How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis.

Ruairi O'Driscoll,[1] Jake Turicchi,[1] Kristine Beaulieu,[1] Sarah Scott,[1] Jamie Matu,[2] Kevin Deighton,[3] Graham Finlayson,[1] R. James Stubbs[1]

[1]Appetite Control and Energy Balance Group, School of Psychology, University of Leeds, Leeds, U.K.

[2]Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Leeds, U.K.

[3]Institute for Sport, Physical Activity & Leisure, Leeds Beckett University, Leeds, U.K.

Corresponding author:
Ruairi O'Driscoll
Appetite Control and Energy Balance Group
University of Leeds,
Leeds, U.K.
LS2 9JT
Psrod@leeds.ac.uk

Word count:
4493

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 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 heat 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.

Registration PROSPERO CRD42018085016.

Keywords*: Energy expenditure, Accelerometer, Meta-analysis, Wrist, Validation.*

Device abbreviations*: Actical (ACT), Actigraph GT3X (AGT3X), Apple watch (AW), Apple Watch series 2 (AWS2), Beurer (BA) Basis b1 (BB1), Bodymedia CORE armband (BMC), Basis Peak (BP), Epson Pulsense (EP), ePulse Personal Fitness Assistant (EPUL), Fitbit Blaze (FB), Fitbit Charge (FC), Fitbit Charge 2 (FC2), Fitbit Charge HR (FCHR), Fitbit Flex (FF), Garmin Forerunner 225 (GF225), Garmin Forerunner 920XT (GF920XT), Garmin Vivoactive (GVA), Garmin Vivofit (GVF), Garmin Vivosmart (GVS), Garmin Vivosmart HR (GVHR), Jawbone UP (JU), Jawbone UP24 (JU24), LifeChek calorie sensor (LC), Mio Alpha (MA), Microsoft band (MB), Misfit Shine (MS), Polar: AW360 (PA360), Nike Fuel band (NF), Polar Loop (PL), Polar: AW200 (PO200), Samsung Gear S (SG), SenseWear Armband (SWA), SenseWear Armband Pro 2 (SWA p2), SenseWear Armband Pro 3 (SWA p3), SenseWear Armband MINI (SWAM), TOMTOM Touch (TT), Vivago (V), Withings Pulse (WP), Withings Pulse O2 (WPO).*

What is already known on this topic?

- Wrist 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 this study adds

- 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.

Introduction

The prevalence of obesity has tripled in the last 40 years [1] and it has been estimated that by 2050, 60% of males and 50% of females may be obese [2]. Obesity is the result of a chronic imbalance between energy intake (EI) and energy expenditure (EE) [3] driven by physiological, psychological and environmental factors.

Doubly-labelled water (DLW) is considered the gold standard for the measurement of free-living EE [4]; however, the considerable costs and analytical requirements limit its feasibility in large cohort studies [5]. 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 [6]. Wearable activity monitors are increasingly popular for the estimation of EE [7].

Wearable devices which use triaxial accelerometry to derive an estimate of EE have been available for research purposes for some time [8]. 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 [9]; however, participant comfort and compliance is a recognised issue [10] 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 [11] facilitated by improved battery longevity and miniaturization of hardware required to produce interpretable data [12]. Recent consumer devices include triaxial accelerometers, heat sensors and photoplethysmography heart rate sensors [13]. This information can be incorporated to improve the estimation of EE relative to accelerometry alone [14]. However, their accuracy compared with criterion measures is questionable [15] and may vary with the type and intensity of activity [16].

This meta-analysis aimed to investigate the accuracy of EE estimates from current wrist 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 [17]. 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 [18], 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 PRISMA diagnostic test accuracy guideline [19] (supplementary material 1) and was prospectively registered in the PROSPERO database (CRD42018085016).

Search strategy

SportDISCUS (EBSCOHost), PubMed, Medline (Ovid), PsycINFO (EBSCOHost), EMBASE (Ovid) and CINAHL (EBSCOHost) were searched for studies published up to 1st December 2017 using terms relevant to the validation of EE estimates from activity monitors against criterion measures with the following strategy ((tracker AND EE) AND validation). The search was updated 15th January 2018. The specific keywords and the full search strategy can be found in supplementary material 2. No language restrictions were applied and in the case of studies available only as an abstract, attempts were made to contact the authors.

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 (kcal), kilojoules (kJ) or megajoules (MJ) from an activity monitor. We considered only wrist or arm-worn devices. There is a clear tendency towards wrist worn devices amongst consumer devices and devices worn on alternative anatomical locations produce different accelerometry patterns and therefore estimates of EE [20]. For criterion validation, we considered DLW, indirect calorimetry devices and metabolic chambers [6].

