How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies

Ruairi O’Driscoll, Jake Turicchi, Kristine Beaulieu, Sarah Scott, Jamie Matu, Kevin Deighton, Graham Finlayson, James Stubbs

ABSTRACT

Objective To determine the accuracy of wrist and arm-worn activity monitors’ estimates of energy expenditure (EE).

Data sources SportDISCUS (EBSCOHost), PubMed, MEDLINE (Ovid), PsycINFO (EBSCOHost), Embase (Ovid) and CINAHL (EBSCOHost).

Design A random effects meta-analysis was performed to evaluate the difference in EE estimates between activity monitors and criterion measurements. Moderator analyses were conducted to determine the benefit of additional sensors and to compare the accuracy of devices used for research purposes with commercially available devices.

Eligibility criteria We included studies validating EE estimates from wrist-worn or arm-worn activity monitors against criterion measures (indirect calorimetry, room calorimeters and doubly labelled water) in healthy adult populations.

Results 60 studies (104 effect sizes) were included in the meta-analysis. Devices showed variable accuracy depending on activity type. Large and significant heterogeneity was observed for many devices ($I^2$ >75%). Combining heart rate or heart sensing technology with accelerometry decreased the error in most activity types. Research-grade devices were statistically more accurate for comparisons of total EE but less accurate than commercial devices during ambulatory activity and sedentary tasks.

Conclusions EE estimates from wrist and arm-worn devices differ in accuracy depending on activity type. Addition of physiological sensors improves estimates of EE, and research-grade devices are superior for total EE. These data highlight the need to improve estimates of EE from wearable devices, and one way this can be achieved is with the addition of heart rate to accelerometry.

PROSPERO registration number CRD42018085016.

INTRODUCTION

The prevalence of obesity has tripled in the last 40 years, and it has been estimated that by 2050, 60% of males and 50% of females may be obese. Obesity is the result of a chronic imbalance between energy intake (EI) and energy expenditure (EE) driven by physiological, psychological and environmental factors. Doubly labelled water (DLW) is considered the gold standard for the measurement of free-living EE; however, the considerable costs and analytical requirements limit its feasibility in large cohort studies. Indirect calorimetry methods represent the most commonly employed criterion measure for assessment of the energy cost of an activity but again are limited to structured activities usually within a laboratory. Wearable activity monitors are increasingly popular for the estimation of EE.

Wearable devices that use triaxial accelerometry to derive an estimate of EE have been available for research purposes for some time. These devices are worn on the hip, thigh or lower back, as proximity to the centre of mass more accurately reflects the energy cost of movement; however, participant comfort and compliance is a recognised issue, and therefore traditional wear devices have limited long-term, free-living measurement capability. Use of wrist-worn activity monitors by both consumers and researchers has dramatically increased, facilitated by improved battery longevity and miniaturisation of hardware required to produce interpretable data. Recent consumer devices include triaxial accelerometers, heat sensors and photoplethysmography heart rate sensors. This information can be incorporated to improve the estimation of EE relative to accelerometry alone. However, their accuracy compared with criterion measures is questionable and may vary with the type and intensity of activity.

This meta-analysis aimed to investigate the accuracy of EE estimates from current wrist-worn or arm-worn devices during different activities. Given the recent popularity wrist and arm-worn activity monitors, it is critical to determine their validity for the estimation of EE. Secondary aims were to investigate the usefulness of specific sensors within devices and compare commercial and research-grade devices. We hypothesised that the addition of physiological data to accelerometry within wearable devices will provide a more accurate estimate of EE compared with criterion measures and that the performance of research-grade devices would be superior to commercial devices.

METHODS

This systematic review and meta-analysis adhered to Preferred Reporting Items for Systematic Reviews and Meta-Analyses diagnostic test accuracy guidelines (online supplementary material 1) and was prospectively registered in the PROSPERO database (CRD42018085016).

