Contents

Original Papers

Enabling Wearable Pulse Transit Time-Based Blood Pressure Estimation for Medically Underserved Areas and Health Equity: Comprehensive Evaluation Study (e27466)
Venu Ganti, Andrew Carek, Hewon Jung, Adith Srivatsa, Deborah Cherry, Levather Johnson, Omer Inan. .......................................................... 2

Comparison of the Validity and Generalizability of Machine Learning Algorithms for the Prediction of Energy Expenditure: Validation Study (e23938)
Ruairi O’Driscoll, Jake Turicchi, Mark Hopkins, Cristina Duarte, Graham Horgan, Graham Finlayson, R Stubbs. ................................. 20
Enabling Wearable Pulse Transit Time-Based Blood Pressure Estimation for Medically Underserved Areas and Health Equity: Comprehensive Evaluation Study

Venu Ganti¹, BS; Andrew M Carek¹, PhD; Hewon Jung¹, MS; Adith V Srivatsa²; Deborah Cherry³, BBA; Levather Neicey Johnson³, BSN, RN; Omer T Inan¹,², PhD

¹School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, United States
²The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, United States
³VSNS, Inc, Atlanta, GA, United States

Corresponding Author:
Venu Ganti, BS
School of Electrical and Computer Engineering
Georgia Institute of Technology
85 5th St NW
Atlanta, GA, 30308
United States
Phone: 1 2406434250
Email: vganti6@gatech.edu

Abstract

Background: Noninvasive and cuffless approaches to monitor blood pressure (BP), in light of their convenience and accuracy, have paved the way toward remote screening and management of hypertension. However, existing noninvasive methodologies, which operate on mechanical, electrical, and optical sensing modalities, have not been thoroughly evaluated in demographically and racially diverse populations. Thus, the potential accuracy of these technologies in populations where they could have the greatest impact has not been sufficiently addressed. This presents challenges in clinical translation due to concerns about perpetuating existing health disparities.

Objective: In this paper, we aim to present findings on the feasibility of a cuffless, wrist-worn, pulse transit time (PTT)–based device for monitoring BP in a diverse population.

Methods: We recruited a diverse population through a collaborative effort with a nonprofit organization working with medically underserved areas in Georgia. We used our custom, multimodal, wrist-worn device to measure the PTT through seismocardiography, as the proximal timing reference, and photoplethysmography, as the distal timing reference. In addition, we created a novel data-driven beat-selection algorithm to reduce noise and improve the robustness of the method. We compared the wearable PTT measurements with those from a finger-cuff continuous BP device over the course of several perturbations used to modulate BP.

Results: Our PTT-based wrist-worn device accurately monitored diastolic blood pressure (DBP) and mean arterial pressure (MAP) in a diverse population (N=44 participants) with a mean absolute difference of 2.90 mm Hg and 3.39 mm Hg for DBP and MAP, respectively, after calibration. Meanwhile, the mean absolute difference of our systolic BP estimation was 5.36 mm Hg, a grade B classification based on the Institute for Electronics and Electrical Engineers standard. We have further demonstrated the ability of our device to capture the commonly observed demographic differences in underlying arterial stiffness.

Conclusions: Accurate DBP and MAP estimation, along with grade B systolic BP estimation, using a convenient wearable device can empower users and facilitate remote BP monitoring in medically underserved areas, thus providing widespread hypertension screening and management for health equity.

(JMIR Mhealth Uhealth 2021;9(8):e27466) doi:10.2196/27466

KEYWORDS
wearable sensing; pulse transit time; cuffless blood pressure; noninvasive blood pressure estimation; health equity; mobile phone
Introduction

Background

Current clinical practice regarding hypertension management and control hinges on the century-old approach of obtaining infrequent cuff-based measurements of blood pressure (BP) in clinical settings. This paradigm of the measurement being anchored to the clinical setting and requiring persons to proactively visit a medical professional to determine their hypertensive status is costly—due to the time and money spent [1]—and considered ineffective—due to the infrequency and error (ie, white coat hypertension) of office BP measurements [2,3]. Hence, we observed remarkable disparities in hypertension detection, treatment, and control across socioeconomic status and race, with populations lacking access to regular office visits and care, having nearly half the awareness of their existing hypertensive status, and enduring up to triple the rates of subsequent cardiac events [4,5]. Technologies that enable frequent, reliable, and accurate measurements of BP in ambulatory settings promise to reduce the global burden of hypertension and offer an opportunity to advance health equity [6]. Leveraging the ubiquity of smartphones and digital health technologies equipped with highly sensitive, miniaturized sensors is essential for the remote monitoring of BP [7].

Existing wearable devices that incorporate noninvasive BP methodologies offer an affordable and efficient means of tracking out-of-office BP [8]. Unfortunately, they commonly use uncomfortable techniques (ie, oscillometry and tonometry) that demand imparting forces on blood vessels to achieve accurate measurements [9-11]. These inconveniences fail to empower users to take control of their health, posing a significant challenge to the widespread routine monitoring of BP. Instead, strategies that compute the pulse transit time (PTT), a measure of arterial stiffness, present a convenient alternative for BP estimation [12].

The PTT, the time the pressure wave propagates along the length of the arterial tree, is a cuffless surrogate for BP and can be acquired noninvasively [12]. In practice, the acquisition of noninvasive PTT requires a combination of sensors—typically an accelerometer, force sensor, light-emitting diode (LED) and photodiode, electrode, or ultrasonic transceiver—placed proximally and distally along the arterial tree and computed from fiducial points in the captured seismocardiogram (SCG), ballistocardiogram, photoplethysmogram (PPG), impedance cardiogram, impedance plethysmogram, or arterial blood pressure (ABP) waveform [13,14]. Despite their inherent convenience, these sensing modalities are naturally of concern when used in populations with intrinsic mechanical, optical, and electrical barriers, stemming from higher melanin levels and body fat percentages.

To the best of our knowledge, noninvasive PTT-based BP estimation has yet to be examined in a diverse population—a gap in our scientific understanding that presents a formidable obstacle to its adoption. Specifically, some medically underserved areas (MUAs), which stand to benefit the most from remote monitoring [15], have a large number of Black and Latino individuals with higher melanin content and obesity rates compared with White individuals [16]. Recent notable data from the Centers for Disease Control and Prevention further stress these concerns by exposing that non-Hispanic Black individuals not only have significantly higher hypertension prevalence than non-Hispanic White individuals but also witness significantly lower hypertension control rates [17,18]. Affordable remote monitoring options have the responsibility to combat not only social determinants of health, such as access to health care and income, but also in turn the existing health disparities that are byproducts of them. As a result, there exists a glaring hole in PTT-based BP monitoring—that this technology has yet to be tested on the population for whom it may be the most valuable, and until now, its continuing practice will only exacerbate existing health disparities.

Objectives

In our previous work, we designed a wearable, multimodal, wrist-worn PTT monitoring device (SeismoWatch) and validated it in both controlled lab [19] and unsupervised home [20] settings, primarily on young, healthy persons with lighter skin. In this paper, we expand upon our previous work with a community-engaged research strategy that leverages expertise from a nonprofit organization serving MUAs in Georgia and evaluated our device in a more diverse population. We present our device’s ability to accurately estimate BP in this diverse population and capture significant demographic-level differences in underlying arterial stiffness that coincide with observations from existing literature, through the calibration coefficients used in our BP estimation model. This work represents the first time that a noninvasive, cuffless, PTT-based wearable device has been extensively evaluated in a community-based diverse population as a potentially reliable and convenient monitoring option toward, ultimately, the remote screening and management of hypertension for health equity.

Methods

Study Protocol

A comprehensive breakdown of the demographics of the study population is presented in Table 1. This study was conducted under a protocol approved by the Georgia Institute of Technology institutional review board (protocol number H19251). The study was separated into two different populations (N=44 participants) referred to throughout this work as follows: (1) a young and healthy homogeneous population (first cohort=26 participants) and (2) an older, entirely Black, higher BMI, metropolitan population (second cohort=18 participants) recruited later through the help of our community outreach partners—a nonprofit organization serving medically underserved persons in the state of Georgia. For the first cohort, 26 (19 males and 7 females) young and healthy volunteers (mean age 26.7 years, SD 3.7; mean weight 73.8 kg, SD 14.1; height 173.9 cm, SD 9.6; and mean BMI 24.2 kg/m², SD 3.2) with no previous history of cardiovascular disease were recruited, and written informed consent was obtained. For the second cohort, 18 (6 males and 12 females) Black participants (mean age 44.1 years, SD 11.7; mean weight 94.4 kg, SD 18.0; mean height 169.6 cm, SD 11.5; and mean BMI 33.2 kg/m², SD 7.6) with no previous history of cardiovascular disease other than
hypertension were recruited from the Atlanta metropolitan area; written informed consent was obtained, and further demographic information was collected post hoc with verbal consent. Both hypertensive status and the use of regular prescription medications were self-reported.

Table 1. Participant demographics and cardiovascular parameters for study participants (grouped by cohort; N=44).

<table>
<thead>
<tr>
<th>Demographics and cardiovascular parameters</th>
<th>Homogenous data set (first cohort; n=26; participant 1-26)</th>
<th>Community outreach (metropolitan Atlanta) data set (second cohort; n=18; participant 27-44)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>26.7 (3.7)</td>
<td>44.1 (11.7)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>19 (73)</td>
<td>6 (33)</td>
<td>N/A</td>
</tr>
<tr>
<td>Female</td>
<td>7 (26)</td>
<td>12 (67)</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Height (cm), mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homogenous data set</td>
<td>173.9 (9.6)</td>
<td>169.6 (11.5)</td>
<td>.19</td>
</tr>
<tr>
<td>Community outreach</td>
<td>73.8 (14.1)</td>
<td>94.4 (18.0)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Weight (kg), mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homogenous data set</td>
<td>33.2 (7.6)</td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Community outreach</td>
<td>24.2 (3.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BMIc (kg/m^2), mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homogenous data set</td>
<td>94.4 (18.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Community outreach</td>
<td>73.8 (14.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Obesity class, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>1^d (4)</td>
<td>2^e (11)</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>N/A</td>
<td>3^f (17)</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>N/A</td>
<td>4^g (22)</td>
<td></td>
</tr>
<tr>
<td><strong>Race, n (%)</strong></td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Black</td>
<td>1^h (4)</td>
<td>18 (100)</td>
<td></td>
</tr>
<tr>
<td>Other race</td>
<td>25 (96)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td><strong>Hypertensive status, n (%)</strong></td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Normotensive</td>
<td>26 (100)</td>
<td>15 (83)</td>
<td></td>
</tr>
<tr>
<td>Hypertensive</td>
<td>N/A</td>
<td>2^i (11)</td>
<td></td>
</tr>
<tr>
<td>Hypotensive</td>
<td>N/A</td>
<td>1^j (6)</td>
<td></td>
</tr>
<tr>
<td><strong>Current medications, n (%)</strong></td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Hydrochlorothiazide (1×day)</td>
<td>N/A</td>
<td>2^k (11)</td>
<td></td>
</tr>
<tr>
<td>Lisinopril (1×day)</td>
<td>N/A</td>
<td>1^l (6)</td>
<td></td>
</tr>
<tr>
<td>Iron supplement</td>
<td>N/A</td>
<td>1^m (6)</td>
<td></td>
</tr>
</tbody>
</table>

^aStatistical significance between groups in values, where applicable, was computed using an unpaired two-tailed t test.

^bN/A: not applicable.

^cObesity classified using the BMI per the guidelines from the National Heart, Lung, and Blood Institute of the National Institutes of Health [21] (I: BMI=30-34.9; II: BMI=35-39.9; III: BMI ≥40).

^dParticipant 23.

^eParticipants 30 and 43.

^fParticipants 38, 40, and 42.

^gParticipants 34, 36, 37, and 41.

^hParticipant 5.

^iParticipants 29 and 37.

^jParticipant 33.

^kParticipants 29 and 37.

^lParticipant 29.

^mParticipant 33.

The concept of the study design is shown in Figure 1. Although not explicitly shown, two versions of the SeismoWatch were used in this study: a previous version of the hardware with comparable sensors was used in the young, homogeneous population (ie, the first cohort), before being adapted for a more robust, portable, and multimodal wearable device used in the metropolitan Atlanta population (ie, the second cohort). Specifically, the data from these cohorts were collected during
two intervals, between which the hardware was revised to incorporate multiwavelength PPGs before investigating the performance of the sensing modality in the underrepresented population. This was essential to assess the efficacy of shorter-wavelength LEDs (ie, those with shallower skin penetration depths) in a Black population. However, in both the correlations in Figure 2 and calibration coefficient comparisons in Figure 3, only the results derived from the infrared (IR) PPGs, available to both devices, were computed and shown. The other key sensing components and reference system components were essentially identical: (1) the first version of the device used an analog version of the accelerometer (ADXL354, Analog Devices) to acquire the SCG, whereas the second version simply used the digital version of the same sensor (ADXL355, Analog Devices) to reduce size and (2) the finger-cuff continuous BP reference system (ccNexfin, Edwards Lifesciences) along with the data acquisition module (MPU150, Biopac Systems) were identical in both studies.