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 [21]. 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 (ROD 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 (ROD 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 standard deviation (SD), standard error (SE) or 95% confidence intervals (95% CI). 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 [22]. Data was extracted to a specialised spreadsheet and entered into Comprehensive Meta-analysis (CMA) (CMA, version 2; Biostat, Englewood, NJ) for analysis. Data was extracted by one author (ROD) 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 [23]. The Downs and Black instrument is an established tool for determination of the quality of a study within a systematic review and meta-analysis [24]. The modified version used in the present study carried a maximum score of 18 and was quantified as: low (≤9, <50%), moderate (>9–14 points, 50–79%), or high (≥15 points, ≥80%) [25]. It contained 17 questions, 10 related to reporting, three to external validity and four to internal validity. The risk of bias assessment was performed independently by two authors (ROD and JT), 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 sizes. Correlation coefficients were based on raw data from previously published studies or were conservatively estimated based on the mean of similar devices (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 (ES) [26] 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 underestimation 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, 0.40-0.80 as moderate and >0.80 as large [27]. 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 [28]. To determine heterogeneity, the I^2 statistic [29] was utilised 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 relative 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 [30]. 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 to DLW.

We conducted moderator analyses by sensors and all devices were grouped based on the inclusion of the following sensor hardware: 1) accelerometry alone (ACC); 2) heart rate alone (HR); 3) accelerometry and heart rate (ACC+HR); 4) accelerometry and heat sensing or galvanic skin response (ACC+HS) and 5) accelerometry, heart rate sensors and heat sensing or galvanic skin response sensors (ACC+HR+HS). Secondly, moderator analyses were conducted by commercial and research-grade devices. Devices produced by Actical,

Actigraph and Bodymedia were considered as research-grade and all other devices included in the analysis were considered commercial devices. Comparisons between each moderator employed a random effects model.

Results

Overview

A total of 64 studies were included in the systematic review (Supplementary 4). Four studies could not be synthesised by meta-analysis as mean difference between activity monitors and criterion measurements were not provided [12,31–33]; thus, 60 studies were included in the meta-analysis (figure 1) [10,13,41–50,20,51–60,34,61–70,35,71–80,36,81–88,37–40]. A total of 1946 participants were included, with a mean age of 35 years (range 20 to 86 years). The mean BMI was 24.9 kg/m^2 (range 21.8 to 31.6 kg/m^2). 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 comprised 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 comprised 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 is shown in supplementary materials 5.

Meta-analysis

Individual study effect sizes and allocation to moderator variables are provided in supplementary materials 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 available. Statistical outputs for each device are presented in supplementary materials 7.

Quality assessment

The modified Downs and Black scores revealed a median score of 13, with one study being classed as low quality [69], 48 classed as moderate and 11 classed as high quality (supplementary materials 8). The questions included in the modified tool and percentage of studies fulfilling each question is shown in supplementary materials 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 -0.04; n=104; p=0.03) and showed significant heterogeneity between devices (I^2 =92.18%; p=<0.001). Significant underestimations relative to criterion measures were observed for the Garmin Vivofit (GVF; ES: -1.09, 95% CI: -1.61 to -0.56; n=5; p<0.001) and the Jawbone UP24 (JU24; ES: -1.16, 95% CI: -1.79 to -0.53; 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) [81]. The Apple watch (AW) 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 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 0.49; p<0.001) [88]. The NF was the only device that did not display significant heterogeneity between studies (I^2 =25.44%; p=0.26), with the remaining devices having I^2 values ≥ 66.91% (all p≤0.05). No device showed evidence of small study effects.

AEE

A forest plot of individual devices during activities classed as AEE is shown in supplementary materials 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 (I^2 =94.94%; p<0.001). The SWA p2 underestimated EE (ES: -0.78, 95% CI: -1.48 to -0.08; n=3; p=0.03) and had moderate, non-significant heterogeneity (I^2 =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; n=55; p=0.62) and significant heterogeneity was observed between devices (I^2 =93.74%; p<0.01). The FCHR (ES: 0.78, 95% CI: 0.27 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) overestimated EE. The GVF underestimated EE (ES: -1.24, 95% CI: -1.86 to -0.62; n=4; p<0.01), however, sensitivity analysis revealed that the removal of two comparisons significantly altered the mean effect (p>0.05) the most influential significantly altered the mean effect to ES: -1.32 (95% CI: -2.73 to 0.08; p=0.07) [34]. Further, 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=5; p<0.01) 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) [56]. 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) [87]. All devices showed significant heterogeneity.