Search strategy SportDISCUS (EBSCOHost), PubMed, MEDLINE (Ovid), PsycINFO (EBSCOHost), Embase (Ovid)
Data were extracted to a specialised spreadsheet and entered into CMA. A second author (JT) verified 100% of extracted data and was cross-checked for data extraction errors. A third author (SS) screened 10% of the data. Studies reporting only accelerometer counts or studies involving post hoc manipulation of the device output were excluded.

**Inclusion criteria**

We considered laboratory or field validation studies conducted in healthy adults (≥18 years) comparing a criterion measure of EE to an estimate of EE in kilocalories, kilojoules or megajoules from an activity monitor. We considered only wrist-worn or arm-worn devices. There is a clear tendency towards wrist-worn devices among consumer devices, and devices worn on alternative anatomical locations produce different accelerometer patterns and therefore estimates of EE. For criterion validation, we considered DLW, indirect calorimetry systems and metabolic chambers.

**Exclusion criteria**

Adults with conditions deemed to produce atypical movement patterns were excluded, including Parkinson’s disease, chronic obstructive pulmonary disease, cerebral palsy and amputees. These conditions are often associated with abnormal gait pattern and thus reduce accuracy in EE estimates. Devices requiring external sensors or components were excluded. Studies reporting only accelerometer counts or studies involving post hoc manipulation of the device output were excluded.

**Study selection**

Two authors (RO and JT) independently assessed 100% of titles and abstracts for potential inclusion, with 10% screened independently by a third author (GF). In the case of disagreements between reviewers, the paper was retrieved in full text, and mutual consensus was reached. Remaining articles were screened independently for inclusion at the full-text level by two authors (RO and JT), with a third author (SS) screening 10%. Similarly, conflicts were resolved by discussion between reviewers.

**Data extraction**

From each of the included studies, characteristics of participants, validation protocol, criterion measure and the devices tested including model, wear site and output were extracted. Mean difference or EE estimates from the criterion measure and the device were extracted, along with SD, SE or 95% CIs. If only SE was provided, SE was converted to SD. If data were not provided, authors were contacted to request the raw data. Where values were only presented in figures, a digitiser tool was used. Data were extracted to a specialised spreadsheet and entered into Comprehensive Meta-analysis (CMA) (version 2; Biostat, Englewood, New Jersey, USA) for analysis. Data were extracted by one author (RO) and was cross-checked for data extraction errors. A second author (JT) verified 100% of extracted data and data entered into CMA.

**Quality assessment**

Risk of bias in included studies was determined using a modified version of the Downs and Black checklist for non-randomised studies. The Downs and Black instrument is an established tool for determination of the quality of a study within a systematic review and meta-analysis. The modified version used in the present study carried a maximum score of 18 and was quantified as: low (≤9 points, <50%), moderate (>9–14 points, 50%–79%) or high (≥15 points, ≥80%). It contained 17 questions, 10 related to reporting, 3 related to external validity and 4 related to internal validity. The risk of bias assessment was performed independently by two authors (ROD and JT), and disagreements were resolved by discussion.

**Statistical analysis**

Descriptive statistics were calculated for studies included within the meta-analysis. EE estimates from the device and criterion, SD or 95% CI, sample sizes and correlation coefficients for within-activity comparisons for each device were used to calculate effect size (ES). Correlation coefficients were based on raw data from previously published studies or were conservatively estimated based on the mean of similar devices (online supplementary material 3). Where a study provided data for more than one comparison for one device, the selected outcomes were pooled to provide a single mean and prevent overpowering of a single study. Hedges’ g and 95% CIs were calculated using CMA, in accordance with the majority of studies in the literature testing the mean bias between activity monitors and criterion measures. A negative ES represents an underestimate relative to the criterion, and a positive value represents an overestimation. Interpretation of ES was as follows: <0.20 as trivial, 0.20–0.39 as small, 0.40–0.59 as moderate and >0.80 as large. A random effects model was employed for all analyses based on the assumption that heterogeneity would exist between included studies due to the variability in study design. To determine heterogeneity, the I² statistic was used and >75% was considered to represent large heterogeneity. To determine susceptibility to bias from one study, a leave-one-out analysis was conducted where the removal of one study would leave at least three studies. The study associated with the greatest change to significance of the effect is reported. To assist interpretation of the error associated with each device, we calculated the percentage error for each device using percentage difference and weight within each meta-analysis.