Figure 1. Concept overview and study design. Sensor information and placement locations for wearable system (blue) and reference system (purple). Noninvasive pulse transit time (PTT) measurement concept overview using seismocardiogram (SCG) and photoplethysmogram (PPG) sensors. Study protocol tasks in chronological order with duration and mean (SD) of mean arterial pressure (MAP) values for each task. Sample filtered signals from the participant with the lowest signal-to-noise ratio (SNR) signals (n=37); a hypertensive, high BMI, older Black female. In order from top to bottom: electrocardiogram (ECG), SCG, infrared PPG, red PPG, green PPG signals measured from the wearable system (blue) and the synchronized ECG, and arterial blood pressure (ABP) signals measured by the reference system (purple). Systolic blood pressure (SBP; top) and diastolic blood pressure (DBP; bottom) plotted across the full protocol for participant 37, with rest periods (green) and perturbations used to modulate BP (red) highlighted in chronological order, and the location where the reference finger-cuff continuous blood pressure (BP) system was paused during the exercise indicated. ABP: arterial blood pressure; BP: blood pressure; DBP: diastolic blood pressure; ECG: electrocardiogram; LED: light-emitting diode; PD: photodiode; PPG: photoplethysmogram; PTT: pulse transit time; SBP: systolic blood pressure; SCG: seismocardiogram.
Figure 2. Wearable pulse transit time (PTT)–based blood pressure (BP) estimation results. Correlation and Bland-Altman plots between PTT-estimated BP and the finger-cuff continuous BP for mean arterial pressure, diastolic blood pressure, and systolic blood pressure estimation. The root mean squared error and the mean absolute difference for each correlation are shown. DBP: diastolic blood pressure; MAD: mean absolute difference; MAP: mean arterial pressure; PTT: pulse transit time; RMSE: root mean square error; SBP: systolic blood pressure.

Figure 3. Participant-specific diastolic blood pressure (DBP) calibration coefficients are significantly different in demographics with typical disparities in arterial stiffness. Boxplots showing the statistically significant (*P<.05; Mann-Whitney U) difference in the DBP K1 and K2 calibration coefficients between participants who are nonobese and obese. Boxplots showing the statistically significant (*P<.05; Mann-Whitney U) difference in the DBP K1 calibration coefficients between male and female participants. Boxplots showing the statistically significant (*P<.05; Mann-Whitney U) difference in the DBP K1 calibration coefficients between participants of other race and Black participants. Boxplots showing the difference in the DBP K1 and K2 calibration coefficients between young and older participants. DBP: diastolic blood pressure.

To acquire a timing reference for the start of a cardiac cycle, while serving as a reference for alignment to the wearable system signals, a wireless electrocardiogram (ECG) module (BN-EL50, Biopac Systems) was attached to the participant in a three-lead configuration with Ag/AgCl gel electrodes as shown in Figure 1. As depicted in Figure 1, a finger-cuff BP sensor
based on the volume-clamp methodology (ccNexfin, Edwards Lifesciences) [22,23] was placed on the same hand as the watch, acquiring a reference measurement of continuous beat-by-beat BP. Although volume-clamping continuous BP devices are not the clinical gold standard for ABP measurements, an arterial line was not feasible due to invasiveness, and a sphygmomanometer was not used because of the need for a trained professional and lack of beat-by-beat BP data. Similarly, semiautomated BP cuffs were not used as they hinge on following strict guidelines to obtain an accurate reading, such as being seated and resting the arm at heart level, which were impossible to satisfy simultaneously while acquiring watch measurements, given the need for the contralateral hand to touch the ECG electrode to activate the PPT mode [20]. In addition, it was recently demonstrated that a volume-clamping–based system had comparable accuracy with noninvasive oscillometric BP cuffs [24]. All reference system sensors were sampled at 1 kHz and interfaced to a computer using a data acquisition system (MPU150, Biopac Systems) and its corresponding software (Acqknowledge, Biopac Systems). The reference system files were saved to a desktop computer for postprocessing.

Participants were asked to change into either a V-cut T-shirt or tank top, if not wearing one already, to acquire the sternal PPGs included in the wearable designs, though not examined in this work. The watch was fitted such that the PPGs faced the radial artery on the ventral side of the wrist. To capture the PTT, the participant performed a simple maneuver to place the watch on the sternum to acquire the SCG for the proximal timing reference, as shown in Figure 1, whereas the PPGs were sampled at both the sternum and wrist. Although this offers a noncontinuous assessment, routine remote BP monitoring using oscillometric devices has already demonstrated clinical value, although it does not offer continuous BP measurement [2]. Specifically, ambulatory BP monitors, due to their superior portability and measurement frequency—comparable with what this wearable device can easily provide [20]—have become invaluable for the screening and management of hypertension [2] such that the added benefit of continuous BP measurement may only be marginal.

In order, the participants went through a 2-minute baseline period while sitting before obtaining another 2-minute baseline measurement while standing. Then, a series of perturbations with varying rest periods in between were used to modulate BP. First, a mental arithmetic exercise was used to increase BP [12], in which participants were given a three-digit integer and were told to add the sum of the digits to the number repeatedly for 1 minute. Then, a cold pressor test was conducted in which participants submerged their hand contralateral to the watch in a bucket of ice water for as long as tolerable or until the full minute. Finally, during the exercise session, the finger cuff was removed to avoid damage, and the participant performed either a stair stepping or bicycling exercise, based on personal preference, for 1 minute. As mentioned in our previous work [20], the new version of the watch enters the PTT measurement mode when the user places a finger from their hand contralateral to the watch on the positive wrist ECG electrode; therefore, we were unable to acquire PTT data during exercise for both cohorts and cold pressor for the participants in the second cohort (ie, second cohort). Although with the newer hardware, we were unable to collect PTT data during the cold pressor perturbation for the second cohort, the effect of the cold pressor—assessed directly after the hand was removed from the ice water (ie, a maximum of 1 minute after immersion)—was still well within its physiological window during the following rest period [25]. Overall, as our device is not designed to offer continuous measurements of BP, examining the effect of these perturbations in the rest period directly following them, similar to our previous work [19], still allowed for a comprehensive evaluation of the methodology in a diverse population. However, PTT data from the first cohort during the cold pressor were still used. As the BP data from the cold pressor test were still acquired for the second cohort, as the continuous BP cuff was still on, the mean arterial pressure (MAP) values were factored into the ones displayed in Figure 1. To do so, a 50 ms moving average filter was applied to the measured continuous BP signal, ensemble averages of 10 heartbeats with 50% overlap were taken, and the BP beat with the highest signal-to-noise ratio (SNR) was selected.

**Signal Processing**

The signal processing pipeline is shown in Figure 4. All signal processing and statistical analyses were performed in MATLAB R2018a (MathWorks). Before preprocessing the SCG and PPG signals acquired from the wearable system, it was imperative to time-align them to the continuous BP signal from the reference system using the respective ECGs to ensure proper temporal comparison. Specifically, the ECGs from each system were first filtered using a digital finite impulse response bandpass filter (BPF; fpass=10-40 Hz) to remove baseline wander due to postural sway and extract the R-wave, which was then identified using a simple peak detection algorithm. Then, cross-correlation was used to align the R-peaks of the two ECG readings by detecting the amount of lead and truncating either the wearable or reference signals depending on the condition. After alignment, the dorsoventral axis of the SCG (ie, z-axis acceleration) and green, red, and IR wrist PPGs were filtered using a digital finite impulse response BPF with bandwidths of 1-40 Hz and 1-8 Hz, respectively, to remove their out-of-band noise and baseline wander due to respiration. In addition, the continuous ABP waveform was smoothed using a 50 ms moving average filter.
Figure 4. Signal processing pipeline. Block diagram of signal processing overview showing signal alignment using electrocardiogram signals acquired from the wearable system (blue) and reference system (purple) before bandpass filtering, heartbeat windowing, and photoplethysmogram (PPG) selection. After beat selection and signal quality assessment, the pulse transit time is computed as the aortic valve opening point of the seismocardiogram to the diastolic foot of the PPG. Calibration is used to estimate blood pressure (BP) using the arterial blood pressure waveform acquired from the continuous BP finger-cuff. Block diagram of the custom PPG selection algorithm, locating beats with greater systolic upstrokes and signal-to-noise ratio (SNR). ABP: arterial blood pressure; AO: aortic valve opening; BP: blood pressure; BPF: bandpass filter; ECG: electrocardiogram; PPG: photoplethysmogram; PTT: pulse transit time; SBP: systolic blood pressure; SCG: seismocardiogram; SNR: signal-to-noise ratio.

Next, the filtered and aligned SCG, PPG, and ABP waveforms were split into separate heartbeats using the detected R-R intervals of the synchronized ECG. Then, these heartbeat-indexed signals were ensemble-averaged using 10-beat windows with 50% overlap before assessing the signal quality to select the highest quality beat per task for each participant, similar to the methods used in our previous works [19,20]. Given the number of high BMI participants in this population, the SCG not only had a lower mean SNR when compared with our previous studies but was also observed to have less variability than the PPG SNR; hence, an emphasis was placed on determining the optimal PPGs first. In addition, upon an initial assessment of signal quality, it was observed that when the PPG signal had the highest SNR, typically, the SCG signal did as well—perhaps because acquiring a clean PPG signal inherently hinges on applying consistent pressure. The optimum PPG was selected using a physiologically inspired algorithm to first identify the beats with the top 10% of systolic upstrokes (ie, maximum of the derivative of the PPG waveform) and then select the remaining beat with the maximum SNR. The SNR
was calculated using a noise-to-signal ratio detection algorithm detailed in Iinan et al [26]. The methods used to determine the timing references for PTT calculation, the foot of the PPG, and the aortic valve opening (AO) point of the SCG were the same as those used in our previous studies [19,20]. Specifically, the foot of the PPG was computed using the intersecting tangent method described in the study by Mukkamala et al [12], and the AO point was assumed to be the first peak in each ensemble-averaged window before the foot of the PPG. Occasionally, the SCG signal was manually annotated to impose realistic constraints for the AO point or to ensure that the same morphological peak was consistently chosen for all tasks per participant. Participant-specific SNR thresholds were set to retain only high-fidelity signals; if the SNR of the SCG, PPG, or ABP beats was not greater than the prescribed cutoff, or if the foot of the PPG was not within a realistic range, then the respective ensemble-averaged waveforms were deemed too noisy for use and that task was not used for PTT calculation. Notably, the continuous reference BP allowed for the ability to evaluate the SNR of the ABP signal and incorporate this quality assessment into our signal processing pipeline to remove beats with low SNR reference measurements. After the entire signal quality assessment process, at least four of the tasks were used for BP estimation per participant. Finally, the PTT was calculated as the difference of the proximal timing reference, AO point of the paired SCG, and distal timing reference, the foot of the selected PPG. In addition to wavelength comparisons, the green and red wavelength PPGs were not used as the IR wavelength wrist PPGs had the highest mean SNR, because of the greater indifference of the IR wavelength to melanin absorption and the ability to capture more pulsatile arteries deeper in the tissue than cutaneous capillaries [12,27].

In addition, the postexercise recovery period was separated into an early and late rest period based on when the BP reached a consistent value. This allowed us to capture both the immediately heightened cardiac output–induced BP increase postexercise and the recovery back to baseline, while opportunely adding another PTT and BP data point for linear regression.

Statistical Analysis

Simple linear regression was performed independently between wearable participant-specific inverse PTT (PTT\(^{-1}\)) and reference diastolic blood pressure (DBP), MAP, or systolic blood pressure (SBP) value pairs, to calculate the calibration coefficients necessary to estimate each of the three BP components per participant; nonlinear models, whereas potentially more accurate, dictate the need for more calibration points [12,28]. Therefore, the resulting calibration coefficients—used to estimate BP from the conventional PTT-based BP estimation model shown in equation 1—are merely the slope (ie, slope calibration coefficient \([K_1]\)) and y-intercept (ie, Y-intercept calibration coefficient \([K_2]\)) of the line of best fit [12]. This was identical to the calibration methods used in our previous work [19,20].

\[
BP = (K_1 / PTT) + K_2 \quad (1)
\]

The mean absolute difference (MAD) was computed from the mean of the absolute value of the difference between the estimated and reference BP. The benchmarks for MAD were chosen based on the Institute for Electronics and Electrical Engineers (IEEE) standard for wearable cuffless BP measuring devices [29]. In addition, the root mean square error (RMSE), calculated as the root mean square of the difference between the estimated BP and measured BP, was computed because of its enhanced sensitivity to outliers.