Cycling

A forest plot of individual devices during cycling is shown in supplementary materials 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 supplementary materials 10. The pooled estimate was not statistically different from criterion measures (ES: -0.08, 95% CI: -0.41 to 0.25; n=38; p=0.65) 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) [87]. 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. The pooled effect was not statistically different from criterion measures (ES: -0.09, 95% CI: -0.51 to 0.32; n=30; 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. The SWA p3 overestimated EE (ES: 0.67, 95% CI: 0.00 to 1.34; p=0.049). Sensitivity analysis revealed that the removal of three studies changed the mean effect, the most influential of which decreased the ES to 0.41 (95% CI: -0.01 to 0.82; p=0.05) [42]. Observed heterogeneity was significant for the AW, SWA p3 and SWAM. The FCHR had moderate, non-significant heterogeneity (I^2 =59.60%; p=0.60).

TEE

A forest plot of individual devices for the measurement of TEE is shown in figure 5. The pooled effect for TEE showed a significant underestimation of EE (ES: -0.68, 95% CI: -1.15 to -0.21; n=16; p= p=0.005) and significant heterogeneity was observed between devices (I^2 =92.17%; p<0.01). 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. 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.

Moderator variable	Subgroup level	*p*-value	Hedges' g (95% CI)
Overall activities			
Sensors	ACC (n=35)	<0.01	-0.36 (-0.55, -0.17)*
	ACC + HR (n=20)		0.06 (-0.18, 0.31)
	ACC + HR + HS (n=3)		-0.99 (-1.65, -0.33)*
	ACC + HS (n=45)		-0.151 (-0.32, 0.01)
Device grade	Commercial (n=58)	0.27	-0.269(-0.42, -0.12)*
	Research (n=46)		-0.141 (-0.31, 0.03)
AEE			
Sensors	ACC (n=8)	0.19	-0.40 (-0.84, 0.04)
	ACC + HR (n=9)		-0.04 (-0.47, 0.38)
	ACC + HS (n=16)		-0.32 (-0.63, -0.01)*
Device grade	Commercial (n=18)	0.62	-0.38 (-0.67, -0.08)*
	Research (n=17)		-0.27 (-0.57, 0.04)
Ambulation and stairs			
Sensors	ACC (n=24)	0.01	-0.23 (-0.51, 0.06)
	ACC + HR (n=10)		0.45 (0.02, 0.87)*
	ACC + HS (n=19)		0.40 (0.08, 0.72)*
Device grade	Commercial (n=35)	0.05	-0.04 (-0.28, 0.20)
	Research (n=20)		0.37 (0.05, 0.68)*
Cycling			
Sensors	ACC (n=3)	<0.01	-3.75 (-4.65, -2.85)*
	ACC + HR (n=9)		-0.04 (-0.47, 0.40)
	ACC + HS (n=9)		-0.41 (-0.84, 0.02)
Device grade	Commercial (n=14)	0.28	-0.82 (-1.30, -0.35)*
	Research (n=9)		-0.41 (-0.99, 0.17)
Running			
Sensors	ACC (n=19)	0.18	-0.06 (-0.364, 0.24)
	ACC + HR (n=7)		0.34 (-0.15, 0.82)
	ACC + HS (n=10)		-0.36 (-0.78, 0.05)
Device grade	Commercial (n=28)	0.08	0.06 (-0.18, 0.30)
	Research (n=10)		-0.36 (-0.76, 0.04)
Sedentary and household			
Sensors	ACC (n=6)	<0.01	-0.65 (-1.16, -0.13)*
	ACC + HR (n=9)		0.14 (-0.28, 0.57)
	ACC + HS (n=13)		0.41 (0.06, 0.75)*
Device grade	Commercial (n=17)	<0.01	-0.27 (-0.59, 0.05)
	Research (n=13)		0.41 (0.05, 0.77)*
TEE (DLW)			
Sensors	ACC (n=5)	<0.01	-1.24(-1.66, -0.81)*
	ACC + HS (n=10)		-0.13(-0.397, 0.32)
Device grade	Commercial (n=6)	<0.01	-1.13(-1.51, -0.76)*
	Research (n=10)		-0.13 (-0.39, 0.14)

Discussion

Given the clinical and consumer uptake of wrist and arm-worn activity monitors which can be used for the estimation of EE, the aims of this meta-analysis were (i) to determine the relative accuracy of current devices, (ii) to investigate the importance of specific sensors within devices and (iii) to 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 [73] and Bodymedia products are widely used in research and have been since 2004 [59]. Whilst it is initially encouraging that the ES for many devices was not significantly different from criterion, the 95% CI observed in many cases indicates the potential for these devices to produce erroneous estimates of mean EE and as such we would be hesitant to consider any device sufficiently accurate. A 10% 'equivalence zone' has been suggested previously [65] and with the exception of the Nike Fuel band, in which all three studies reported a mean error <10% [65,79,82], no device pooled in this meta-analysis consistently met this criteria. The SenseWear armband Mini was the most accurate device overall but error reported in studies ranged from -21.27% [87] to 14.76% [39]. Studies in this analysis followed the manufacturer's instructions for setup, with researchers ensuring the position of the device and characteristics such as height, weight, sex and age were correct. In free-living environments the lack of researcher presence could yield greater error than observed in this analysis [17], as indicated by the moderate, significant underestimation for the pooled effect in the TEE subgroup.

