and females did not differ in the median number of days that sleep was tracked, standardised $U = .319$, $p = .750$. Linear regression was conducted to explore the relationship between the number of nights that sleep was tracked and age. Neither age in years [$\beta = -.181$, $t(227) = -1.257$, $p = .210$] or the quadratic age term [$\beta = .001$, $t(227) = .810$, $p = .419$] were significant.

Usability

Median completion times for surveys and active tasks were as expected (see Supplementary Table S1). 183 participants started the optional end of study usability module, and 180 provided complete data regarding usability. Figure 3 summarizes the usability feedback. Most of the sample reported that the assessment schedule was manageable in the context of their daily life, and that they wore an Apple Watch (Apple Inc) during the study. Due to a lack of variability in these responses, we were not able to look for age or sex differences in these variables.

As shown in Supplementary Table S2, Independent-Samples Mann–Whitney *U*-tests showed that the distribution of responses to usability questions were the same for males and females for most items rated. However, males and females differed in the frequency of responses to two items: The comfort of the Apple

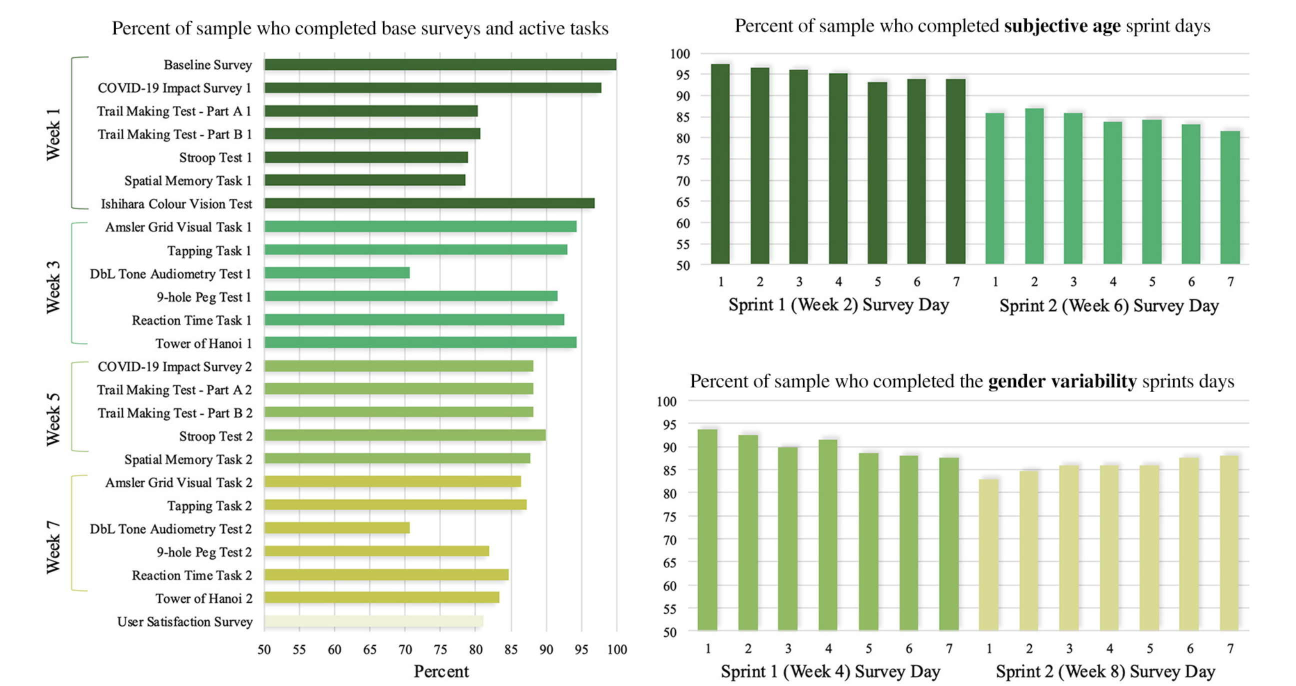


FIGURE 1
Percent of Labs Without Walls Sample (*N* = 228) who completed each survey, cognitive task, sensory task, and sprint.