We stratified the entire study population for the demographic comparisons of the calibration coefficients shown in Figure 3, based on four factors (ie, obesity, sex, race, and age) known to affect arterial stiffness [30-35] and therefore the PTT. The participants were split into nonobese and obese groups based on the guidelines from the National Heart, Lung, and Blood Institute of the National Institutes of Health defining a BMI \(\geq 30\) kg/m\(^2\) as obese [21]. Thus, the nonobese group had a BMI \(\leq 30\). To assess differences due to age, we separated the participants into younger (aged \(\leq 40\) years) and older groups (aged \(> 40\) years). Statistical significance \((P<.05)\) between demographic data for each cohort was assessed using an unpaired two-sample, two-tailed \(t\) test, as shown in Table 1.

For the demographic DBP calibration coefficient comparisons, a one-sample Kolmogorov-Smirnov test was used on each data point to test for normality, which determined that none of the data for the comparisons were normally distributed. Then, a Mann-Whitney U test (ie, Wilcoxon Rank Sum test) was used to assess statistical significance \((P<.05)\) among the unpaired data.

For the PPG wavelength DBP estimation comparisons—only applicable to the second cohort population due to the differences in hardware used—first, a one-sample Kolmogorov-Smirnov test was used on each data point to test for normality, which determined that none of the data for the comparisons were normally distributed. Then, a Wilcoxon Signed Rank test was used to assess statistical significance \((P<.05)\) among the paired data.

Results

Multimodal Engineering Mechanics of the SeismoWatch

The previous version of the watch, not shown in this work, consisted of a 3D printed case embedded with an accelerometer, PDs, and IR LEDs. All sensors were connected to a small external circuit box with straps for the participant to wear around the waist. The output of the analog accelerometer (ADXL354, Analog Devices) was connected to an analog front end (AFE) in the circuit box. To amplify the SCG signal and prevent saturation of the alternating current components owing to the varying direct current levels, the AFE separated the direct current and alternating current components using a low pass \((fc=1 \text{ Hz}; \text{G}=10 \text{ dB})\) and BPF \((fpass=0.2 \text{ Hz}-40 \text{ Hz})\) in parallel. An analog adder recombined both components. For PPG measurements, the cathode of the PDs (S2386-18k, Hamamatsu Photonics) was connected to a transimpedance amplifiers configured as a low-pass filter \((fc=12 \text{ Hz}; \text{G}=110 \text{ dB})\) followed by...
by gain and filter stages ($f_{\text{pass}}=0.5-12$ Hz; $G=59$ dB). Finally, the ECG was acquired by placing 3 copper dry electrodes on the wrist band of the watch with 2 on the inside in contact with the wrist and 1 on the outside to place the index and middle finger. The 2 on the inside act as the right leg drive electrode and the positive lead, whereas the outside electrode is the negative lead. All electrodes were connected to an AFE (AD8232, Analog Devices) for ECG measurements. A microcontroller (Teensy 3.6, PJRC LLC) sampled the output of the accelerometer, PPG, and ECG AFE at 1 kHz. An onboard SD card was used to store the raw data for postprocessing, and a 1.2 Ah lithium-ion rechargeable battery was used to power the system. All instrumentation details were adopted from our previous work, with minor revisions [19].

The updated hardware, pictured in Figure 5, added modalities of sensing (ie, a gyroscope), included multiple wavelengths of LEDs for comparison with IR, improved the form factor for comfort and ease of use, and featured embedded systems innovations leveraged in this study. A more thorough description of the revised hardware is available in our most recent work [20]. An example of the serviceable automatic LED current scaling algorithm, detailed in our previous work [20], is highlighted in Figure 5. This proved to be an integral part of enabling this work; by adaptively adjusting the LED drive current, we were able to prevent saturation and variable PPG signal quality caused by varying contact pressure and, more importantly, prominent differences in skin tone among participants.

Figure 5. Pertinent multimodal hardware block diagram and adaptive light-emitting diode (LED) scaling. Main board with ATSAM4LS8 microcontroller ($\mu$C), ADXL355 triaxial accelerometer, BMG250 triaxial gyroscope, and BME280 environmental sensor using the serial peripheral interface for fast communication supporting higher sample rates. Sensor board used to acquire wrist photoplethysmogram (PPG) and electrocardiogram signals. Automatic LED current scaling in operation during data collection: showing an increase in contact pressure and subsequent saturation of the photodiode, mitigated by an automatic decrease in LED current and overall consequential improvement in PPG signal quality. ECG: electrocardiogram; LED: light-emitting diode; PD: photodiode; PPG: photoplethysmogram; RLD: right leg drive; SD: Secure Digital; SPI: Serial Peripheral Interface.
Human Subject Studies in a Diverse Population

All applicable results are presented as mean (SD). Figure 2 illustrates the correlation and Bland-Altman plots for our wearable PTT-based BP estimation of MAP, DBP, and SBP across all participants (N=44). The MAD was 2.90 mm Hg, 3.39 mm Hg, and 3.56 mm Hg for DBP, MAP, and SBP, respectively. The mean RMSE was 3.41 (SD 2.01) mm Hg, 3.95 (SD 2.42) mm Hg, and 6.28 (SD 3.44) mm Hg for DBP, MAP, and SBP, respectively. DBP and MAP estimation had better 95% CIs than SBP at 7.99 mm Hg, 9.42 mm Hg, and 14.59 mm Hg, respectively. The mean Pearson correlation coefficient (PCC) was 0.67 (SD 0.16), 0.63 (SD 0.31), and 0.50 (SD 0.41) for PTT-based DBP, MAP, and SBP estimation, respectively.

The MAD for the individual study populations (first cohort=26 participants and second cohort=18 participants) was 2.69 mm Hg and 3.20 mm Hg, 3.21 mm Hg and 3.64 mm Hg, and 5.17 mm Hg and 5.63 mm Hg for DBP, MAP, and SBP estimation, respectively. The mean RMSE for the individual study populations (first cohort=26 participants and second cohort=18 participants) was 3.19 (SD 1.64) mm Hg and 3.73 (SD 2.48) mm Hg, 3.78 (SD 2.06) mm Hg and 4.18 (SD 2.90) mmHg, and 6.26 (SD 3.25) mm Hg and 6.32 (SD 3.80) mm Hg for DBP, MAP, and SBP estimation, respectively. The mean PCC for the individual study populations (first cohort=26 participants and second cohort=18 participants) was 0.69 (SD 0.15) and 0.65 (SD 0.17), 0.68 (SD 0.23) and 0.55 (SD 0.38), and 0.38 (SD 0.33) and 0.39 (SD 0.49) for DBP, MAP, and SBP estimation, respectively.

The MAD for the 19 Black participants was 3.18 mm Hg, 3.72 mm Hg, and 5.84 mm Hg for DBP, MAP, and SBP estimation, respectively. The mean RMSE for all 19 Black participants was 3.72 (SD 2.41) mm Hg, 4.29 (SD 2.86) mm Hg, and 6.69 (SD 4.03) mm Hg for DBP, MAP, and SBP estimation, respectively. The mean PCC for all 19 Black participants was 0.64 (SD 0.17), 0.53 (SD 0.38), and 0.37 (SD 0.48) for DBP, MAP, and SBP estimation, respectively.

The MAD for the 10 participants who were obese was 2.69 mmHg, 3.17 mm Hg, and 5.02 mm Hg for DBP, MAP, and SBP estimation, respectively. The mean RMSE for all 10 participants who were obese was 3.28 (SD 2.59) mm Hg, 3.69 (SD 3.00) mm Hg, and 5.71 (SD 4.18) mm Hg for DBP, MAP, and SBP estimation, respectively. The mean PCC for all 10 participants who were obese was 0.65 (SD 0.18), 0.52 (SD 0.48), and 0.39 (SD 0.58) for DBP, MAP, and SBP estimation, respectively.

Figure 3 depicts the boxplots of the DBP calibration coefficients from our estimation model, K₁ and K₂, for four different demographic factors known to affect arterial stiffness: obesity, sex, race, and age. The DBP K₁ and K₂ values for nonobese (N=34) versus obese (N=10) participants are 2.38 (SD 1.99) mm Hg/s versus 1.20 (SD 0.88) mm Hg/s and 61.02 (SD 18.03) mm Hg versus 74.31 (SD 5.14) mm Hg, respectively. The DBP K₁ and K₂ values for male (N=25) versus female (N=19) participants are 2.65 (SD 2.18) mm Hg/s versus 1.40 (SD 0.98) mm Hg/s and 60.16 (SD 20.64) mm Hg versus 69.14 (SD 8.29) mm Hg, respectively. The DBP K₁ and K₂ values for non-Black (N=25) versus Black (N=19) participants are 2.63 (SD 2.21) mm Hg/s versus 1.44 (SD 0.94) mm Hg/s and 60.66 (SD 20.29) mm Hg versus 68.49 (SD 9.98) mm Hg, respectively. The DBP K₁ and K₂ values for young (N=31) versus older (N=13) participants are 2.38 (SD 2.09) mm Hg/s versus 1.47 (SD 0.87) mm Hg/s and 61.96 (SD 19.03) mm Hg versus 69.00 (SD 9.22) mm Hg, respectively.

Both K₁ and K₂ were significantly different between the nonobese and obese populations (P=0.045 and P=0.008, respectively). The female K₁ values were significantly (P=0.04) lower than those of their male counterparts. The K₁ values for Black participants were significantly (P=0.047) lower than those of the other races.

For the participants in the second cohort—all Black—with whom we used the newer version of the hardware [20] that included green and red LEDs in addition to the IR, the PCC for DBP estimation was 0.38 (SD 0.34), 0.59 (SD 0.44), and 0.65 (SD 0.17) when using the green λ=526 nm, red λ=660 nm, and IR λ=950 nm wavelength PPGs for the distal timing reference, respectively. The PCC for the IR and red wavelength PPGs was significantly (P=0.01 and P=0.048) higher than that of the green wavelength PPGs. However, the corresponding mean DBP RMSE was 3.95 (SD 2.53) mm Hg, 3.11 (SD 2.33) mm Hg, and 3.73 (SD 2.48) mm Hg for green, red, and IR, respectively.

Discussion

Principal Findings

To the best of our knowledge, this is the first study to accurately estimate DBP and MAP using noninvasive PTT measurements acquired from a holistic population, with considerable differences in body fat percentage, melanin levels, and vascular stiffness inherently influencing the measured levels, and BMI. Our results demonstrated that K₁ and K₂, for four different demographic factors known to affect arterial stiffness: obesity, sex, race, and age. The DBP K₁ and K₂ values for nonobese (N=34) versus obese (N=10) participants are 2.38 (SD 1.99) mm Hg/s versus 1.20 (SD 0.88) mm Hg/s and 61.02 (SD 18.03) mm Hg versus 74.31 (SD 5.14) mm Hg, respectively. The DBP K₁ and K₂ values for male (N=25) versus female (N=19) participants are 2.65 (SD 2.18) mm Hg/s versus 1.40 (SD 0.98) mm Hg/s and 60.16 (SD 20.64) mm Hg versus 69.14 (SD 8.29) mm Hg, respectively. The DBP K₁ and K₂ values for non-Black (N=25) versus Black (N=19) participants are 2.63 (SD 2.21) mm Hg/s versus 1.44 (SD 0.94) mm Hg/s and 60.66 (SD 20.29) mm Hg versus 68.49 (SD 9.98) mm Hg, respectively. The DBP K₁ and K₂ values for young (N=31) versus older (N=13) participants are 2.38 (SD 2.09) mm Hg/s versus 1.47 (SD 0.87) mm Hg/s and 61.96 (SD 19.03) mm Hg versus 69.00 (SD 9.22) mm Hg, respectively.

Both K₁ and K₂ were significantly different between the nonobese and obese populations (P<0.05 and P<0.008, respectively). The female K₁ values were significantly (P=0.04) lower than those of their male counterparts. The K₁ values for Black participants were significantly (P=0.047) lower than those of the other races.

For the participants in the second cohort—all Black—with whom we used the newer version of the hardware [20] that included green and red LEDs in addition to the IR, the PCC for DBP estimation was 0.38 (SD 0.34), 0.59 (SD 0.44), and 0.65 (SD 0.17) when using the green λ=526 nm, red λ=660 nm, and IR λ=950 nm wavelength PPGs for the distal timing reference, respectively. The PCC for the IR and red wavelength PPGs was significantly (P<0.01 and P<0.048) higher than that of the green wavelength PPGs. However, the corresponding mean DBP RMSE was 3.95 (SD 2.53) mm Hg, 3.11 (SD 2.33) mm Hg, and 3.73 (SD 2.48) mm Hg for green, red, and IR, respectively.