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 [13,45,52], 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 [6] so error here could have a great impact on the error in TEE. The Fitbit Charge HR 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 [90], 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 [11] and has been used to improve estimates of EE [92]. Montoye et al have shown that accelerometers worn on the wrists and thigh can be used to predict activity type [93]. The SenseWear software employs complex pattern-recognition algorithms to determine activity type [45] 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 [94] and as this technology develops, activity-specific EE prediction equations may offer the opportunity to reduced errors associated with activity types.

Sensors

A 2012 review concluded that multisensory and triaxial accelerometry devices improve estimates of EE, relative to uniaxial devices [21]. Due to recent technological advancements, triaxial accelerometry, as well as heart rate or heat sensing technology are commonplace in newer devices [48]. 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 EE [95] and may offer a means for further reducing the error observed in wrist 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 multi-sensor 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 or arm-worn research-grade devices for the validation of newer devices. Comparisons to criterion measures such as DLW or indirect calorimetry are more appropriate when absolute accuracy is required [6]. Further, it is important to highlight that other research-grade devices, for instance the Actiheart, which is worn on the chest [95], are likely to be more accurate than research-grade devices included in this study [48]. 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 which 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 [97]. 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 [98]. 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

devices differ from criterion measures [56]. There is also the possibility that the discrepancy between device estimates relates to populations studied [16] for example, a higher BMI [35,40] or age related changes in movement patterns [69]. 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 suggests 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 generalized to populations with conditions that produce abnormal gait patterns.

Lastly, there is a lag between product release and testing in research environments [40] 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 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 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 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.

Funding

The research was funded by a University of Leeds PhD studentship. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflicting interests

None

Reference list

1 Ells LJ, Demaio A, Farpour-Lambert N. Diet, genes, and obesity. *BMJ* 2018;**360**:k7. doi:10.1136/BMJ.K7

2 Agha M, Agha R. The rising prevalence of obesity. *Int J Surg Oncol* 2017;**2**:e17. doi:10.1097/IJ9.0000000000000017

3 Carneiro IP, Elliott SA, Siervo M, *et al.* Is Obesity Associated with Altered Energy Expenditure? *Adv Nutr* 2016;**7**:476–87. doi:10.3945/an.115.008755.findings

4 Seale JL, Conway JM, Canary JJ. Seven-day validation of doubly labeled water method using indirect room calorimetry. *J Appl Physiol* 1993;**74**:402–9. doi:10.1152/jappl.1993.74.1.402

5 Delany JP. Measurement of energy expenditure. Pediatr. Blood Cancer. 2012;**58**:129–34. doi:10.1002/pbc.23369

6 Hills AP, Mokhtar N, Byrne NM. Assessment of Physical Activity and Energy Expenditure: An Overview of Objective Measures. *Front Nutr* 2014;**1**:1–16. doi:10.3389/fnut.2014.00005

7 Dhurandhar N V, Schoeller D, Brown AW, *et al.* Energy Balance Measurement: When Something is Not Better than Nothing. *Int J Obes* 2015;**39**:1109–13. doi:10.1038/ijo.2014.199

8 Lyden K, Kozey SL, Staudenmeyer JW, *et al.* A comprehensive evaluation of commonly used accelerometer energy expenditure and MET prediction equations. *Eur J Appl Physiol* 2011;**111**:187–201. doi:10.1007/s00421-010-1639-8

9 Chen KY, Acra SA, Majchrzak K, *et al.* Predicting energy expenditure of physical activity using hip- and wrist-worn accelerometers. *Diabetes Technol Ther* 2003;**5**:1023–33. doi:10.1089/152091503322641088

10 Diaz KM, Krupka DJ, Chang MJ, *et al.* Validation of the Fitbit One® for physical activity measurement at an upper torso attachment site. *BMC Res Notes* 2016;**9**:213. doi:10.1186/s13104-016-2020-8