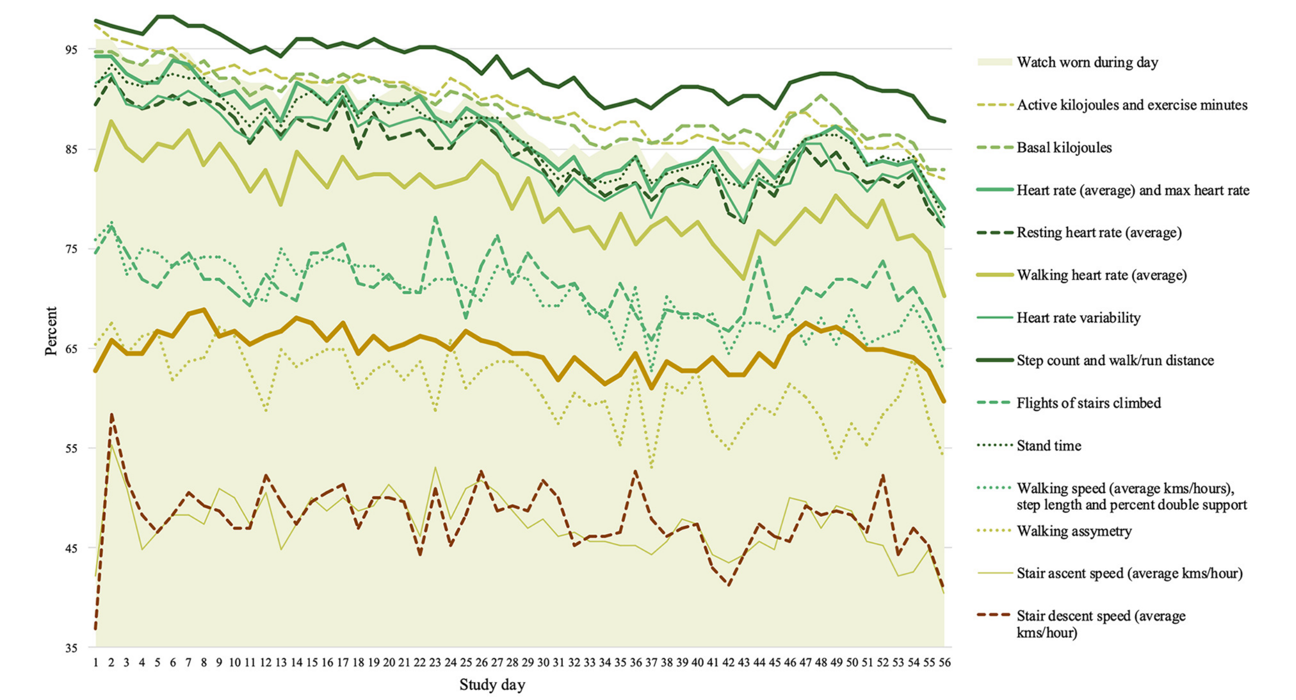


FIGURE 2
Percent of the total Labs Without Walls sample (*N* = 228) who wore the Apple Watch and provided health and environmental data per study day.