Accurately Estimating BP in a Diverse Population Using a Multimodal Wearable Device

We demonstrated the performance of our wrist-worn PTT-based device when used to estimate BP within a diverse population over the course of multiple unique perturbations. Our results for MAP and DBP passed the acceptable benchmarks for the BP estimation error set by the IEEE standard on wearable cuffless BP estimation devices (MAD≤5 mm Hg) [29]. We were still able to achieve a reliable correlation between PTT and BP even with several demographic factors such as age, melanin levels, and BMI inherently influencing the measured optical-PPG and mechanical-SCG signals.

https://mhealth.jmir.org/2021/8/e27466
The DBP estimation remained the most accurate, similar to our previous studies [19,20]; the foot of the PPG waveform, used as the distal timing reference, indicates the arrival of the pulse wave during end **diastole**. Similarly, the SBP estimation continued to perform the worst, as the peak of the pulse wave is the fiducial marker of the PPG that occurs during **systole**; however, the peak is not frequently extracted, as its true timing can be confounded by wave reflection interference, leading to unreliable PTT estimates [12]. Recent studies have demonstrated that the PTT computed using the diastolic foot of the PPG outperforms that using the systolic maximum for both DBP and SBP [37].

The DBP RMSE was relatively similar at both low and high values of DBP, which indicates that the diastolic foot was a dependable timing reference for calculating the PTT, irrespective of inherent participant-specific differences in BP. Although our SBP estimation was just outside the acceptable limits set forth by the IEEE standard (ie, MAD=5.36 mm Hg vs 5.00 mm Hg) [29], this error translates to a grade B classification [29] and therefore would still be clinically recommended for monitoring SBP [36]. Furthermore, the SBP range studied in this work was greater than 100 mm Hg, substantially higher than that reported in previous studies in the literature for wearable cuffless BP estimation, and a combination of different perturbations was used to modulate BP. Using a single perturbation would have led to an improved correlation [12,14], as in our previous work where we had only used exercise [19]. However, a comprehensive evaluation of this methodology would be incomplete without a procedure consisting of a wide variety of perturbations with different known physiological responses and pathways to modulate BP [14]. In addition, as noted in Figure 1, the exercise perturbation did not apparently produce a marked difference in BP due to several factors: (1) technical limitations in rapid calibration for the reference measurement (ie, finger-cuff continuous BP) and increased motion artifacts following exercise led to a greater percentage of beat removal in the early exercise section than any other task and (2) exercise does not necessarily consistently modulate BP in a predictable manner due to differences in participant-specific vasoactivity and contractility [12,38].

Only the DBP was examined for further analyses conducted below because, as previously mentioned, the distal timing reference used in this work (ie, the foot of the PPG waveform) occurs during diastole and therefore provides the most reliable estimation of DBP out of the three BP components [12]. The dependability of the diastolic foot and our robust DBP estimation were necessary before performing in-depth analyses with confidence. Although elevated SBP is considered to be the greatest predictor of future cardiovascular risk [39,40], elevated DBP has nonetheless been shown to independently increase the risk of subsequent cardiac events [39,41]. In addition, DBP is a greater contributor to MAP, which in older patients with isolated systolic hypertension, when compared with an equivalent increase in pulse pressure, has been shown to be a comparable independent predictor of both stroke and all-cause mortality [42]. Finally, DBP has been shown to be a more significant predictor than SBP of new-onset hypertension in adults aged ≤50 years [40,43-45]. This suggests that accurate DBP estimation using a wearable device can efficiently be used to incentivize people to make healthy lifestyle modifications **earlier in life**, central to the World Health Organization’s effort to reduce the global prevalence of hypertension [46].

**Essential Device Novelties Enabling Reliable PTT Computation**

For the first time, we demonstrated that noninvasive PTT measurements are reliable estimators of BP across a wide range of skin tones and BMI. Both DBP and MAP estimation for the 10 participants who are obese and 19 Black participants in this study were well under the IEEE requirement [29]. This was enabled by the highly sensitive hardware, multisensor approach, and automated LED current scaling that our custom wearable device offers [20]. The PPG array and adaptive LED current scaling allowed us to automatically mitigate poor signal quality issues due to misplacement, inherent differences in skin tone, and applied pressure that typically corrupts PPG signals. However, the most integral components of our PPG hardware were the IR wavelength LEDs.

We leveraged longer wavelengths of light for deeper penetration into the tissue to robustly acquire the PPG signal from arteries located deeper than the cutaneous vascular bed [12]. Cutaneous arteries are greatly affected by the changes in vascular tone expected from the perturbations we used to modulate BP herein (ie, cold pressor and exercise). Furthermore, as IR PPGs are more susceptible to motion artifacts than lower wavelength ones [12,47], our PPG-first signal quality assessment not only avoided these motion artifact corrupted waveforms because of their low SNR but also avoided moments where the SCG quality would naturally suffer as well. However, even the red PPGs had a considerably larger SD in their PCC than the IR PPGs, possibly because the IR wavelength, when compared with red, is less sensitive to the oxygen content of hemoglobin and has approximately half the skin absorption coefficient in Black individuals [12,27]. Despite statistically significant differences in the PCC using IR and red PPGs rather than green PPGs, the actual DBP RMSEs were comparable. This implies that when using the green PPGs for participants with a low PCC, our signal quality assessment algorithm removed beats with greater BP variability, resulting in a lower SD of DBP and consequent RMSE. Although even green wavelength PPGs have demonstrated the ability to reliably extract heart rate across a wide variety of skin tones [48], our data suggest that these shorter wavelengths cannot be used to dependably compute the PTT in a diverse population. In addition, although unconventional, our watch was placed on the ventral side of the wrist, which allows for both higher quality, convenient SCG acquisition and enhanced PPG SNR due to viable access to the radial artery and less melanin content than the dorsal side [49]. Therefore, existing smartwatches, beginning to slowly incorporate cuffless, noninvasive BP methodologies, may face even greater difficulties in achieving accurate PPG measurements across a broad range of skin tones.

Finally, our physiologically inspired PPG selection algorithm—to first select the PPG signals with the greatest systolic upstrokes—had an important role in reducing the BP estimation error. PPG waveforms with greater systolic upstrokes...
(ie, maximum derivative of the PPG waveform) offer improved PTT estimates and are key indicators of BP stemming from larger, more pulsatile, elastic arteries with greater distensibility [12,50]. In addition, several recent machine learning (ML) approaches to use the PPG signal for BP estimation have shown that the systolic upstroke is one of the most important features of the waveform [51,52]. Hence, the selection algorithm, by extracting information from these more reliable and clinically important arteries, was a central part of our ability to notice the demographic differences in arterial stiffness rooted in our calibration coefficients.

**Calibration Coefficients Capture Demographic Differences in Arterial Stiffness**

We observed that the participant-specific calibration coefficients used in the standard linear PTT-BP estimation model for DBP, shown in equation 1 (ie, $K_1$ and $K_2$), are significantly different between subpopulations with large variations in demographic factors known to affect arterial stiffness. We selected the four demographic categories (ie, obesity, sex, race, and age) based on the literature, emphasizing these as major determinants of differences in arterial stiffness and therefore risk factors for hypertension [31,34,35,53,54].

The $K_1$ value (ie, the slope of the line of best fit) is indicative of the underlying baseline vascular stiffness, whereas $K_2$ (ie, the intercept) represents the inherently correlated bias in baseline BP [12,55,56]. At the same BP, persons with greater arterial stiffness have inherently faster pulse wave velocities (PWVs) and therefore shorter PTTs than persons with normal arterial stiffness [12]. The $K_1$ value mitigates these differences in PTT-based estimation by capturing the intrinsic participant-specific arterial stiffness to output similar BP values. Therefore, with increasing arterial stiffness, we expected to find a lower $K_1$ value and a higher $K_2$ value, as observed in the PWV literature [55,56].

Obesity was the only comparison for which the differences in the $K_1$ and $K_2$ calibration coefficients were statistically significant. This coincides with the literature stating that obesity is one of the greatest age-normalized risk factors and contributors to arterial stiffness [57]. Otherwise, only the $K_1$ values in the sex and race comparisons were statistically significant between the groups. Although it has been shown that both females and Black individuals have greater arterial stiffness than similar-age males and White individuals [31,34,35], these two comparisons should be re-evaluated after increasing our recruitment. Approximately 47% (9/19 participants) of both the female and Black population were also obese. The age comparison was not statistically significant, although the older population followed a similar trend of a lower $K_1$ and higher $K_2$. This finding is not surprising, as significant differences in arterial stiffness and substantial augmentations in arterial remodeling are typically examined in participants aged ≥50 years [32,58].

**Limitations and Future Work**

**Refining Population Demographics and Investigating PTT, $K_1$, and $K_2$ as Potential Digital Biomarkers of Arterial Stiffness**

Overall, although this data set captured a more representative population in the range of end users for which consistent BP monitoring is recommended [59], our PTT-based device should be further tested in an exclusively older (ie, age ≥50 years), morbidly obese (ie, BMI >40 kg/m$^2$), and hypertensive population—with even distributions across sex, race, and skin tones along the Fitzpatrick scale—to truly understand the limits of this technology and supplement the findings herein.

Early vascular remodeling due to the demographic factors investigated in this work, not to mention socioeconomic factors affecting MUAs [15,16], predispose individuals who are obese and Black individuals to greater lifetime cardiovascular risk [30,35,57,60,61]. Therefore, future PTT-based BP estimation studies should closely monitor the calibration coefficients, $K_1$ and $K_2$, as potential intermediate digital biomarkers for longitudinal monitoring and the comparison of arterial stiffness among different persons [7]. Eventually, even PTT measurements, as PWV is already an independent predictor of arterial stiffness [62], may indicate subclinical differences in vascular resistance due to early stage arterial remodeling, the main precursor to hypertension [32].

**Reducing the Burden of Calibration**

Consistent recalibration poses a practical concern for PTT-based BP estimation. Hence, future studies should focus on evaluating the timeframe for which participant-specific calibration curves can reliably estimate BP and whether interparticipant and population-level curves can be sufficient. However, given the value of interpreting the calibration coefficients presented in this work, caution should be exercised due to the trade-off of sacrificing this potential usefulness when using generalized interparticipant models. Furthermore, the individual effects of the perturbations used to modulate BP in this experiment should be scrutinized, along with other exercises shown to substantially change BP [63-65]. The goal is to use perturbations that can consistently be leveraged to increase the dynamic range of BP measurements for calibration—critical to achieving optimal estimations at home in our previous work [20] and are achievable in low-resource settings.

**Leveraging ML and Hardware Advancements for Robust SCG AO Detection**

Similarly, to the instrumental role of the physiologically inspired PPG selection algorithm in this work, further exploration into automated SCG fiducial point detection algorithms may help extract the most informative SCG signals. Specifically, the SCG can be greatly affected by inaccurate placement of the watch; however, recent advancements using ML techniques have shown that the SCG waveform is modulated in a predictable manner during these placement inaccuracies [66]. Therefore, by interpreting these findings, one might be able to convert the measured SCG to the archetypal SCG or use a template-matching localization approach [67] for each
participant before extracting salient features from the optimal waveform.

In addition, annotating the exact AO point can be challenging because the signal not only has appreciable interparticipant variability, especially in a population with considerable differences in BMI, but can also be corrupted by motion artifacts. Although our technique for extracting the AO point has led to a high correlation between PTT and BP, in both our recent work [20] and this one, for a few sessions, we manually annotated the SCG to impose realistic constraints for the range of the pre-ejection period (PEP) and selected a consistent morphological peak across all tasks per participant. Eventually, robust identification of this timing reference is necessary for reliable automatic PTT computation, as the main advantage of using the PTT over the pulse arrival time (ie, the time from the R-wave of the ECG to the diastolic foot of the PPG) for BP estimation is its ability to account for changes in the nonnegligible cardio-electromechanical delay, that is, the PEP [12,68]. Furthermore, examining the other sensor data available at our disposal, such as filtering the SCG in a higher bandwidth (ie, fpass=30-125 Hz) to retain the phonocardiogram signal indicative of valve closures, using the three-axis gyrocardiogram or simply the other axes of the SCG, could prove to help with improving PEP estimation as shown in previous work [19,69].

Conclusions
We have demonstrated that a wrist-worn device, using noninvasive PTT estimates, can reliably and conveniently track BP in a diverse population. Leveraging the ubiquity of wearable devices can empower users to make healthy lifestyle modifications such as exercise, which can contribute to a significant reduction in arterial stiffness [30,70] by providing consistent feedback on progress [71-73]. In addition, digital health technologies that accurately estimate BP could potentially be used to titrate BP medications for patients with hypertension from the comfort of their homes [7,74]. In addition to these broader impacts, the knowledge gained from this study—especially when combined with the advent of low-profile, flexible electronics capable of robustly detecting physiological biosignals [75-78]—represents a significant step toward the unobtrusive monitoring of BP in ambulatory settings and health equity for persons in MUAs.

Acknowledgments
This work was funded in part by the National Institute of Health, National Institute of Biomedical Imaging and Bioengineering under grant U01EB018818-01, by the National Institute of Biomedical Imaging and Bioengineering Point-of-Care Technologies Research Network, Atlanta Center for Microsystems-Engineered Point-of-Care Technologies under grant 1U54EB027690, and by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR002378.