11 Wright SP, Hall Brown TS, Collier SR, *et al.* How consumer physical activity monitors could transform human physiology research. *Am J Physiol - Regul Integr Comp Physiol* 2017;**312**:R358–67. doi:10.1152/ajpregu.00349.2016

12 Shcherbina A, Mikael Mattsson C, Waggott D, *et al.* Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *J Pers Med* 2017;**7**:1–12. doi:10.3390/jpm7020003

13 Woodman JA, CROUTER SE, BASSETT DR, *et al.* Accuracy of Consumer Monitors for Estimating Energy Expenditure and Activity Type. *Med Sci Sports Exerc* 2017;**49**:371–7. doi:10.1249/MSS.0000000000001090

14 Silva AM, Santos DA, Matias CN, *et al.* Accuracy of a combined heart rate and motion sensor for assessing energy expenditure in free-living adults during a double-blind crossover caffeine trial using doubly labeled water as the reference method. *Eur J Clin Nutr* 2015;**69**:20–7. doi:10.1038/ejcn.2014.51

15 Stahl SE, An H-S, Dinkel DM, *et al.* How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough? *BMJ Open Sport Exerc Med* 2016;**2**:e000106. doi:10.1136/bmjsem-2015-000106

16 Koehler K, Drenowatz C. Monitoring energy expenditure using a multi-sensor device-Applications and limitations of the sense wear armband in athletic populations. *Front Physiol* 2017;**8**:1–7. doi:10.3389/fphys.2017.00983

17 Evenson KR, Goto MM, Furberg RD, *et al.* Systematic review of the validity and reliability of consumer-wearable activity trackers. *Int J Behav Nutr Phys Act* 2015;**12**:159. doi:10.1186/s12966-015-0314-1

18 Brage SS, Westgate K, Franks PW, *et al.* Estimation of free-living energy expenditure by heart rate and movement sensing: A doubly-labelled water study. *PLoS One* 2015;**10**:e0137206. doi:10.1371/journal.pone.0137206

19 McInnes MDF, Moher D, Thombs BD, *et al.* Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies. *JAMA* 2018;**319**:388. doi:10.1001/jama.2017.19163

20 Nelson MB, Kaminsky LA, Dickin DC, *et al.* Validity of Consumer-Based Physical Activity Monitors for Specific Activity Types. *Med Sci Sports Exerc* 2016;**48**:1619–28. doi:10.1249/MSS.0000000000000933

21 Van Remoortel H, Giavedoni S, Raste Y, *et al.* Validity of activity monitors in health and chronic disease: a systematic review. *Int J Behav Nutr Phys Act* 2012;**9**:84. doi:10.1186/1479-5868-9-84

22 Rohatgi. WebPlotDigitizer - Extract data from plots, images, and maps. 2017.https://automeris.io/WebPlotDigitizer/ (accessed 6 Mar 2018).

23 Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998;**52**:377–84. doi:10.1136/JECH.52.6.377

24 Deeks JJ, Dinnes J, D'Amico R, *et al.* Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;**7**:iii–x, 1-173.http://www.ncbi.nlm.nih.gov/pubmed/14499048 (accessed 23 Feb 2018).

25 MacDonald H V, Johnson BT, Huedo-Medina TB, *et al.* Dynamic Resistance Training as Stand-Alone Antihypertensive Lifestyle Therapy: A Meta-Analysis. *J Am Heart Assoc* 2016;**5**. doi:10.1161/JAHA.116.003231

26 Hedges L V. Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *J Educ Stat* 1981;**6**:107. doi:10.2307/1164588

27 Cohen J. *Statistical power analysis for the behavioral sciences*. Academic Press 1977. https://www.sciencedirect.com/science/book/9780121790608 (accessed 23 Feb 2018).

28 Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009;**172**:137–59. doi:10.1111/j.1467-985X.2008.00552.x

29 Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;**21**:1539–58. doi:10.1002/sim.1186

30 Egger M, Smith GD, Schneider M, *et al.* Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;**315**:629–34. doi:10.1136/bmj.315.7109.629

31 Lopez GA, Brønd JC, Andersen LB, *et al.* Validation of SenseWear Armband in children, adolescents, and adults. *Scand J Med Sci Sport* 2018;**28**:487–95. doi:10.1111/sms.12920

32 Reeve MD, Pumpa KL, Ball N. Accuracy of the SenseWear Armband Mini and the BodyMedia FIT in resistance training. *J Sci Med Sport* 2014;**17**:630–4. doi:10.1016/j.jsams.2013.08.007

33 Machač S, Procházka M, Radvanský J, *et al.* Validation of Physical Activity Monitors in Individuals with Diabetes: Energy Expenditure Estimation by the Multisensor SenseWear Armband Pro3 and the Step Counter Omron HJ-720 Against Indirect Calorimetry During Walking. *Diabetes Technol Ther* 2013;**15**:413–8. doi:10.1089/dia.2012.0235