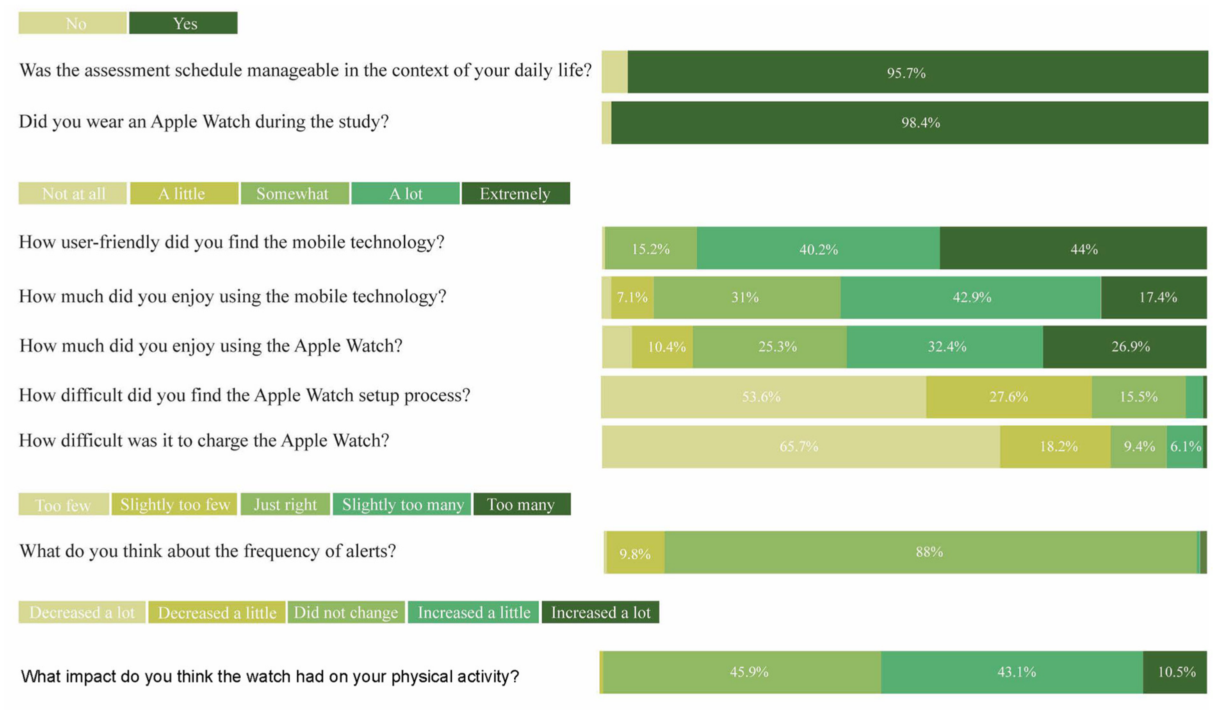


FIGURE 3
Usability feedback on the Labs Without Walls study, as a percentage of participants who endorsed each response option.

Watch (which was rated as slightly more comfortable by a higher proportion of males), and difficulty charging the Apple Watch (which was rated as marginally more difficult by males).

Separate linear regression analyses were conducted to explore the relationship between usability outcomes and age. Only the regression model predicting perceptions of the frequency of alerts was significant, $F(2, 178) = 7.076$, $p = .001$, and accounted for approximately 7.4% of the variance in user satisfaction ($R^2 = .074$). Age was a significant predictor, $\beta = 0.23$, $t(178) = 2.390$, $p = .018$ such that higher age was associated with slightly greater satisfaction with the frequency of alerts. The quadratic age term was not significant, $\beta = -0.00$, $t(178) = -1.854$, $p = .065$. Age did not predict any other usability outcome.

Summary of qualitative feedback on the study

122 participants provided an optional typed response when asked if they had any other feedback they would like to share regarding the study. Of those, 24 provided comments on the study devices, including the iPhone and/or Apple Watch (Apple Inc).

11 participants noted difficulty completing tasks or surveys based on device characteristics,

> Some tasks were not really suitable for phones with a smaller screen (I have an iPhone 7)

> I found the fine motor task didn't work particularly well on my phone for some reason. Really enjoyed everything else though!

Three participants also noted that the Apple Watch required charging more often than expected,

> I found the watch ran out of charge a bit often.

12 participants commented on a potential or experienced technical issue. Of the specific technical issues mentioned, five participants experienced minor display issues with the custom keyboards used to respond to some survey questions or active tasks,

> The correction button wasn't visible so if a typo occurred the task had to be cancelled & restarted.

Two participants experienced an issue with the sleep tracking,

> Not sure why but the sleeping tracker didn't work. I accepted and then later that day it said it was finished.

27 comments mentioned cognitive or sensory tasks. Six participants noted how much they enjoyed the cognitive and sensory tasks,

> All the activities were so fun!

However, 13 participants noted difficulty with the Tower of Hanoi (disk stacking) task, or the 9-hole peg task,

> I didnt understand what to do in the disc stacking task. Ive done similar things in the past without a problem. But this one had me stumped.

> Some tasks required getting the technique using a phone right first (e.g., the pinching and moving the dot).