Authors’ Contributions
VGG co-developed the newer version of the hardware, performed human subject studies on the second cohort, analyzed the collected data, and coproduced the manuscript. AMC developed the older version of the hardware, co-developed the newer version of the hardware, co-conducted human subject studies on the first cohort, and assisted in advising the analysis and editing of the manuscript. HJ co-conducted human subject studies on the first cohort and assisted in both human subject studies on the second cohort and editing of the manuscript. AVS assisted in both conducting human subject studies on the second cohort and editing of the manuscript. DC assisted in both participant recruitment for the second cohort and editing of the manuscript. LNJ assisted in advising the study, participant recruitment for the second cohort, and editing of the manuscript. OTI guided the study and coproduced the manuscript.

Conflicts of Interest
OTI is a cofounder of and scientific advisor at Cardiosense, Inc, and a scientific advisor at Physiowave, Inc. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References


42. Domanski MJ, Davis BR, Pfeffer MA, Kastantin M, Mitchell GF. Isolated systolic hypertension: prognostic information provided by pulse pressure. Hypertension 1999 Sep;34(3):375-380. [doi: 10.1161/01.hyp.34.3.375] [Medline: 10489379]


Abbreviations

- ABP: arterial blood pressure
- AFE: analog front end
- AO: aortic valve opening
- BP: blood pressure
- BPF: bandpass filter
- DBP: diastolic blood pressure
- ECG: electrocardiogram
- IEEE: Institute for Electronics and Electrical Engineers
- IR: infrared
- LED: light-emitting diode
- MAD: mean absolute difference
- MAP: mean arterial pressure
- ML: machine learning
- MUA: medically underserved area
- PCC: Pearson correlation coefficient
- PEP: pre-ejection period
- PPG: photoplethysmogram
- PTT: pulse transit time
- PWV: pulse wave velocity
- RMSE: root mean square error
- SBP: systolic blood pressure
- SCG: seismocardiogram
- SNR: signal-to-noise ratio
Comparison of the Validity and Generalizability of Machine Learning Algorithms for the Prediction of Energy Expenditure: Validation Study

Ruairi O'Driscoll¹, MSc; Jake Turicchi¹, BSc; Mark Hopkins², PhD; Cristiana Duarte¹, PhD; Graham W Horgan³, PhD; Graham Finlayson¹, PhD; R James Stubbs¹, PhD

¹Appetite Control and Energy Balance Group, School of Psychology, University of Leeds, Leeds, United Kingdom
²School of Food Science and Nutrition, Faculty of Environment, University of Leeds, Leeds, United Kingdom
³Biomathematics & Statistics Scotland, Aberdeen, United Kingdom

Corresponding Author:
Ruairi O'Driscoll, MSc
Appetite Control and Energy Balance Group
School of Psychology
University of Leeds
Woodhouse
Leeds
United Kingdom
Phone: 44 113 343 2846
Email: psrod@leeds.ac.uk

Abstract

Background: Accurate solutions for the estimation of physical activity and energy expenditure at scale are needed for a range of medical and health research fields. Machine learning techniques show promise in research-grade accelerometers, and some evidence indicates that these techniques can be applied to more scalable commercial devices.

Objective: This study aims to test the validity and out-of-sample generalizability of algorithms for the prediction of energy expenditure in several wearables (ie, Fitbit Charge 2, ActiGraph GT3-x, SenseWear Armband Mini, and Polar H7) using two laboratory data sets comprising different activities.

Methods: Two laboratory studies (study 1: n=59, age 44.4 years, weight 75.7 kg; study 2: n=30, age=31.9 years, weight=70.6 kg), in which adult participants performed a sequential lab-based activity protocol consisting of resting, household, ambulatory, and nonambulatory tasks, were combined in this study. In both studies, accelerometer and physiological data were collected from the wearables alongside energy expenditure using indirect calorimetry. Three regression algorithms were used to predict metabolic equivalents (METs; ie, random forest, gradient boosting, and neural networks), and five classification algorithms (ie, k-nearest neighbor, support vector machine, random forest, gradient boosting, and neural networks) were used for physical activity intensity classification as sedentary, light, or moderate to vigorous. Algorithms were evaluated using leave-one-subject-out cross-validations and out-of-sample validations.

Results: The root mean square error (RMSE) was lowest for gradient boosting applied to SenseWear and Polar H7 data (0.91 METs), and in the classification task, gradient boost applied to SenseWear and Polar H7 was the most accurate (85.5%). Fitbit models achieved an RMSE of 1.36 METs and 78.2% accuracy for classification. Errors tended to increase in out-of-sample validations with the SenseWear neural network achieving RMSE values of 1.22 METs in the regression tasks and the SenseWear gradient boost and random forest achieving an accuracy of 80% in classification tasks.

Conclusions: Algorithms trained on combined data sets demonstrated high predictive accuracy, with a tendency for superior performance of random forests and gradient boosting for most but not all wearable devices. Predictions were poorer in the between-study validations, which creates uncertainty regarding the generalizability of the tested algorithms.

(JMIR Mhealth Uhealth 2021;9(8):e23938) doi:10.2196/23938

KEYWORDS
bioenergetics; energy balance; accelerometers; machine learning; validation
Introduction

Background

Participation in physical activity results in increased energy expenditure [1] and represents a key modifiable risk factor for cardiovascular disease, obesity, diabetes mellitus, cancer, and mortality [2]. Thus, longitudinal, unobtrusive, and accurate measurement of intraday physical activity energy expenditure would be highly valuable for health research. Activity trackers offer a scalable means for the continuous collection of physical activity data in free-living environments and, by extension, the measurement of energy expenditure. Unfortunately, the accuracy of activity trackers varies greatly between devices and activities [3,4], which limits their use when quantifying energy balance and activity behaviors.

The potential of machine learning techniques to model the complex interactions of accelerometer data, physiological variables, and the rate of energy expenditure has been recognized for some time. Rothney et al [5] trained an artificial neural network using raw accelerometer data as input to predict the energy expenditure in a whole-body calorimeter chamber. Poher et al [6] used quadratic discriminant analysis and a hidden Markov model to classify activity and subsequently estimated the proportion of time performing different activities. Research groups have built on these early findings and have reported highly accurate algorithms for a variety of activities [7-11]. Researchers often take two broad approaches when modeling physical activities: first, attempting to predict the rate of energy expenditure, and second, classifying a minute as sedentary activity, light physical activity, or moderate-to-vigorous physical activity (MVPA), both of which are important for health research. Regression approaches can be used to derive the total energy expenditure for a subject and this can subsequently be incorporated into energy balance models to calculate energy intake [12]. Alternatively, accurately determining the time an individual spends in broader categories of activity or the intensity of that activity can be important for public health guidance. For example, successful weight maintenance in the National Weight Control Registry and weight management recommendations are often defined based on the time an individual spends in MVPA [13]. Machine learning algorithms have the potential to enhance physical activity assessment beyond that of traditional count-based methods, which despite being more accessible, may not be sufficiently accurate for the assessment of energy expenditure and intensity classifications [14].

Recently, we demonstrated in a laboratory validation study that accelerometer and physiological sensor outputs can be modeled using random forests to predict the rate of energy expenditure (as a multiple of resting energy expenditure) in commercial and research-grade activity monitors. We demonstrated a low error in the prediction of energy expenditure [15]. The number of activities in which energy expenditure was measured in this study was limited, and the generalizability of these algorithms remains uncertain. A method for continued refinement of predictive algorithms is to obtain more than one data set [16] to provide larger, more diverse training data with more activities. More data present a new optimization problem, which (because of different assumptions made by different algorithms) means that there is no guarantee that any algorithm will minimize error on all problems [17]. For machine learning models to be used in general health research settings, it is critical to evaluate the generalizability of prediction algorithms. The extent to which an algorithm will generalize is influenced by the characteristics of the sample, activity types, size, and quality of the training data. One approach that addresses each of these limitations is to evaluate prediction algorithms on different samples using data collected under different conditions. In addition to generalizability, a combination of heterogeneous data sets collected under different experimental conditions may help to increase the accuracy of predictions [18].

Objectives

In this study, two distinct data sets of concurrent inputs from multiple wearable devices (ie, Fitbit Charge 2, ActiGraph GT3-x, SenseWear Armband Mini, and a polar chest strap) and measured energy expenditure (indirect calorimetry) are combined to develop predictive models of minute-level energy expenditure and physical activity. We aim to evaluate classification and regression algorithms to (1) predict the rate of energy expenditure and (2) classify a single minute as sedentary activity, light physical activity, or MVPA. Algorithms were validated using leave-one-subject-out cross-validation (LOSO) and out-of-sample validation. Concurrently, we evaluated the SenseWear armband, a device that has been shown to outperform accelerometer-based monitors when classifying activity minutes [19] and is one of the most accurate wrist or arm-based monitors for estimating energy expenditure [3].

Methods

Studies

This study aggregated the data collected as part of two separate studies at the Human Appetite Research Unit, University of Leeds. Participants were recruited from the local area using word-of-mouth and recruitment emails. Participants must have been at least 18 years of age, have been able to attend the research laboratory at the required intervals, be able to ambulate without assistance, they must not have been taking medications known to alter metabolic rate, and participants must not have had any cardiovascular, metabolic, renal disorders, illness, or injury that would increase the risk of medical events during physical activity. Both studies were approved by the University of Leeds, School of Psychology Ethics Committee (PSC-407 and PSC-744 for study 1 and 2, respectively), and all participants provided informed consent before participation in the study. The participant information for the samples is shown in Table 1. Study 2 had proportionately more males, lower age, lower average percentage of fat mass (FM), and a higher resting metabolic rate (RMR) on average.

https://mhealth.jmir.org/2021/8/e23938

JMIR MHEALTH AND UHEALTH O'Driscoll et al

(0)
The details of study 1 have been published previously [15]. The protocol of study 1 consisted of 10 activities, each performed for 5 minutes in the following order: sitting, standing, treadmill walking and incline walking (4 km/h), jogging, and incline jogging (6-8 km/h). Participants then rested for 3 minutes and transitioned to a cycle ergometer for low- and moderate-intensity cycling. After another period of recovery, participants performed a folding and sweeping task. Owing to a variation in physical fitness, the jogging task (n=49), incline jogging (n=30), and moderate cycling tasks (n=58) were not performed by all participants.

In study 2 (total energy expenditure from wearable devices study), participants visited the lab and refrained from eating or consuming caffeine for at least 4 hours. This exercise visit is the first of three visits to the laboratory conducted as part of a wider project. Weight and height were obtained from a SECA 704s stadiometer and electronic scale (SECA, Germany), and subsequently, an activity protocol was performed. All activities were performed in 5-minute increments, and the order was identical for all participants. First, resting tasks were performed where participants lay supine, sat in a backed chair, and then stood. Next, after a 2-minute unstructured transitional period, participants performed seated typing, standing ironing, and wiping surfaces while standing. After another 2-minute transition, participants walked on a treadmill at 4 km/h, walked at an incline of 5% at 4 km/h, and subsequently jogged at 7 km/h. The participants then rested for 10 minutes. After the unstructured resting period, participants performed low-intensity and moderate-intensity cycling, low-intensity and moderate-intensity rowing, and low-intensity and moderate-intensity cross-training (elliptical), with 1-minute transitions between each, and the intensity of the tasks was determined by a self-selected perceived exertion. In study 2, one participant did not perform rowing or elliptical tasks.

**Body Composition Assessment**

In both studies, body composition was estimated using air displacement plethysmography (BodPod, Life Measurement, Inc), n=57 in study 1 and n=30 in study 2. Study 2 is part of a wider study in which participants visited the laboratory three times, the first of which was the laboratory validation reported here. Body composition was measured at a subsequent visit to the laboratory in a fasting state.

### Energy Expenditure

This study used metabolic equivalents (METs) as the outcome variable, which served to eliminate the proportion of energy expenditure attributable to RMR. We first established the RMR of each participant, which was measured in the fasting state, before any exercise. In both studies, RMR was determined from VO$_2$ and VCO$_2$ data collected through a ventilated hood indirect calorimeter system (gas exchange measurement; Nutren Technology Ltd). In study 1, RMR was measured before exercise testing, and in study 2, which occurred on a subsequent visit to the laboratory. After researchers explained the procedures to the participants and an initial calibration process (approximately 10 minutes), VO$_2$ and VCO$_2$ were measured for 30 minutes in the supine position. The RMR was established from the VO$_2$ and VCO$_2$ of the 5-minute block with the lowest coefficient of variation [20]. If RMR data were unavailable (n=3 across both studies), we approximated the RMR with BMI-specific equations [21]. During the activity sessions, energy expenditure was obtained from a stationary metabolic cart (Vyntus CPX, Jaeger-CareFusion), and these data were expressed relative to the measured RMR of each subject to derive METs. Definitions of METs are inconsistent [22] and we took an individualized approach to METs calculations because the standard definition of METs may have limited applicability in some subjects [23].