**Exploration of small study effects**

To examine small study effects, data were visually inspected with funnel plots and subsequently quantified by using Egger’s linear regression intercept. A statistically significant Egger’s statistic indicates the presence of a small study effect.

**Moderators and subgroups**

As well as overall, which represents a combination of all subgroups, subgroup meta-analyses were performed for specific activities/categories: (1) activity energy expenditure (AEE), which included comparisons of EE estimates from the device to a criterion during non-specific exercise protocols, circuits, arm ergometer, rowing and resistance exercises; (2) ambulation and stair climbing; (3) cycling; (4) running; (5) sedentary behaviours and household tasks; and (6) total energy expenditure (TEE), representing comparisons with DLW.

We conducted moderator analyses by sensors, and all devices were grouped based on the inclusion of the following sensor hardware: (1) accelerometer alone (ACC); (2) heart rate (HR) alone; (3) accelerometer and heart rate (ACC+HR); (4) accelerometer and heat sensing or galvanic skin response (ACC+HS); and (5) accelerometer, heart rate sensors and heat...
RESULTS
Overview
A total of 64 studies were included in the systematic review (online supplementary material 4). Four studies could not be synthesised by meta-analysis as mean difference between activity monitors and criterion measurements were not provided; thus, 60 studies were included in the meta-analysis (figure 1). A total of 1946 participants were included, with a mean age of 35 years (range 20–86 years). The mean body mass index (BMI) was 24.9 kg/m² (range 21.8–31.6 kg/m²). Within the included studies, 104 comparisons between devices and a criterion were included. This represented 58 commercial and 46 research-grade device comparisons. ACC was composed of 35 comparisons, 1 in HR devices, 20 in ACC+HR devices, 45 in ACC+HS and 3 in ACC+HR+HS. With regard to activity performed, 35 comparisons were classed as AEE, ambulation and stairs included 55 comparisons, 23 were cycling tasks and 38 were running tasks. Sedentary and low-intensity was composed of 30 comparisons and TEE included 16 comparisons.

Devices
A total of 40 devices were tested in the included studies. One device was forearm-worn, 6 were worn on the upper arm (triceps) and 33 were wrist worn. Characteristics of the devices, number of studies and weighted percentage error for each device are shown in online supplementary material 5.

Meta-analysis
Individual study effect sizes and allocation to moderator variables are provided in online supplementary material 6. A minimum of three comparisons were required for meta-analysis and, as such, we report pooled ES for individual devices or moderators where three or more comparisons were

Figure 1 Flow diagram of study selection.
Figure 2  Pooled Hedges’ g and 95% CIs for estimates of energy expenditure relative to criterion measures per device over all activities. Total refers to number of effect sizes. A negative Hedges’ g statistic represents an underestimation, and a positive Hedges’ g represents an overestimation. ACT, Actical; AGT3X, Actigraph GT3X; AW, Apple watch; AW52, Apple Watch series 2; BA, Beurer AS80; BMC, Bodymedia CORE armband; BP, Basis Peak; EP, Epson Pulsense; EPUL, ePulse Personal Fitness Assistant; FB, Fitbit Blaze; FC, Fitbit Charge; FC2, Fitbit Charge 2; FCHR, Fitbit Charge HR; FF, Fitbit Flex; GF225, Garmin Forerunner 225; GF920XT, Garmin Forerunner 920XT; GVA, Garmin Vivoactive; GVF, Garmin Vivofit; GVS, Garmin vivosmart; LC, LifeChek calorie sensor; MA, Mio Alpha; MB, Microsoft band; MS, Misfit Shine; NF, Nike FuelBand; PL, Polar Loop; Polar AW200, Polar: AW200; PA360, Polar: AW360; SG, Samsung Gear S; SWA, SenseWear Armband; SWA p2, SenseWear Armband Pro 2; SWA p3, SenseWear Armband Pro 3; SWAM, SenseWear Armband Mini; TT, TOMTOM Touch; V, Vivago; WP, Withings Pulse; WPO, Withings Pulse O2.