48 comments addressed the surveys. Most noted was the repetitious daily surveys within the survey sprints, which 37 of those who commented found to be monotonous, irritating, or boring.

> Doing the surveys became really repetitive, as they were the same questions for days on end.

A small number of participants provided feedback on specific survey inclusions, such as

> …mood options were limited.

> I enjoyed it and was intrigued by the two survey questions that involved gender and did not seem to fit with the idea of perceptions of aging.

21 comments offered suggestions for future content inclusions. Suggestions included:

The ability to add contextual comments on days with a scheduled survey or task,

> The option to explain some answers may be useful. For example my first hearing test was effected by a fire alarm. Survey results effected yesterday by a reaction to COVID vaccination

The addition of definitions for core constructs of interest,

> Some definition of terms would have been helpful in the survey to ensure reporting on what the researcher wants

The ability to pause the schedule or reset the schedule to an earlier point if interrupted for a block of days,

> I was in a mobile/internet black spot for 10 days. I would have liked to rewind back to that block to allow me to fully participate

And the provision of personalised results following surveys or tasks,

> Results on each of the tests. Ie you do/do not have colour vision issues, hearing is better in left or right.

Two participants lamented the lack of face-to-face contact with the research team. For one participant, the…

> Lack of any face to face researcher/subject meeting meant lack of commitment to study.

38 participants shared positive feedback and/or notes of thanks.

Discussion

This study evaluated the feasibility, adherence, and usability of the Labs Without Walls research app and paired Apple Watches for 8-week remote micro-longitudinal research. We found strong evidence that this approach is feasible and effective. This supports the growing use of Apple's ResearchKit for cost-effective and accessible app-based research [e.g. (7, 22)]. Our ability to remotely recruit a diverse sample of adults across Australia, including older adults, demonstrates the potential of mobile technologies to reach hard-to-reach participants, regardless of age or location, and is in-line with global research showing that digital research participation is acceptable to people of a range of ages [e.g., (23)].

The 88.16% study completion rate exceeds the average reported in a recent review (11). Despite a large sample and intensive 8-week period, strategies like customizable task notifications likely boosted retention. Age and sex did not impact completion rates, contrary to some previous research that has shown higher completion rates among females compared to males (15).

Adherence ranged from 100% for the baseline survey to 70.61% for Tone Audiometry tests. In-line with advice from Broekhuis et al. (18), poorer adherence for these tests may be due to their longer duration and specific testing requirements. However, all tasks exceeded the 60% completion benchmark. While there was a slight decline over 8 weeks, it was less substantial than in other longer studies [e.g., (19)].

As anticipated, completeness of passively collected health and environmental data varied depending on the frequency of the behaviours being measured. Completeness was excellent for many measures across the full 8-week schedule, including greater than 80% completeness on each study day for wearing the Apple Watch (Apple Inc), active and basal kilijoules, resting heart rate, heart rate variability, and step count. Completeness was lower for measures that required dedicated periods of specific activities (e.g., walking speed, walking assymetry, walking heart rate, and stairs climbed). The poorest completeness was seen for stair ascent and stair descent speeds—the least regular of physical activity patterns studied. Future studies should consider the impact of behaviour regularity on missing data in ambulatory assessment studies. The optional sleep week had a 57.89% completion rate, with no age or sex differences. Future studies should consider lower completion rates for optional elements when planning sample sizes.

The usability ratings were extremely positive. 95.7% of participants found the intensive schedule manageable. This is similar to previous research with smaller, age-restricted samples and less intensive testing schedules [e.g., (24)]. The app was rated user-friendly and enjoyable, and 88% found the alert frequency "just right". Over 98% wore the Apple Watch, with most reporting no setup or charging difficulties. However, improving support for less tech-savvy participants is an opportunity.

A significant proportion of participants reported increased physical activity due to wearing the Apple Watch. While Labs Without Walls was designed as an observational study, providing the watch may have unintentionally influenced behavior. Longer studies may find that there is an initial increase in activity followed by a return to baseline for most people, however, research is needed to explore this hypothesis. Alternatively, future studies could seek to recruit participants who already own and use an Apple Watch into their research studies. Doing so would reduce or remove the novelty associated with the devices which we believe was the reason for the increased activity in the current study, while also lowering the cost of conducting similar research by removing the need for device purchase and/or postage.