### Devices

Accelerometer and physiological data were collected using various sensors in both protocols. The Polar H7 chest strap (Polar Electro) was used to measure the heart rate. An ActiGraph GT3-X accelerometer (ActiGraph) and a Fitbit Charge 2 (Fitbit Inc) were attached securely to the nondominant wrist. Participants also wore the SenseWear Armband Mini (BodyMedia Inc) on the upper arm.

### Data Aggregation

The sensor outputs were obtained from the device-specific software and aggregated to the minute level and time matched to the criterion energy expenditure data. Data loss attributable...
to device malfunction was as follows: in study 1, Fitbit data of 2 participants, ActiGraph data of 1 participant, and polar heart rate data of 1 participant were lost. In study 2, 1 SenseWear and 1 Fitbit data set were lost because of device failure. Given the slightly different data availability in each model, our results report the number of minutes used and the number of participants. All minutes in which energy expenditure data were available (ie, face mask was not removed) were included in this analysis, and the aggregation of the data sets by time was conducted in Python 3.7.6 and R version 3.6.3 (R Core Team).

For activity-specific analyses, we grouped activities into broader categories. Activities of daily living, which involved folding, sweeping, typing, ironing, and wiping surfaces. Distinct categories were assigned for cycling, elliptical, rowing, running, and walking. The sedentary activities involved all sitting, standing, and supine tasks. The transitional category refers to unstructured resting or transitional minutes.

**Features**

Predictive models were built for Fitbit, ActiGraph, and SenseWear, and the features used in each model are listed in Table 2. Each device used a combination of subject-level features, accelerometer features, and physiological features, which have been related to the rate of energy expenditure in previous studies [3,5,24-26]. The features varied depending on the feature availability of each device. Where small (limit of 5 minutes) heart rate gaps existed (eg, loss of signal between the respective heart rate sensor and the skin), we used linear interpolation to fill gaps. As activity in the preceding minutes influences the rate of energy expenditure at the measurement point [27], some time-lagged features were computed; for steps (Fitbit and SenseWear), vector magnitude (ActiGraph), Fitbit heart rate (Fitbit), and polar heart rate (SenseWear and ActiGraph), the change from t-1 minutes for each minute up to t-5 minutes were included as predictive features. In addition, the mean and SD of the current and last 5 minutes were used as predictive features. If time-lagged variables could not be computed due to missing data (ie, for the first minutes for each subject), we imputed backward using the next available observation.

As a constant variance is important for some of the algorithms tested in this study, all numeric features were standardized before training using the following formula:

\[
z = \frac{x - \mu}{sd}
\]

(1)

where \(\mu\) and \(sd\) refer to the variable mean and SD, respectively.
Table 2. Predictive features used in each of the models.

<table>
<thead>
<tr>
<th>Device and category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fitbit</strong></td>
<td></td>
</tr>
<tr>
<td>Subject features</td>
<td>Gender, age, height, weight, and sitting heart rate</td>
</tr>
<tr>
<td>Acceleration features</td>
<td>steps mean, steps difference (t-1, t-2, t-3, t-4, and t-5 minutes); steps mean and SD of last 5 minutes</td>
</tr>
<tr>
<td>Physiological features</td>
<td>Fitbit heart rate features:</td>
</tr>
<tr>
<td></td>
<td>Fitbit heart rate above sitting heart rate, Fitbit heart rate percentage of maximum heart rate, Fitbit heart rate mean, Fitbit heart rate difference (t-1, t-2, t-3, t-4, and t-5 minutes), and Fitbit heart rate mean and SD of last 5 minutes</td>
</tr>
<tr>
<td><strong>ActiGraph</strong></td>
<td></td>
</tr>
<tr>
<td>Subject features</td>
<td>Gender, age, height, and weight</td>
</tr>
<tr>
<td>Acceleration features</td>
<td>X, Y, Z features:</td>
</tr>
<tr>
<td></td>
<td>minimum, maximum, mean, SD; median crossings; 10th, 25th, 50th, 75th, 90th percentiles; correlations (XY, XZ, YZ); dominant frequency; dominant frequency magnitude</td>
</tr>
<tr>
<td></td>
<td>First order differential of X, Y, Z features:</td>
</tr>
<tr>
<td></td>
<td>minimum, maximum, mean, SD; median crossings; 10th, 25th, 50th, 75th, and 90th percentiles; correlations (XY, XZ, YZ); dominant frequency; dominant frequency magnitude</td>
</tr>
<tr>
<td></td>
<td>Vector magnitude features:</td>
</tr>
<tr>
<td></td>
<td>vector magnitude mean; vector magnitude difference (t-1, t-2, t-3, t-4, and t-5 minutes); vector magnitude mean and SD of last 5 minutes</td>
</tr>
<tr>
<td>Physiological features</td>
<td>Polar heart rate features:</td>
</tr>
<tr>
<td></td>
<td>polar heart rate above sitting heart rate; polar heart rate percentage of maximum heart rate; polar heart rate mean; polar heart rate difference (t-1, t-2, t-3, t-4, and t-5 minutes); polar heart rate mean and SD of last 5 minutes</td>
</tr>
<tr>
<td><strong>SenseWear</strong></td>
<td></td>
</tr>
<tr>
<td>Subject features</td>
<td>Gender, age, height, and weight</td>
</tr>
<tr>
<td>Acceleration features</td>
<td>X, Y, Z features:</td>
</tr>
<tr>
<td></td>
<td>peaks, mean of absolute differences, average;</td>
</tr>
<tr>
<td></td>
<td>Steps features:</td>
</tr>
<tr>
<td></td>
<td>steps mean; steps difference (t-1, t-2, t-3, t-4, and t-5 minutes); steps mean and SD of last 5 minutes</td>
</tr>
<tr>
<td>Physiological features</td>
<td>Polar heart rate features:</td>
</tr>
<tr>
<td></td>
<td>polar heart rate above sitting heart rate; polar heart rate percentage of maximum heart rate; polar heart rate mean; polar heart rate difference (t-1, t-2, t-3, t-4, and t-5 minutes); polar heart rate mean and SD of last 5 minutes; and SenseWear sensors: near body temperature average, Galvanic skin response average, skin temperature average</td>
</tr>
</tbody>
</table>

For each device, the subject characteristics, acceleration features, and physiological features are listed.

**Algorithms**

The SenseWear outputs a MET estimate that we evaluated in this study (SenseWear manufacturer). We also tested several machine learning algorithms for regression and classification tasks, which are described below. In the regression tasks, algorithms predicted a MET value for each minute, and in the classification tasks, algorithms classified activity categories for each minute. The activity classifications were as follows: sedentary activity (≤1.5 METs), light physical activity (>1.5 and <3 METs), and MVPA (≥3.0 METs) [18,28,29]. For each algorithm, the hyperparameters were informed by a random search through a range of potential hyperparameters in the preliminary tuning experiments. Random search iterates over a grid of randomly selected combinations of hyperparameters, rather than exploring every possible combination of features, and therefore offers a significant computational advantage over a grid-search approach [30]. Each random search was conducted with the RandomizedSearchCV class in Scikit Learn [31], using three-fold cross-validation. The specific parameters for each algorithm are detailed in Multimedia Appendix 1, and except for the neural network models (explained in the following section), the scoring or loss criterion was the default loss or scoring metrics within Scikit Learn. All algorithms were trained using Keras-GPU [32] or Scikit Learn [31].

**Random Forest**

The random forest algorithm was used for regression and classification tasks [33]. Random forests involve training of multiple decision trees on data subsamples. Importantly, when splitting these decision trees, only a subsample of the potential predictors is used, which serves to decorrelate the trees. The predictions of each tree can then be combined to produce a majority vote (classification) or continuous prediction (regression). The optimal hyperparameters of the algorithm were estimated in the tuning experiments and included the number of trees, number of samples required to split a tree, number of samples per leaf, total predictors, and the depth of
trees. In regression, the quality of a split was assessed with mean square error, and in classification, Gini impurity was used. Algorithms were implemented using the RandomForestClassifier and RandomForestRegressor classes in Scikit Learn [31].

**Gradient Boosting**

For the regression and classification tasks, we used the gradient boosting algorithm. Similar to random forests, this algorithm is a tree-based ensemble method. However, where random forests may be considered to use a bagging approach, gradient boosting uses boosting to learn. Boosting involves the sequential growth of small (weak) decision trees. Each tree is trained using the residuals of the previous estimator and subsequently added to the fitted function to update the residuals. In the boosting phase, a learning rate parameter penalizes the contribution of each tree to the overall model, thereby slowing the learning [34]. The gradient boosting hyperparameters were tuned in the random search experiments and included the number of boosting stages, the maximum depth of the estimators, learning rate, number of samples required to split a node, the number of samples per leaf, and the maximum number of predictors. In the regression, the loss function was least squares, and in classification, deviance was used. Algorithms were implemented using the GradientBoostingClassifier and GradientBoostingRegressor classes in Scikit Learn [31].

**Neural Networks**

The third algorithm, used in both regression and classification tasks, was artificial neural networks. Neural networks allow complex, nonlinear functions to be modeled and comprise layers of interconnected neurons. At each neuron, inputs are subjected to a numerical activation function, and then passed through subsequent hidden layers of neurons to an output layer [34,35]. In the training process, the interneuronal weights of the network are refined relative to a loss function (ie, mean square error or cross-entropy). Neural networks in the classification studies used the sparse categorical cross-entropy loss function, and in the regression setting, the loss was the mean square error. We tuned the learning rate of each network, the number of layers, and the number of neurons. Neural networks hidden layers used the relu activation function, and classification models used a softmax activation in the output layer, both classification and regression networks used the Adam optimizer.

**K-Nearest Neighbors**

For classification tasks, we tested the K-nearest neighbor (KNN) algorithm. This algorithm assigns a given point to a particular class based on the majority point of the k nearest neighbors, where the neighbors of a given point are defined by a distance metric (ie, Euclidian, Minkowski, or Manhattan) [34]. Hyperparameters adjusted in the training process included the number of neighbors in each neighborhood (k), distance metrics, and the weight applied to each of the observations in a neighborhood. KNN was implemented with Scikit Learn [31], using the KNeighborsClassifier class.

**Support Vector Machine**

The final classification model tested was a support vector machine classifier with a radial basis function [35]. A support vector machine aims to find a separating hyperplane between classes by maximizing the distance between the points and the hyperplane. In this study, we tuned the regularization parameter (C) and gamma, which defines the magnitude of the effect of specific training examples. The support vector machine classifier was implemented with the SVC class in Scikit Learn [31].

**Statistical Analyses**

We conducted two validation approaches for all the analyses and algorithms. First, LOSO validations, where algorithms are trained on all but the data of 1 participant, and the participant is held back for validation. This process was repeated until all participants had served as the validation participant once. Second, we used an out-of-sample validation in which the entire data set from one study was used as training data, and the second study was used as an out-of-sample validation. Regression algorithms were evaluated by root mean square error (RMSE), mean absolute percentage error (MAPE) with the Metrics package in R and concordance correlation coefficient (CCC) with DescTools. Agreement statistics were calculated at the minute level; however, for visualization purposes, we computed the RMSE at the level of individuals and plotted these values. Equivalence tests were used to determine if the true METs and predicted METs were statistically equivalent; tests used equivalence bounds of 10%, and to be considered equivalent, the 90% CI must fall within the equivalence bounds. Finally, linear mixed models with a random intercept of subject ID were used to investigate differences in RMSE between the models. Comparisons were conducted using the Lme4 [36] package in R, with P values adjusted by the Bonferroni method in post hoc comparisons. For classification tasks, we report the κ statistic, which compares the accuracy of the predictions to that of a random system. We also report accuracy, where accuracy is the proportion of cases that were classified correctly and the F1 score. All classification statistics were calculated using the Caret [37] package in R. A P value of <.05 was used to determine statistical significance, where P values were reported.