Meta Analysis Overall

Available. Statistical outputs for each device are presented in online supplementary material 7.

Quality assessment

The modified Downs and Black scores revealed a median score of 13, with one study being classed as low quality,49 48 classed as moderate and 11 classed as high quality (online supplementary material 8). The questions included in the modified tool and percentage of studies fulfilling each question are shown in online supplementary material 9.

Overall

A forest plot of individual devices over all activities is shown in figure 2. Overall, devices underestimated EE (ES: −0.23, 95% CI −0.44 to −0.03; n=104; p=0.03) and showed significant heterogeneity between devices (I²=92.18%; p<0.001). Significant underestimations relative to criterion measures were observed for the Garmin Vivofit (GVF; ES: −1.09, 95% CI −1.60 to −0.57; n=5; p<0.001) and the Jawbone UP24 (ES: −1.16, 95% CI −1.78 to −0.54; n=3; p<0.001). The SenseWear Armband Pro3 (SWA p3) also underestimated EE (ES: −0.32, 95% CI −0.62 to −0.01; n=12; p=0.04). Sensitivity analysis revealed that the removal of six comparisons altered the significance of the SWA p3 (p>0.05), the most influential of which decreased the ES to −0.19 (95% CI −0.50 to 0.11; p=0.21). The Apple Watch, Bodymedia CORE armband (BMC), Fitbit Charge HR (FCHR), Fitbit Flex (FF), Jawbone UP (JU), Nike FuelBand (NF), SenseWear Armband (SWA), SenseWear Armband Pro2 (SWA p2) and SenseWear Armband Mini (SWAM) did not differ significantly from criterion measures. However, sensitivity analysis showed the FCHR differed significantly with the removal of one study (ES: 0.34, 95% CI 0.20 to 0.56).
Figure 3  Pooling Hedges’ g and 95% CIs for estimates of energy expenditure relative to criterion measures per device for ambulation and stair climbing. Total refers to number of effect sizes. A negative Hedges’ g statistic represents an underestimation, and a positive Hedges’ g represents an overestimation. AGT3X, Actigraph GT3X; AW, Apple Watch; BA, Beurer AS80; BMC, Bodymedia CORE armband; BP, Basis Peak; EPUL, ePulse Personal Fitness Assistant; FC, Fitbit Charge; FCHR, Fitbit Charge HR; FF, Fitbit Flex; GF225, Garmin Forerunner 225; GF920XT, Garmin Forerunner 920XT; GVA, Garmin Vivoactive; GVF, Garmin Vivofit; GVS, Garmin vivosmart; JU, Jawbone UP; JU24, Jawbone UP24; MB, Microsoft band; NF, Nike FuelBand; PL, Polar Loop; Polar AW200, Polar: AW200; SWA, SenseWear Armband; SWA p2, SenseWear Armband Pro 2; SWA p3, SenseWear Armband Pro 3; SWAM, SenseWear Armband Mini; V, Vivago; WP, Withings Pulse; WPO, Withings Pulse O2.

Activity energy expenditure
A forest plot of individual devices during activities classed as AEE is shown in online supplementary material 10. For AEE, the pooled estimate of all devices was a non-significant tendency to underestimate EE compared with criterion measures (ES: −0.34, 95%CI −0.71 to 0.04; n=35; p=0.08) and significant heterogeneity was observed between devices (I²=76.34%; p<0.001). The SWA p2 underestimated EE (ES: −0.78, 95% CI −1.48 to −0.08; p=0.03) and had moderate, non-significant heterogeneity (I²=64.19%; p=0.06). The BMC, NF, SWA and SWAM did not differ significantly from criterion measures but all displayed significant heterogeneity. No device showed evidence of small study effects.