We recommend open-ended feedback as a core element of usability testing in future studies. In this study, participant feedback highlighted opportunities for improvement. Among the most prominent constructive feedback was that participants found the repeated daily surveys to be monotonous. In hindsight, this is understandable given that the questions each day were the same for 7 days at a time, with only a brief justification provided to participants for why the questions were being asked. This daily-diary style approach was important to be able to answer our research questions, however, future research may be able to disrupt the perceived monotony by reducing the overall length of daily surveys and providing greater transparency to participants about the purpose of the sprints which may increase their perceived value and thereby decrease boredom. We note that future research would likely benefit from consultation with community members regarding survey and task scheduling and approaches to improving interest in repetitive aspects of research apps prior to launching full scale data collection.

Qualitative feedback also suggested that participants' experiences of the app as well as data quality could be impacted by device screen size and characteristics of certain active tasks. For example, several participants who joined the study with smaller iPhone devices (e.g., iPhones 6 and 7) found some of the game-like tasks difficult to complete on the smaller screens. This appears to be particularly true for tasks that involved using the touch screen function in a precise way (such as the Tower of Hanoi or the Hole Peg task, both pre-built within Apple's ResearchKit). Future remote research that aims to include such active tasks may benefit from recruiting participants who own iPhones with larger screens.

While the open-source resources provided by Apple's ResearchKit lowered the cost to conduct the study (compared to developing an app from the ground up), there were still considerable additional costs including additional iOS development, return-paid postage of study devices, and cloud storage of study data. Now that the Labs Without Walls app is built and validated as an acceptable research tool, it offers a scalable and relatively cost-effective means of conducting high volume remote research.

Finally, we also acknowledge the inherent selection bias in this study which results from offering a research app that is not compatible with Android devices. There were several important considerations that guided the decision to develop the app for iOS and not Android or both. First, iPhones are the most common smartphone device in the Australian market (25). Second, developing apps is expensive, and associated costs can more than double once you consider developing for both iOS and Android. In this case, developing for iOS was the most cost effective option given the availability of open-source ResearchKit tasks (26) which substantially reduced development time—particularly for the game-like cognitive and sensory tasks we administered from the ResearchKit library—and a grant of Apple Watch devices (Apple Inc) which can be more seamlessly integrated with an iOS app. Additionally, there is evidence for response latency differences across operating systems and devices, which is particularly evident among Android devices given the much greater variability in device manufacturers (27). This calls into question the current comparability of performance across devices on some tasks. There is no doubt that technological innovations in the coming years will remediate some of the concerns above and make it more feasible to build and administer research apps that are equivalent across both iOS and Android devices. We believe this will be an important future step in improving the accessibility of app- and wearable-based research.

Conclusion

This study provides strong evidence for the feasibility, adherence, and usability of the Labs Without Walls research app and paired Apple Watch devices. By addressing challenges and incorporating participant feedback, future research can further enhance the accessibility and impact of app-based studies in the field of aging research.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The study involving humans was approved by UNSW Human Research Ethics Committee. The study was conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

BB: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Writing – original draft, Writing – review & editing. SZ: Data curation, Investigation, Project administration, Writing – review & editing. DA: Methodology, Software, Writing – review & editing. LZ: Project administration, Writing – review & editing. RE: Funding acquisition, Methodology, Project administration, Writing – review & editing. KA: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing, Investigation.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This project is funded by the University of New South Wales Ageing Futures Institute. BB, LZ, DA and KJA are supported by Australian Research Council Laureate Fellowship FL190100011. SZ is supported by an Australian Government Research Training Program (RTP) scholarship and Australian Research Council Centre of Excellence in Population Ageing Research (Project number: CE170100005).

Acknowledgments

The research team would like to thank the research volunteers who joined the Labs Without Walls study from around Australia. This project received assistance with the design and set-up of backend data storage architecture from UNSW Sydney Research Technology Services (ResTech). ResTech have also partially supported the cost of hosting the backend storage of study data via their cloud pilot scheme. Apple Watch (Apple Inc) devices were provided by Apple Inc. following a successful application to their Investigator Support Program (https://www.researchandcare.org/resources/investigator-support-program/). Apple Inc is not involved in the conduct of the research.