34 Alsubheen SA, George AM, Baker A, *et al.* Accuracy of the vivofit activity tracker. *J Med Eng Technol* 2016;**40**:298–306. doi:10.1080/03091902.2016.1193238

35 Bai Y, Hibbing P, Mantis C, *et al.* Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit Charge HR. *J Sports Sci* 2018;**36**:1734–41. doi:10.1080/02640414.2017.1412235

36 Benito PJ, Neiva C, González-Quijano PS, *et al.* Validation of the SenseWear armband

in circuit resistance training with different loads. *Eur J Appl Physiol* 2012;**112**:3155–9. doi:10.1007/s00421-011-2269-5

37 Berntsen S, Hageberg R, Aandstad A, *et al.* Validity of physical activity monitors in adults participating in free-living activities. *Br J Sports Med* 2010;**44**:657–64. doi:10.1136/bjsm.2008.048868

38 Berntsen S, Stafne SN, Maørkved S. Physical activity monitor for recording energy expenditure in pregnancy. *Acta Obstet Gynecol Scand* 2011;**90**:903–7. doi:10.1111/j.1600-0412.2011.01172.x

39 Bhammar DM, Sawyer BJ, Tucker WJ, *et al.* Validity of SenseWear(R) Armband v5.2 and v2.2 for estimating energy expenditure. *J Sports Sci* 2016;**34**:1830–8. doi:10.1080/02640414.2016.1140220

40 Boudreaux BD, Hebert EP, Hollander DB, *et al.* Validity of Wearable Activity Monitors during Cycling and Resistance Exercise. *Med Sci Sports Exerc* 2018;**50**:624–33. doi:10.1249/MSS.0000000000001471

41 Brazeau AS, Karelis AD, Mignault D, *et al.* Accuracy of the SenseWear Armband??? during ergocycling. *Int J Sports Med* 2011;**32**:761–4. doi:10.1055/s-0031-1279768

42 Brazeau AS, Suppere C, Strychar I, *et al.* Accuracy of energy expenditure estimation by activity monitors differs with ethnicity. *Int J Sports Med* 2014;**35**:847–50. doi:10.1055/s-0034-1371837

43 Brazeau AS, Beaudoin N, Bélisle V, *et al.* Validation and reliability of two activity monitors for energy expenditure assessment. *J Sci Med Sport* 2016;**19**:46–50. doi:10.1016/j.jsams.2014.11.001

44 Brugniaux J V., Niva A, Pulkkinen I, *et al.* Polar activity watch 200: A new device to accurately assess energy expenditure. *Br J Sports Med* 2010;**44**:245–9. doi:10.1136/bjsm.2007.045575

45 Calabro MA, Lee JM, Saint-Maurice PF, *et al.* Validity of physical activity monitors for assessing lower intensity activity in adults. *Int J Behav Nutr Phys Act* 2014;**11**:119. doi:10.1186/s12966-014-0119-7

46 Calabro MA, Kim Y, Franke WD, *et al.* Objective and subjective measurement of energy expenditure in older adults: A doubly labeled water study. *Eur J Clin Nutr* 2015;**69**:850–5. doi:10.1038/ejcn.2014.241

47 Casiraghi F, Lertwattanarak R, Luzi L, *et al.* Energy Expenditure Evaluation in Humans and Non-Human Primates by SenseWear Armband. Validation of Energy Expenditure Evaluation by SenseWear Armband by Direct Comparison with Indirect

Calorimetry. *PLoS One* 2013;**8**:e73651. doi:10.1371/journal.pone.0073651

48 Chowdhury EA, Western MJ, Nightingale TE, *et al.* Assessment of laboratory and daily energy expenditure estimates from consumer multisensor physical activity monitors. *PLoS One* 2017;**12**:e0171720. doi:10.1371/journal.pone.0171720

49 Colbert LH, Matthews CE, Havighurst TC, *et al.* Comparative validity of physical activity measures in older adults. *Med Sci Sports Exerc* 2011;**43**:867–76. doi:10.1249/MSS.0b013e3181fc7162

50 Correa JB, Apolzan JW, Shepard DN, *et al.* Evaluation of the ability of three physical activity monitors to predict weight change and estimate energy expenditure. *Appl Physiol Nutr Metab* 2016;**41**:758–66. doi:10.1139/apnm-2015-0461

51 Diaz KM, Krupka DJ, Chang MJ, *et al.* Fitbit®: an Accurate and Reliable Device for Wireless Physical Activity Tracking. *Int J Cardiol* 2016;**185**:138–40. doi:10.1016/j.ijcard.2015.03.038.FITBIT