Ambulation and stairs
A forest plot of individual devices during ambulation and stair climbing is shown in figure 3. The pooled estimate of all devices did not differ from criterion measures (ES: −0.09, 95% CI −0.45 to 0.27; p=0.01). The FCHR (ES: 0.78, 95% CI 0.28 to 1.29; n=5; p=0.002) and FF (ES: 1.10, 95% CI 0.43 to 1.77; n=3; p=0.001) underestimated EE. The GVF underestimated EE (ES: −1.24, 95% CI −1.86 to −0.62; n=4; p<0.001), however, sensitivity analysis revealed that the removal of two comparisons significantly altered the mean effect (p>0.05) the most influential study significantly altered the mean effect to ES: −1.32 (95% CI −2.73 to 0.08; p=0.07). Furthermore, there was evidence of small study effects (intercept=−13.76, 95% CI −19.72 to −7.80; p=0.01). The SWA overestimated EE (ES: 0.79, 95% CI 0.25 to 1.33; n=3; p=0.004) and sensitivity analysis revealed that the removal of four comparisons significantly altered the mean effect (p>0.05) the most influential significantly altered the mean effect to ES: 0.33 (95% CI −0.26 to 0.92; p=0.28).36 The AW, JU, SWA p3 and SWAM did not differ significantly from criterion measures. The mean effect of the SWAM was significantly altered by the removal of two studies; the removal of the most influential study yielded a significant overestimation (ES: 0.57, 95% CI 0.20 to 0.94; p=0.003). All devices showed significant heterogeneity.

Cycling
A forest plot of individual devices during cycling is shown in online supplementary material 10. The pooled estimate of all devices was significantly lower than criterion measures (ES: −0.73, 95% CI −1.39 to −0.06; n=23; p=0.03) and significant.
heterogeneity was observed between devices ($I^2=94.74\%; p<0.01$). The SWA did not differ significantly from criterion but showed significant heterogeneity ($I^2=89.39\%; p<0.001$). The SWA p3 did not differ from criterion measures and showed moderate heterogeneity ($I^2=54.95\%; p=0.11$).

Running
A forest plot of individual devices during running is shown in [online supplementary material 10](http://bjsm.bmj.com/). The pooled estimate was not statistically different from criterion measures ($ES: -0.08, 95\%CI -0.41 to 0.25; n=38; p=0.63$) and significant heterogeneity was observed between devices ($I^2=92.05\%; p<0.001$). The FCHR, GVF and SWA did not differ from criterion measures. Sensitivity analysis revealed the removal of one study changed the overall effect for the FCHR ($ES: 0.59, 95\%CI 0.28 to 0.90; p<0.001$). Significant heterogeneity was observed for the FCHR ($I^2=66.8\%; p=0.03$) and SWA ($I^2=96.79\%; p<0.001$) but not for the GVF ($I^2=46.39\%; p=0.15$).

Sedentary and household tasks
A forest plot of individual devices during sedentary and household tasks is shown in [figure 4](http://bjsm.bmj.com/). The pooled effect was not statistically different from criterion measures ($ES: -0.09, 95\%CI -0.51 to 0.32; n=38; p=0.66$) and significant heterogeneity was observed between devices ($I^2=94.84\%; p<0.001$). The AW, FCHR and SWAM were not statistically different from criterion measures. Sensitivity analysis revealed that the removal of three studies changed the mean effect, the most influential of which decreased the ES to $-0.01$ (95\%CI $-0.82$ to $-0.01$; $p=0.05$). Observed heterogeneity was significant for the AW, SWA p3 and SWAM. The FCHR had moderate, non-significant heterogeneity ($I^2=59.60\%; p=0.06$).

Meta Analysis Sedentary and Household

Figure 4 Pooled Hedges’ g and 95% CI for estimates of energy expenditure relative to criterion measures per device for sedentary and household tasks. Total refers to number of effect sizes. A negative Hedges’ g statistic represents an underestimation, and a positive Hedges’ g represents an overestimation. AW, Apple Watch; BMC, Bodymedia CORE armband; BP, Basis Peak EPUL, ePulse Personal Fitness Assistant; FCHR, Fitbit Charge HR; FF, Fitbit Flex; GF225, Garmin Forerunner 225; GVF, Garmin Vivofit; JU, Jawbone UP; JU24, Jawbone UP24; MB, Microsoft band; SWA p2, SenseWear Armband Pro 2; SWA p3; SenseWear Armband Pro 3; SWAM, SenseWear Armband Mini; V, Vivago; WP, Withings Pulse.