Conflict of interest

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

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2025.1520971/full#supplementary-material

References

1. Brady B, Zhou S, Ashworth D, Zheng L, Eramudugolla R, Huque MH, et al. A technology-enriched approach to studying microlongitudinal aging among adults aged 18 to 85 years: protocol for the labs without walls study. *JMIR Res Protoc.* (2023) 12(1):e47053. doi: 10.2196/47053

2. Gerstorf D, Hoppmann CA, Ram N. The promise and challenges of integrating multiple time-scales in adult developmental inquiry. *Res Hum Dev.* (2014) 11(2):75–90. doi: 10.1080/15427609.2014.906727

3. Rickenbach M, Almeida DM. Micro-longitudinal research in life-span psychology. *Res Hum Dev.* (2019) 16(2–3):94–106. doi: 10.1080/15427609.2019.1600057

4. Sliwinski MJ. Measurement-burst designs for social health research. *Soc Personal Psychol Compass.* (2008) 2(1):245–61. doi: 10.1111/j.1751-9004.2007.00043.x

5. Baltes PB. Theoretical propositions of life-span developmental psychology: on the dynamics between growth and decline. *Dev Psychol.* (1987) 23(5):611. doi: 10.1037/0012-1649.23.5.611

6. Baltes PB, Nesselroade JR. Paradigm lost and paradigm regained: critique of dannefer's portrayal of life-span developmental psychology. *Am Sociol Rev.* (1984) 49(6):841–7. doi: 10.2307/2095533

7. Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci Data.* (2016) 3(1):1–9. doi: 10.1038/sdata.2016.11

8. Munos B, Baker PC, Bot BM, Crouthamel M, de Vries G, Ferguson I, et al. Mobile health: the power of wearables, sensors, and apps to transform clinical trials. *Ann N Y Acad Sci.* (2016) 1375(1):3–18. doi: 10.1111/nyas.13117

9. Fischer F, Kleen S. Possibilities, problems, and perspectives of data collection by mobile apps in longitudinal epidemiological studies: scoping review. *J Med Internet Res.* (2021) 23(1):e17691. doi: 10.2196/17691

10. World Health Organization. Monitoring and Evaluating Digital Health Interventions: A Practical Guide to Conducting Research and Assessment. (2016). Available online at: https://iris.who.int/bitstream/handle/10665/252183 (Accessed June 01, 2024).

11. Oakley-Girvan I, Yunis R, Longmire M, Ouillon JS. What works best to engage participants in mobile app interventions and e-health: a scoping review. *Telemed E Health.* (2022) 28(6):768–80. doi: 10.1089/tmj.2021.0226

12. Pratap A, Renn BN, Volponi J, Mooney SD, Gazzaley A, Arean PA, et al. Using mobile apps to assess and treat depression in Hispanic and Latino populations: fully remote randomized clinical trial. *J Med Internet Res.* (2018) 20(8):e10130. doi: 10.2196/10130

13. Morrison LG, Hargood C, Lin SX, Dennison L, Joseph J, Hughes S, et al. Understanding usage of a hybrid website and smartphone app for weight management: a mixed-methods study. *J Med Internet Res.* (2014) 16(10):e3579. doi: 10.2196/jmir.3579

14. Sims T, Reed AE, Carr DC. Information and communication technology use is related to higher well-being among the oldest-old. *J Gerontol Series B*. (2016) 72:761–70. doi: 10.1093/geronb/gbw130

15. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. *Assessment*. (2023) 30(3):825–46. doi: 10.1177/10731911211002995

16. Schmitt DP, Realo A, Voracek M, Allik J. Why can't a man be more like a woman? Sex differences in big five personality traits across 55 cultures. *J Pers Soc Psychol*. (2008) 94:168–82. doi: 10.1037/0022-3514.94.1.168