52 Dondzila C, Garner D. Comparative accuracy of fitness tracking modalities in quantifying energy expenditure. *J Med Eng Technol* 2016;**40**:325–9. doi:10.1080/03091902.2016.1197978

53 Erdogan A, Cetin C, Karatosun H, *et al.* Accuracy of the Polar S810i(TM) Heart Rate Monitor and the Sensewear Pro Armband(TM) to Estimate Energy Expenditure of Indoor Rowing Exercise in Overweight and Obese Individuals. *J Sports Sci Med* 2010;**9**:508–16.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3761702&tool=pmcentrez&rendertype=abstract

54 Fruin ML, Rankin JW. Validity of a multi-sensor Armband in estimating rest and exercise energy expenditure. *Med Sci Sports Exerc* 2004;**36**:1063–9. doi:10.1249/01.MSS.0000128144.91337.38

55 Furlanetto KC, Bisca GW, Oldemberg N, *et al.* Step Counting and Energy Expenditure Estimation in Patients With Chronic Obstructive Pulmonary Disease and Healthy Elderly: Accuracy of 2 Motion Sensors. *Arch Phys Med Rehabil* 2010;**91**:261–7. doi:10.1016/j.apmr.2009.10.024

56 Gastin PB, Cayzer C, Dwyer D, *et al.* Validity of the ActiGraph GT3X+ and BodyMedia SenseWear Armband to estimate energy expenditure during physical activity and sport. *J Sci Med Sport* Published Online First: July 2017. doi:10.1016/j.jsams.2017.07.022

57 Heiermann S, Khalaj Hedayati K, Müller MJ, *et al.* Accuracy of a portable multisensor

body monitor for predicting resting energy expenditure in older people: A comparison with indirect calorimetry. *Gerontology* 2011;**57**:473–9. doi:10.1159/000322109

58 Imboden MT, Nelson MB, Kaminsky LA, *et al.* Comparison of four Fitbit and Jawbone activity monitors with a research-grade ActiGraph accelerometer for estimating physical activity and energy expenditure. *Br J Sports Med* 2017;:bjsports-2016-096990. doi:10.1136/bjsports-2016-096990

59 Jakicic JM, Marcus M, Gallagher KI, *et al.* Evaluation of the SenseWear Pro Armband to Assess Energy Expenditure during Exercise. *Med Sci Sport Exerc* 2004;**36**:897–904. doi:10.1249/01.MSS.0000126805.32659.43

60 Johannsen DL, Calabro MA, Stewart J, *et al.* Accuracy of armband monitors for measuring daily energy expenditure in healthy adults. *Med Sci Sports Exerc* 2010;**42**:2134–40. doi:10.1249/MSS.0b013e3181e0b3ff

61 Kim Y, Welk GJ. Criterion validity of competing accelerometry-based activity monitoring devices. *Med Sci Sports Exerc* 2015;**47**:2456–63. doi:10.1249/MSS.0000000000000691

62 King GA, Torres N, Potter C, *et al.* Comparison of activity monitors to estimate energy cost of treadmill exercise. *Med Sci Sports Exerc* 2004;**36**:1244–51. doi:10.1249/01.MSS.0000132379.09364.F8

63 Koehler K, Braun H, De MarIes M, *et al.* Assessing energy expenditure in male endurance athletes: Validity of the sensewear armband. *Med Sci Sports Exerc* 2011;**43**:1328–33. doi:10.1249/MSS.0b013e31820750f5

64 Lee CM, Gorelick M, Mendoza A. Accuracy of an infrared led device to measure heart rate and energy expenditure during rest and exercise. *J Sports Sci* 2011;**29**:1645–53. doi:10.1080/02640414.2011.609899

65 Lee J-MM, Kim Y-WY, Welk GJ. Validity of consumer-based physical activity monitors. *Med Sci Sports Exerc* 2014;**46**:1840–8. doi:10.1249/MSS.0000000000000287

66 MacKey DC, Manini TM, Schoeller DA, *et al.* Validation of an armband to measure daily energy expenditure in older adults. *Journals Gerontol - Ser A Biol Sci Med Sci* 2011;**66 A**:1108–13. doi:10.1093/gerona/glr101

67 Martien S, Seghers J, Boen F, *et al.* Energy expenditure in institutionalized older adults: Validation of SenseWear Mini. *Med Sci Sports Exerc* 2015;**47**:1265–71. doi:10.1249/MSS.0000000000000529