Total energy expenditure
A forest plot of individual devices for the measurement of TEE is shown in [figure 5](http://bjsm.bmj.com/). The pooled effect for TEE showed a significant underestimation of EE ($ES: -0.68, 95\%CI -1.15 to -0.21; n=16; p=0.005$), and significant heterogeneity was observed between devices ($I^2=92.71\%; p<0.001$). The SWA p3 did not differ significantly from criterion measures and showed significant heterogeneity ($I^2=94.20\%; p=0.001$).

Moderator analyses
The results of moderator analyses are shown in [table 1](http://bjsm.bmj.com/). Overall, there was a significant difference between sensors ($p=0.003$). Pooled estimate of EE from ACC+HR and ACC+HS was not statistically different from criterion, but ACC+HS showed a non-significant tendency for underestimation, and ACC and ACC+HR+HS both significantly underestimated EE. In the AEE comparison, there was no statistical difference between sensors, but ACC+HS significantly underestimated EE, ACC showed a non-significant tendency for underestimation and ACC+HR did not differ significantly from criterion measures. During ambulation and stair climbing, a significant difference between sensors was observed, with estimates of EE from ACC+HR and ACC+HS being significantly higher than criterion. In cycling, significant differences were observed between sensors, with ACC devices underestimating EE. During running activities, none of the pooled mean estimates were significantly different from criterion. For sedentary and household tasks, a significant difference was observed between sensors; ACC+HR was not different from criterion measures, whereas ACC and ACC+HS underestimated and overestimated EE, respectively. For TEE, sensors differed significantly; ACC underestimated EE, whereas ACC+HS did not differ significantly from criterion.
When analysed by commercial and research-grade devices, no significant difference was observed overall, for AEE, cycling or running. For both the ambulation and stairs comparison and the sedentary and household tasks comparison, commercial devices were closer to criterion measurements, with research-grade devices significantly overestimating. For TEE, research-grade devices were superior, with commercial devices significantly underestimating EE.

**DISCUSSION**

Given the clinical and consumer uptake of wrist-worn and arm-worn activity monitors that can be used for the estimation of EE, this meta-analysis aims to: (i) determine the relative accuracy of current devices; (ii) investigate the importance of specific sensors within devices; and (iii) compare commercial and research-grade devices.

For devices with sufficient comparisons to be analysed separately from the main pooled effect, significant error relative to criterion measures was observed for Garmin, Fitbit, Jawbone and Bodymedia products. Garmin, Fitbit and Jawbone represent a major share of the commercial wearable market, and Bodymedia products are widely used in research and have been since 2004. While it is initially encouraging that the ES for Bodymedia products is widely used in research and have been since 2004.59 While it is initially encouraging that the ES for Bodymedia products.

<table>
<thead>
<tr>
<th>Device</th>
<th>Total</th>
<th>Lower Limit</th>
<th>Upper Limit</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>EP</td>
<td>1</td>
<td>-0.68</td>
<td>-1.58</td>
<td>0.21</td>
</tr>
<tr>
<td>FF</td>
<td>1</td>
<td>-0.48</td>
<td>-1.35</td>
<td>0.38</td>
</tr>
<tr>
<td>GVF</td>
<td>1</td>
<td>-1.62</td>
<td>-2.63</td>
<td>-0.62</td>
</tr>
<tr>
<td>JU24</td>
<td>1</td>
<td>-1.96</td>
<td>-2.97</td>
<td>-0.95</td>
</tr>
<tr>
<td>MS</td>
<td>1</td>
<td>-0.88</td>
<td>-1.85</td>
<td>-0.29</td>
</tr>
<tr>
<td>SWA</td>
<td>2</td>
<td>0.14</td>
<td>-0.46</td>
<td>0.75</td>
</tr>
<tr>
<td>SWA p2</td>
<td>1</td>
<td>-0.23</td>
<td>-1.07</td>
<td>0.60</td>
</tr>
<tr>
<td>SWA p3</td>
<td>5</td>
<td>-0.25</td>
<td>-0.64</td>
<td>0.13</td>
</tr>
<tr>
<td>SWAM</td>
<td>2</td>
<td>-0.04</td>
<td>-0.64</td>
<td>0.56</td>
</tr>
<tr>
<td>WPO</td>
<td>1</td>
<td>-1.68</td>
<td>-2.62</td>
<td>-0.75</td>
</tr>
<tr>
<td>Overall</td>
<td>16</td>
<td>-0.68</td>
<td>-1.15</td>
<td>-0.21</td>
</tr>
</tbody>
</table>