17. Westerhof GJ, Miche M, Brothers AF, Barrett AE, Diehl M, Montepare JM, et al. The influence of subjective aging on health and longevity: a meta-analysis of longitudinal data. *Psychol Aging*. (2014) 29(4):793. doi: 10.1037/a0038016

18. Broekhuis M, van Velsen L, Hermens H. Assessing usability of eHealth technology: a comparison of usability benchmarking instruments. *Int J Med Inf*. (2019) 128:24–31. doi: 10.1016/j.ijmedinf.2019.05.016

19. Pathiravasan CH, Zhang Y, Trinquart L, Benjamin EJ, Borrelli B, McManus DD, et al. Adherence of mobile app-based surveys and comparison with traditional surveys: eCohort study. *J Med Internet Res*. (2021) 23(1):e24773. doi: 10.2196/24773

20. IBM Corporation. *IBM SPSS Statistics for Windows, Version 27.0*. New York, NY: IBM Corp (2020).

21. Australian Bureau of Statistics. 2021 Census Community Profiles: Australia. (2021). Available online at: https://www.abs.gov.au/census/find-census-data/community-profiles/2021 (Accessed June 01, 2024).

22. Chan YFY, Bot BM, Zweig M, Tignor N, Ma W, Suver C, et al. The asthma mobile health study, smartphone data collected using ResearchKit. *Sci Data*. (2018) 5(1):1–11. doi: 10.1038/sdata.2018.96

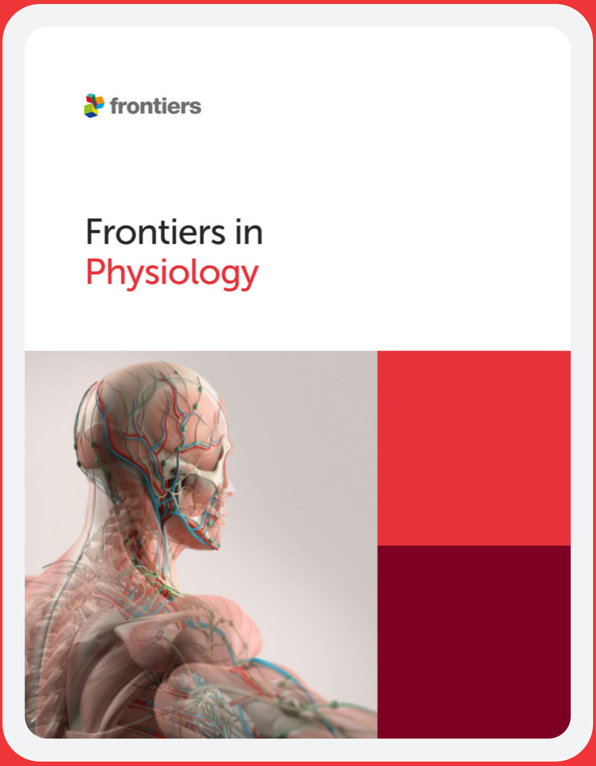
23. Dumbari NM, Gever VC. Online vs face-to-face research participation: which do research respondents prefer? *Mdooter J Commun Digit Technol*. (2025) 2(1):1–7. https://www.mdooterj.com/index.php/mdooterj/article/view/6

24. Brewster PW, Rush J, Ozen L, Vendittelli R, Hofer SM. Feasibility and psychometric integrity of mobile phone-based intensive measurement of cognition in older adults. *Exp Aging Res*. (2021) 47(4):303–21. doi: 10.1080/0361073X.2021.1883160

25. Statista. Most popular smartphone brands in Australia as of September 2023. (2023). Available online at: https://www.statista.com/forecasts/1370998/most-popular-smartphone-brands-in-australia (Accessed June 01, 2024).

26. Apple Inc. Apple ResearchKit framework. (2023). Available online at: https://www.researchandcare.org/researchkit/http://researchkit.org/

27. Nicosia J, Wang B, Aschenbrenner AJ, Sliwinski MJ, Yabiku ST, Roque NA, et al. To BYOD or not: are device latencies important for bring-your-own-device (BYOD) smartphone cognitive testing? *Behav Res Methods*. (2023) 55(6):2800–12. doi: 10.3758/s13428-021-01626-1


frontiers
Frontiers in
Physiology