**Results**

**Regression**

A total of 89 participant activity sessions were included in this sample, and all models could be evaluated on at least 5448 minutes of data in the LOSO validations. The regression algorithms predicting energy expenditure are presented for minute-level data in Table 3 and are visually displayed in Figure 1. Our results demonstrate that the greatest error in METs was observed for the manufacturer-provided SenseWear estimates, with MAPE and RMSE values of 34.54 and 1.86, respectively. For ActiGraph, the RMSE was lowest for gradient boosting (0.93 METs), which also achieved the lowest MAPE of any ActiGraph model (17.88%). For the Fitbit models, the random forest and gradient boosting had equal RMSE (1.36 METs), but a slightly lower MAPE was achieved by the random forest. For the SenseWear, the gradient boost had the lowest RMSE value (0.91 METs), and this was the lowest RMSE of all those tested. The neural network models were associated with a greater overall RMSE for the ActiGraph, Fitbit, and SenseWear models.
Activity-specific MET predictions are presented in Multimedia Appendix 2, and the RMSE is shown in Figure 2. For all activities tested, tree-based models (gradient boost or random forest) applied to ActiGraph or SenseWear data were superior, as measured by RMSE. The manufacturer estimates of SenseWear had the highest RMSE for all activities aside from sedentary activities, in which only the ActiGraph gradient boost and random forest had a lower RMSE. Notably, all Fitbit models overestimated sedentary activities and had the highest RMSE in this category. The pairwise comparisons between models are presented in Multimedia Appendix 3 for each of the comparisons shown in Figure 1 and Figure 2. An example of the model predictions for a single subject is shown in Figure 3.

Table 4 shows the statistics for the between-study predictions. Notably larger errors were observed relative to the LOSO validations, with the Fitbit gradient boost reaching a RMSE of 1.92 METs (neural network) when study 1 was used as the training data. To estimate the relative importance of each of the features used in each model, permutation importance has been reported in Multimedia Appendix 4.

Table 3. Leave-one-subject-out cross-validation results for each of the regression models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Participants, n (%)</th>
<th>Minutesa</th>
<th>Predicted (METs), mean (SD)</th>
<th>True (METs), mean (SD)</th>
<th>MAPEc</th>
<th>RMSEd</th>
<th>CCCe (95% CI)</th>
<th>Equivalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWA manufacturer</td>
<td>5533</td>
<td>88 (99)</td>
<td>3.8 (2.49)</td>
<td>4.04 (2.59)</td>
<td>34.54</td>
<td>1.86</td>
<td>0.73 (0.72-0.74)</td>
<td></td>
</tr>
<tr>
<td>AG gradient boost</td>
<td>5517</td>
<td>87 (98)</td>
<td>4.04 (2.35)</td>
<td>4.04 (2.59)</td>
<td>17.88</td>
<td>0.93</td>
<td>0.93 (0.93-0.93)</td>
<td>Equivalenti</td>
</tr>
<tr>
<td>AG neural network</td>
<td>5517</td>
<td>87 (98)</td>
<td>4.05 (2.55)</td>
<td>4.04 (2.59)</td>
<td>21.65</td>
<td>1.06</td>
<td>0.9 (0.9-0.91)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>AG random forest</td>
<td>5517</td>
<td>87 (98)</td>
<td>4.05 (2.32)</td>
<td>4.04 (2.59)</td>
<td>18.36</td>
<td>0.94</td>
<td>0.93 (0.92-0.93)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB gradient boost</td>
<td>5448</td>
<td>86 (97)</td>
<td>4.03 (2.19)</td>
<td>4.01 (2.58)</td>
<td>30.22</td>
<td>1.36</td>
<td>0.84 (0.83-0.84)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB neural network</td>
<td>5448</td>
<td>86 (97)</td>
<td>4.02 (2.28)</td>
<td>4.01 (2.58)</td>
<td>32.27</td>
<td>1.45</td>
<td>0.82 (0.82-0.83)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB random forest</td>
<td>5448</td>
<td>86 (97)</td>
<td>4.03 (2.14)</td>
<td>4.01 (2.58)</td>
<td>30.10</td>
<td>1.36</td>
<td>0.84 (0.83-0.84)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA gradient boost</td>
<td>5492</td>
<td>87 (98)</td>
<td>4.04 (2.39)</td>
<td>4.04 (2.6)</td>
<td>17.83</td>
<td>0.91</td>
<td>0.93 (0.93-0.94)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA neural network</td>
<td>5492</td>
<td>87 (98)</td>
<td>4.05 (2.47)</td>
<td>4.04 (2.6)</td>
<td>19.56</td>
<td>0.96</td>
<td>0.93 (0.92-0.93)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA random forest</td>
<td>5492</td>
<td>87 (98)</td>
<td>4.05 (2.35)</td>
<td>4.04 (2.6)</td>
<td>18.25</td>
<td>0.92</td>
<td>0.93 (0.93-0.93)</td>
<td>Equivalent</td>
</tr>
</tbody>
</table>

aMinutes refers to the number of minutes the algorithms are validated on.

bMETs: metabolic equivalents.

cMAPE: mean absolute percentage error.

dRMSE: root mean square error.

eCCC: concordance correlation coefficient CCC is presented with 95% CIs.

fSWA: SenseWear.

gThe model is not statistically equivalent to the criterion.

hAG: ActiGraph.

iEquivalent implies that the model is statistically equivalent to the criterion.

jFB: Fitbit.
Figure 1. Boxplots demonstrating the root mean square error overall for each of the tested models. AG: ActiGraph; FB: Fitbit; RMSE: root mean square error; SWA: SenseWear.

Figure 2. Boxplots demonstrating the root mean square error for each of the tested models in specific activity categories. ADL: activities of daily living; AG: ActiGraph; FB: Fitbit; RMSE: root mean square error; SWA: SenseWear.
Figure 3. A time series plot showing metabolic equivalents predicted by the models tested in this study (colored solid line) and indirect calorimeter (black dashed line), for a single subject in study 2. The x-axis represents the measurement time. Minutes 1-15=sedentary; minutes 16-17=transitional/unstructured; minutes 18-32=activities of daily living (typing, wiping surfaces, and ironing); minutes 33-34=transitional/unstructured; minutes 35-44=walking; minutes 45-49=running; minutes 50-59=transitional/unstructured; minutes 60-69=cycling; minutes 71-80=rowing; and minutes 82-91=elliptical. Participants performed cycling, rowing, and elliptical tasks at self-selected low and moderate intensity for 5 minutes each. AG: ActiGraph; FB: Fitbit; METs: metabolic equivalents; SWA: SenseWear.
Table 4. Out-of-sample results for each of the regression models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training data</th>
<th>Minutes\textsuperscript{a}</th>
<th>Predicted (METs\textsuperscript{b}), mean (SD)</th>
<th>True (METs), mean (SD)</th>
<th>MAPE\textsuperscript{c}</th>
<th>RMSE\textsuperscript{d}</th>
<th>CCC\textsuperscript{e} (95% CI)</th>
<th>Equivalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG\textsuperscript{f} gradient boost</td>
<td>Study 1</td>
<td>2690</td>
<td>4.03 (1.9)</td>
<td>3.93 (2.66)</td>
<td>36.35</td>
<td>1.37</td>
<td>0.82 (0.81-0.83)</td>
<td>Equivalent\textsuperscript{g}</td>
</tr>
<tr>
<td>AG neural network</td>
<td>Study 1</td>
<td>2690</td>
<td>4.07 (2.48)</td>
<td>3.93 (2.66)</td>
<td>29.75</td>
<td>1.33</td>
<td>0.87 (0.86-0.88)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>AG random forest</td>
<td>Study 1</td>
<td>2690</td>
<td>3.97 (1.79)</td>
<td>3.93 (2.66)</td>
<td>39.50</td>
<td>1.51</td>
<td>0.78 (0.77-0.79)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB\textsuperscript{h} gradient boost</td>
<td>Study 1</td>
<td>2630</td>
<td>3.76 (1.7)</td>
<td>3.88 (2.65)</td>
<td>47.55</td>
<td>1.89</td>
<td>0.64 (0.62-0.66)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB neural network</td>
<td>Study 1</td>
<td>2630</td>
<td>3.65 (1.86)</td>
<td>3.88 (2.65)</td>
<td>47.40</td>
<td>1.92</td>
<td>0.65 (0.63-0.67)</td>
<td>—\textsuperscript{i}</td>
</tr>
<tr>
<td>FB random forest</td>
<td>Study 1</td>
<td>2630</td>
<td>3.76 (1.66)</td>
<td>3.88 (2.65)</td>
<td>47.45</td>
<td>1.87</td>
<td>0.64 (0.63-0.66)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA\textsuperscript{j} gradient boost</td>
<td>Study 1</td>
<td>2633</td>
<td>3.92 (2.13)</td>
<td>3.94 (2.68)</td>
<td>27.35</td>
<td>1.23</td>
<td>0.87 (0.86-0.88)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA neural network</td>
<td>Study 1</td>
<td>2633</td>
<td>3.88 (2.26)</td>
<td>3.94 (2.68)</td>
<td>27.07</td>
<td>1.22</td>
<td>0.88 (0.87-0.89)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA random forest</td>
<td>Study 1</td>
<td>2633</td>
<td>3.91 (2.07)</td>
<td>3.94 (2.68)</td>
<td>29.54</td>
<td>1.28</td>
<td>0.86 (0.85-0.87)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>AG gradient boost</td>
<td>Study 2</td>
<td>2827</td>
<td>4.46 (2.14)</td>
<td>4.15 (2.52)</td>
<td>31.49</td>
<td>1.36</td>
<td>0.83 (0.82-0.84)</td>
<td>—</td>
</tr>
<tr>
<td>AG neural network</td>
<td>Study 2</td>
<td>2827</td>
<td>4.24 (2.56)</td>
<td>4.15 (2.52)</td>
<td>29.00</td>
<td>1.42</td>
<td>0.84 (0.83-0.85)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>AG random forest</td>
<td>Study 2</td>
<td>2827</td>
<td>4.45 (2.1)</td>
<td>4.15 (2.52)</td>
<td>31.47</td>
<td>1.38</td>
<td>0.82 (0.81-0.84)</td>
<td>—</td>
</tr>
<tr>
<td>FB gradient boost</td>
<td>Study 2</td>
<td>2818</td>
<td>4.11 (2.06)</td>
<td>4.13 (2.51)</td>
<td>34.38</td>
<td>1.66</td>
<td>0.74 (0.72-0.75)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB neural network</td>
<td>Study 2</td>
<td>2818</td>
<td>4.01 (2.04)</td>
<td>4.13 (2.51)</td>
<td>33.10</td>
<td>1.56</td>
<td>0.77 (0.75-0.78)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB random forest</td>
<td>Study 2</td>
<td>2818</td>
<td>4.21 (2.04)</td>
<td>4.13 (2.51)</td>
<td>33.79</td>
<td>1.62</td>
<td>0.75 (0.73-0.77)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA gradient boost</td>
<td>Study 2</td>
<td>2859</td>
<td>4.15 (2.13)</td>
<td>4.14 (2.51)</td>
<td>24.90</td>
<td>1.25</td>
<td>0.86 (0.85-0.87)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA neural network</td>
<td>Study 2</td>
<td>2859</td>
<td>3.94 (2.36)</td>
<td>4.14 (2.51)</td>
<td>25.65</td>
<td>1.25</td>
<td>0.87 (0.86-0.88)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA random forest</td>
<td>Study 2</td>
<td>2859</td>
<td>4.2 (2.13)</td>
<td>4.14 (2.51)</td>
<td>25.72</td>
<td>1.26</td>
<td>0.85 (0.84-0.86)</td>
<td>Equivalent</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Minutes refers to the number of minutes the algorithms are validated on.
\textsuperscript{b}METs: metabolic equivalents.
\textsuperscript{c}MAPE: mean absolute percentage error.
\textsuperscript{d}RMSE: root mean square error.
\textsuperscript{e}CCC: concordance correlation coefficient CCC is presented with 95% CIs.
\textsuperscript{f}AG: ActiGraph.
\textsuperscript{g}Equivalent implies that the model is statistically equivalent to the criterion.
\textsuperscript{h}FB: Fitbit.
\textsuperscript{i}The model is not statistically equivalent to the criterion.
\textsuperscript{j}SWA: SenseWear.

**Classification**

Figure 4 presents the results of the LOSO classification experiments for all classification algorithms and the SenseWear manufacturer estimates. Classes were slightly imbalanced, approximately 19.4% sedentary activity, 22.4% light physical activity, and 58.2% MVPA with small differences between devices due to data availability. The highest accuracy for Fitbit models was the random forest (78.21%), for the ActiGraph models, the random forest achieved the highest accuracy (84.56%), and for the SenseWear models, the gradient boosting algorithm (85.49%) was the most accurate.

Multimedia Appendix 5 provides class-specific statistics for each model. Models tended to perform worse in light activity with F1 scores ranging from 0.20 (SenseWear neural network) to 0.66 (SenseWear gradient boost). In sedentary activities, the F1 score was improved with a range of 0.54 (Actigraph support vector machine) to 0.83 (four models). For MVPA, the F1 score ranged from 0.80 (Actigraph support vector machine) to 0.93 (three models).
**Between-Study Predictions**

The between-study classification accuracies are listed in Table 5. In most cases, when study 1 served as the training data, lower accuracy was observed. When study 1 served as the training data, the accuracy ranged from 0.55 (ActiGraph support vector machine) to 0.80 (two models). When study 2 served as the training data, the accuracy ranged from 0.65 (ActiGraph support vector machine) to 0.79 (three models).
Table 5. Between-study classification results for each of the classification models.