An accurate yet affordable measure of TEE, with a measure of change in energy storage, could theoretically be used to retrospectively determine free-living EI in large cohorts.89 In this context, TEE may be considered the most important activity subgroup in this meta-analysis; however, the most variable and unpredictable component of TEE is EE during activity.6 In agreement with previous studies, we have shown that the accuracy of devices differs by activity and this may be related to the inability of devices to differentiate between activity types. For a device to accurately estimate TEE between individuals, it must accurately estimate the energy cost of a wide range of activities; however, some activities may require greater focus. The majority of EE is attributable to rest or non-exercise activity, so error here could have a great impact on the error in TEE. The FCHR was the most tested commercial device in this analysis, and it showed a trivial, non-significant ES overall and during sedentary tasks but a moderate to large and significant overestimation during ambulatory activity. Considering that ambulatory activity is central to public health guidelines worldwide, the implications of this finding may be great for estimates of TEE.

The observed error for different activity types may be because current algorithms do not take physical activity type or bodily posture into account.91 Indeed, activity recognition is considered an important direction for wearable technology and has been used to improve estimates of EE.92 Montoye et al93 have shown that accelerometers worn on the wrists and thigh can be used to predict activity type. The SenseWear software employs complex pattern-recognition algorithms to determine activity type, which likely contributed to the trivial or small ES observed for the SenseWear Armband Mini in all comparisons. The challenges associated with activity recognition have been reviewed recently, and as this technology develops, activity-specific EE prediction equations may offer the opportunity to reduce errors associated with activity types.
Recent technological advancements, triaxial accelerometry, as well as heart rate or heat-sensing devices improve estimates of EE, relative to uniaxial devices.21 A 2012 review concluded that multisensory and triaxial accelerometers improve estimates of EE, relative to uniaxial devices.21 We hypothesised that the addition of this technology to accelerometry would improve estimates of EE. Overall, this meta-analysis shows that the inclusion of heart rate or heat sensors in devices can improve estimates of EE relative to accelerometry alone. Indeed, it is established that accelerometry is limited for non-weight-bearing activities,84 and accelerometry underestimated EE during cycling activities in our analysis. Significant underestimations were also observed during sedentary and household tasks and TEE, which is likely a product of the limited arm movements associated with these activities.

Accelerometry and heart rate devices moderately overestimated EE during ambulation and stair climbing. Some of this error may be attributable to the individual variability in the relationship between heart rate and EE. Individual calibration of this relationship in the Actiheart device is associated with improved estimates of EE54 and may offer a means for further reducing the error observed in wrist-worn and arm-worn devices. An alternative explanation for this is the variability in estimates of heart rate from photoplethysmography heart rate sensors. A recent study reported a small mean error of −5.9 bpm in the Fitbit Charge 2 but wide limits of agreement of −28.5 to 16.8 bpm,96 and this variability is a common finding.35 40

Device grade

The third aim of this meta-analysis was to compare commercial and research-grade devices. Commercial devices may be developed with affordability and comfort as a primary focus, and as a consequence, it may be unreasonable to expect commercial devices to match the validity of research-grade devices. Recent consumer monitors share similar technology with established research-grade multisensor devices,48 and this is partially reflected in our results. A benefit of research-grade devices for TEE was observed, but commercial devices were statistically superior in ambulation and during sedentary tasks. Our results question the use of wrist-worn or arm-worn research-grade devices for the validation of newer devices. Comparisons with criterion measures such as DLW or indirect calorimetry are more appropriate when absolute accuracy is required.6 Furthermore, it is important to highlight that other research-grade devices, for instance the Actiheart, which is worn on the chest,65 are likely to be more accurate than research-grade devices included in this study.6 Further research is needed to establish whether research-grade devices that are worn in other locations such as the chest, hip or thigh outperform consumer-based devices.

Limitations

Separate pooled analyses to determine the accuracy of individual activity monitors were performed for a limited number of devices due to the small number of comparisons available for the remaining devices (i.e., less than three comparisons). This limitation is inevitable considering the large number of activity monitors included in this review. Nevertheless, the inclusion of all devices in the overall pooled analysis provides an extensive and robust evaluation of the difference in EE outcomes between activity monitors and criterion measures.

The majority of analyses conducted within this review demonstrated large heterogeneity within and between devices that remained after moderating by specific devices and activity. Such heterogeneity is not unexpected and in many cases may be attributable to disparity in the protocols employed.95 Indirect calorimetry systems were the most commonly used criterion measure, but EE estimates may differ by up to 5.2% depending on the equations used.96 EE is likely to be elevated in the period following higher intensity exercise, and the inclusion of only the steady state period may influence the extent to which results are generalizable to the broader population.

Sensors

A 2012 review concluded that multisensory and triaxial accelerometry devices improve estimates of EE, relative to uniaxial devices.81 Due to recent technological advancements, triaxial accelerometry, as well as heat rate or heat-sensing technology is commonplace in newer devices.48 We hypothesised that the addition of this technology to accelerometry would improve estimates of EE. Overall,
which devices differ from criterion measures. There is also the possibility that the discrepancy between device estimates relates to populations studied for example, a higher BMI or age-related changes in movement patterns. As few devices currently provide open access to EE algorithms, the potential for this to create heterogeneity remains uncertain. Despite this, the statistically significant outcomes in many cases suggest a consistent direction in effect sizes for many comparisons, and the differences in statistical outcomes between devices are supported by the magnitude of effect sizes.

External validity was low in 46 studies pooled in this meta-analysis, which must be considered when interpreting the present results. It must also be noted that the present analysis was limited to healthy individuals, and therefore, our results cannot be generalised to populations with conditions that produce abnormal gait patterns.

Lastly, there is a lag between product release and testing in research environments, and some of the devices included in this meta-analysis are no longer in production so the continued validation of newer devices is imperative.

**CONCLUSION**

This meta-analysis collated studies evaluating the validity of EE estimates by wrist-worn or arm-worn devices. Devices vary in accuracy depending on activity type and the significant heterogeneity means caution must be exercised when interpreting these results. Devices with heart rate sensors often produced better estimates than devices using accelerometry only; however, this was not consistent across all activities. Wrist-worn and arm-worn research-grade devices were more accurate than commercial devices for estimates of TEE, but researchers should be aware that such devices do not guarantee superior accuracy. Future research should aim to understand and reduce the error in EE estimates from wrist-worn or arm-worn devices in different activity types. This may be achieved through activity recognition techniques, incorporating physiological measures and exploring the potential for individual calibration of these relationships.

**What is already known**

- Wrist-worn or arm-worn devices incorporating multiple sensors are increasingly common, and many devices provide estimates of energy expenditure. It is important to determine their validity overall and in different activity types.
- It is not clear which specific sensors or combinations of sensors provide the most accurate estimates of energy expenditure.
- It is unclear whether research-grade devices are more accurate than commercial devices.

**What are the new findings**

- The accuracy in energy expenditure estimates from activity monitors varies between activities.
- Larger error is observed from devices employing accelerometry alone; the addition of heart rate sensing improves estimates of energy expenditure in most activities.
- In some activity types, research-grade devices are not superior to commercial devices.

**REFERENCES**


Review