<table>
<thead>
<tr>
<th>Training data and model</th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG(^a) gradient boost</td>
<td>0.75</td>
<td>0.79</td>
</tr>
<tr>
<td>AG k-nearest neighbors</td>
<td>0.61</td>
<td>0.72</td>
</tr>
<tr>
<td>AG neural network</td>
<td>0.72</td>
<td>0.75</td>
</tr>
<tr>
<td>AG random forest</td>
<td>0.74</td>
<td>0.79</td>
</tr>
<tr>
<td>AG support vector machine</td>
<td>0.55</td>
<td>0.80</td>
</tr>
<tr>
<td>FB(^b) gradient boost</td>
<td>0.67</td>
<td>0.65</td>
</tr>
<tr>
<td>FB k-nearest neighbors</td>
<td>0.68</td>
<td>0.73</td>
</tr>
<tr>
<td>FB neural network</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>FB random forest</td>
<td>0.67</td>
<td>0.73</td>
</tr>
<tr>
<td>FB support vector machine</td>
<td>0.67</td>
<td>0.73</td>
</tr>
<tr>
<td>SWA(^c) gradient boost</td>
<td>0.80</td>
<td>0.73</td>
</tr>
<tr>
<td>SWA k-nearest neighbors</td>
<td>0.74</td>
<td>0.78</td>
</tr>
<tr>
<td>SWA neural network</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td>SWA random forest</td>
<td>0.80</td>
<td>0.76</td>
</tr>
<tr>
<td>SWA support vector machine</td>
<td>0.68</td>
<td>0.79</td>
</tr>
</tbody>
</table>

\(^a\)AG: ActiGraph.  
\(^b\)FB: Fitbit.  
\(^c\)SWA: SenseWear.

Discussion

Principal Findings

This study aggregated two laboratory data sets to build on previous work demonstrating the potential for machine learning algorithms to produce accurate estimates of METs and intensity classes in a diverse set of activities and participants. In both regression and classification settings, we observed the smallest errors in energy expenditure predictions when applying tree-based algorithms (ie, random forest and gradient boosting) to SenseWear and ActiGraph outputs with the RMSE and classification errors generally being higher for Fitbit models. In almost all cases, the error was smaller than the SenseWear manufacturer estimates, and in out-of-sample generalizability experiments, we observed greater error and lower accuracy when compared with the LOSO validations. We believe that
this is the first study to classify the intensity of activity using machine learning algorithms in Fitbit devices. In Fitbit models, we demonstrated accuracies up to approximately 78% (κ=0.6), with superior performance observed for sedentary activity and MVPA classifications, but these were generally less accurate than ActiGraph and SenseWear models, where up to approximately 85% accuracy (κ=0.74) was achieved. Taken together, and if these results are verified in free-living, ecologically valid examples, these findings imply that highly accurate estimates of energy expenditure, sedentary activity, and MVPA behaviors can be estimated by the wearables tested here.

Algorithm Accuracy

We used neural networks, random forests, and gradient boosting in regression tasks. In previous studies, neural networks and random forests have been shown to be effective in modeling energy expenditure [8,9], and our results confirm this to an extent. The RMSE values observed in the trained models ranged from 0.91 METs to 1.45 METs, which improve upon the SenseWear manufacturer value of approximately 1.86 METs. However, when the average METs in this study were considered (approximately 4 METs), it was evident that the energy expenditure prediction could be further improved. It is noteworthy that neural networks resulted in the highest RMSE for all 3 devices and performed particularly poorly for Fitbit models. Similarly, Kate et al [38] showed that neural networks resulted in bias significantly different from 0, compared with bagged decision trees and numerous other algorithms, which were not statistically different. Despite the utility of deep neural networks to model highly nonlinear functions, in some use cases, the no free lunch theorems broadly state that there will not be an optimal algorithm for all tasks [17]. Indeed, for our data sets, tree-based ensemble models are generally superior for both learning tasks. It may be that a higher RMSE can be reduced by larger training sets [39].

We generated lagged accelerometer and heart rate variables for each model because the rate of energy expenditure depends on the rate of work in preceding minutes [27], and the relative importance of these metrics is evidenced in the variable importance analyses. Including time-lagged features allows for a clearer distinction between minutes that are relatively similar in their accelerometer pattern but differ in their measured energy expenditure, that is, sitting for a prolonged period versus sitting immediately after running. Transitional minutes were on average approximately 3 METs (largely attributable to the activity in the preceding minutes), compared with sedentary minutes, which averaged approximately 1.3 METs, yet the error statistics were generally comparable with those observed in sedentary minutes, indicating that algorithms could distinguish between those minutes. More advanced neural network architectures (ie, recurrent neural networks) [40] may further the ability of models to capture the temporal dependencies of energy expenditure.

Generalization

Although many studies have reported low errors when using machine learning approaches in the estimation of energy expenditure or classification of activity, external (out-of-sample) validations are rarer and the opportunity to identify cases of overfitting has been limited. Therefore, we used an out-of-sample validation between the two data sets. In all cases, we observed performance degradation when compared with the LOSO validations. Some of this reduction in accuracy is probably attributable to differences in protocols, activities, and participants, which means that algorithms do not have similar minutes on which to train. In addition, it is possible that the algorithms overfit the data. Overfitting occurs when a complex model learns the noise in the training data, which does not represent the true underlying function between the inputs and the output [41]. Previous studies have used out-of-sample validation or validation in free-living environments [10,42,43], and when compared with laboratory validations, errors may increase. Concerning the classification of physical activity intensity in multiple samples, a previous study reported reductions in out-of-sample accuracy relative to the within-sample validated models, in some algorithm and data set comparisons [44]. However, the machine learning models still outperformed the Euclidean norm minus one GGIR classification method in out-of-sample testing. In another comprehensive generalizability study, five lab-based heterogeneous data sets were used to predict exercise intensity. This study found that when models were applied to a different data set than those they were generated on, model accuracy decreased from 72-95% to 41-60% [18]. These drops are notably higher than those in this study, and this is probably attributable to the greater differences in the accelerometer models, wear position, and samples across the five data sets. However, caution must be exercised in a comparison between studies, as the balance of classes is likely to differ and therefore influence some evaluation metrics.

Classification

Our LOSO validations demonstrated a relatively high predictive accuracy (75-85%). However, research-grade device models (ActiGraph and SenseWear) were superior. Fitbit devices provide estimates of time in each category (ie, sedentary, light, and MVPA), but the criteria and algorithms remain proprietary. Feehan et al [45] compared estimates of time in intensities with devices such as ActiGraph and Actical, and concluded that more than 80% of studies reported errors >10% with mean differences ranging between 44% and 632% for estimations of activity above light intensity. Importantly, the devices used for comparison in many studies have varying cut points and are not necessarily gold standards. Our results indicate that the application of machine learning to intensity classification can refine the large errors observed in previous studies. Despite the promising results, we emphasize that laboratory studies have limited ecological validity, and future research should seek to address this. Whole-room indirect calorimetry would likely allow more realistic behaviors to be studied while providing a gold standard comparator.

Strengths and Limitations

A strength of this study is the aggregation of two data sets to provide a more comprehensive and variable data set on which to train models, although the measures (sensors and indirect calorimetry) were the same between studies. The tested cohorts differed demographically, and the protocols were heterogeneous,
which provides a good estimate of the applicability of the tested models. Combining data sets also leads to a larger number of participants (n=89), which is a larger sample size than much of the previous literature [7,9,10,44,46,47]. In general, an increase in training observations is considered a mechanism for enhancing performance [41], and the results of this study provide some evidence that this is the case in both commercial and research-grade accelerometers.

Another strength of this study is the testing of numerous algorithm and device combinations. A previous study developed a multilayer neural network that was trained on a wearable system including a vest for electrocardiogram measurements and 4 accelerometers (one on each wrist and thigh) [47]. Despite the small bias, this is unlikely to be a feasible means of assessing free-living energy balance behaviors. Participant discomfort and sensor removal present additional biases (ie, missing data), which may require additional modeling approaches to address [48-50]. The threshold of practicality varies depending on the size, duration, computational resources, and specific aims of the research study. Therefore, the development of three models with varying requirements is a central advantage of this study.

Testing both classification and regression algorithms in the same devices enhances the use of the results of this study. One area of future work is to explore combined classification and regression approaches, similar to the branched models of the Actiheart [51] or stacked ensemble approaches. This may be effective in producing refined estimates of total daily energy expenditure in free-living subjects, given that most of a day comprises resting or sedentary minutes and some of our models slightly overestimate sedentary activities, although depending on the classification or regression methods, this could incur additional computational costs when applying this to larger data sets. Future work in our lab will examine the application of such models to free-living environments against a doubly labeled water criterion.

A limitation of this study is the lack of a true testing set. Rather, we attempt to develop an unbiased estimate of the true test error by (1) testing on unseen participants and (2) testing on an unseen data set. In the former, the within-subject data are generally more correlated than the between-subject data, and this method represents the closest approximation of how such a model would perform in practice [8]. In the latter, this is extended so that the training and testing sets comprised different participants and protocols. Beyond these validation approaches, the ultimate test of the results presented here is a free-living validation for energy expenditure and intensity classes. The total daily energy expenditure can be validated using the doubly labeled water method over a 7- to 14-day period [52], and the results presented in this paper are part of a wider project in which we aim to validate model predictions in free-living. Although free-living validations are critical, the resolution required to evaluate activity-specific errors can only be obtained from indirect calorimetry. Regarding activity categories, no gold standard method exists to validate time in sedentary activity, light physical activity, and MVPA outside of a controlled environment, and the generalizability of classification models to free-living studies is somewhat uncertain. The authors have highlighted the limitations of accelerometer data collected within a laboratory [53,54]; the activities performed in a free-living environment are more diverse, which further necessitates the need for more naturalistic (ie, free-living) validation studies or at least validation studies conducted over several days using diverse activity protocols in a residential facility. Next, to replicate predictions made by the present algorithms in free-living subjects, measured RMR may be required, which increases the researcher and participant burden. A suitable alternative in the absence of measured RMR would be prediction equations derived from BMI, age, height, and gender, rather than assuming a resting value of 3.5 ml O$_2$/kg/min [55,56]. Finally, our use of the measured RMR to calculate METs may contribute to differences between the tested algorithms and the SenseWear manufacturer.

Conclusions
This study builds on previous work from our lab and others, demonstrating that machine learning techniques can be used to learn the complexities of human movement and physiological data in the study of human energy expenditure. Classification and regression errors were greater when comparisons were made between studies. Single-sample, cross-sectional studies generating energy expenditure models show acceptable accuracy; however, it is likely that these models are overfitted to a given sample, and thus, improving generalizability is essential. To extend the utility of energy expenditure estimates beyond lab conditions, more cross testing between data sets is required, in addition to validation in free-living samples by doubly labeled water.

Acknowledgments
This research was funded by an Engineering and Physical Sciences Research Council impact acceleration grant (EP/R511717/1).

Authors’ Contributions
ROD, JT, MH, GF, CD, and RJS designed the study. ROD and JT collected the data. ROD, JT, and GWH analyzed the data. ROD, JT, MH, CD, GWH, GF, and RJS contributed to writing and reviewing the manuscript.

Conflicts of Interest
RJS consults for Slimming world UK through Consulting Leeds, which is a wholly owned subsidiary of the University of Leeds. The other authors declare no conflicts of interest.
Multimedia Appendix 1
Hyperparameters used in each of the models.
[DOCX File , 30 KB - mhealth_v9i8e23938_app1.docx ]

Multimedia Appendix 2
Leave-one-subject-out cross-validation results for each of the regression models in each of the activity categories.
[DOCX File , 32 KB - mhealth_v9i8e23938_app2.docx ]

Multimedia Appendix 3
Between-model comparisons for root mean square error in each of the tested activity types.
[DOCX File , 69 KB - mhealth_v9i8e23938_app3.docx ]

Multimedia Appendix 4
Permutation importance analysis for Fitbit, SenseWear, and Actigraph datasets.
[DOCX File , 124 KB - mhealth_v9i8e23938_app4.docx ]

Multimedia Appendix 5
Leave-one-subject-out cross-validation results for each of the classification models in each of the intensity categories.
[DOCX File , 22 KB - mhealth_v9i8e23938_app5.docx ]

References


47. Lu K, Yang L, Seoane F, Abtahi F, Forsman M, Lindecrantz K. Fusion of heart rate, respiration and motion measurements from a wearable sensor system to enhance energy expenditure estimation. Sensors 2018 Sep 14;18(9) [FREE Full text] [doi: 10.3390/s18093092] [Medline: 31697686]


Abbreviations

KNN: k-nearest neighbor

LOSO: leave-one-subject-out cross-validation

MAPE: mean absolute percentage error

MET: metabolic equivalent

MVPA: moderate-to-vigorous physical activity

RMR: resting metabolic rate

RMSE: root mean square error