68 McMinn D, Rowe DA, Murtagh S, *et al.* The effect of a school-based active

commuting intervention on children's commuting physical activity and daily physical activity. *Prev Med (Baltim)* 2012;**54**:316–8. doi:10.1016/j.ypmed.2012.02.013

69 Melanson EL, Dykstra JC, Szuminsky N. A novel approach for measuring energy expenditure in free-living humans. *2009 Annu Int Conf IEEE Eng Med Biol Soc* 2009;**2009**:6873–7. doi:10.1109/IEMBS.2009.5333124

70 Montoye AHK, Mitrzyk JR, Molesky MJ. Comparative Accuracy of a Wrist-Worn Activity Tracker and a Smart Shirt for Physical Activity Assessment. *Meas Phys Educ Exerc Sci* 2017;**21**:201–11. doi:10.1080/1091367X.2017.1331166

71 Murakami H, Kawakami R, Nakae S, *et al.* Accuracy of Wearable Devices for Estimating Total Energy Expenditure. *JAMA Intern Med* 2016;**176**:702. doi:10.1001/jamainternmed.2016.0152

72 Papazoglou D, Augello G, Tagliaferri M, *et al.* Evaluation of a multisensor armband in estimating energy expenditure in obese individuals. *Obesity* 2006;**14**:2217–23. doi:10.1038/oby.2006.260

73 Price K, Bird SR, Lythgo N, *et al.* Validation of the Fitbit One, Garmin Vivofit and Jawbone UP activity tracker in estimation of energy expenditure during treadmill walking and running. *J Med Eng Technol* 2017;**41**:208–15. doi:10.1080/03091902.2016.1253795

74 Reece JD, Barry V, Fuller DK, *et al.* Validation of the SenseWear Armband as a Measure of Sedentary Behavior and Light Activity. *J Phys Act Heal* 2015;**12**:1229–37. doi:10.1123/jpah.2014-0136

75 Rousset S, Fardet A, Lacomme P, *et al.* Comparison of total energy expenditure assessed by two devices in controlled and free-living conditions. *Eur J Sport Sci* 2015;**15**:391–9. doi:10.1080/17461391.2014.949309

76 Ryan J, Gormley J. Measurement of energy expenditure by activity monitors. *Phys Ther Rev* 2013;**18**:239–62. doi:10.1179/1743288X13Y.0000000063

77 Slinde F, Bertz F, Winkvist A, *et al.* Energy expenditure by multisensor armband in overweight and obese lactating women validated by doubly labeled water. *Obesity* 2013;**21**:2231–5. doi:10.1002/oby.20363

78 Smith KM, Lanningham-Foster LM, Welk GJ, *et al.* Validity of the SenseWear® armband to predict energy expenditure in pregnant women. *Med Sci Sports Exerc* 2012;**44**:2001–8. doi:10.1249/MSS.0b013e31825ce76f

79 Stackpool C, Porcari JP, Mikat RP GC and FC. The accuracy of various activity trackers in estimating steps taken and energy expenditure. *J Fit Res* 2013;**53**:1689–99.

doi:10.1017/CBO9781107415324.004

80 St-onge M, Mignault D, Allison DB, *et al.* Evaluation of a portable device to measure daily energy expenditure. *Am Soc Nutr* 2007;**85**:742–9. doi:10.1093/ajcn/85.3.742

81 Soric M, Mikulic P, Misigoj-Durakovic M, *et al.* Validation of the Sensewear Armband during recreational in-line skating. *Eur J Appl Physiol* 2012;**112**:1183–8. doi:10.1007/s00421-011-2045-6

82 Tucker WJ, Bhammar DM, Sawyer BJ, *et al.* Validity and reliability of Nike + Fuelband for estimating physical activity energy expenditure. *BMC Sports Sci Med Rehabil* 2015;**7**:14. doi:10.1186/s13102-015-0008-7

83 Vanhelst J, Mikulovic J, Bui-Xuan G, *et al.* Comparison of two ActiGraph accelerometer generations in the assessment of physical activity in free living conditions. *BMC Res Notes* 2012;**5**:187. doi:10.1186/1756-0500-5-187

84 Van Hoye K, Mortelmans P, Lefevre J. Validation of the SenseWear Pro3 armband using an incremental exercise test. *J Strength Cond Res* 2014;**28**:2806–14. doi:10.1519/JSC.0b013e3182a1f836

85 Van Hoye K, Boen F, Lefevre J. Validation of the SenseWear Armband in different ambient temperatures. *J Sports Sci* 2015;**33**:1007–18. doi:10.1080/02640414.2014.981846

86 Vernillo G, Savoldelli A, Pellegrini B, *et al.* Validity of the SenseWear Armband to Assess Energy Expenditure in Graded Walking. *J Phys Act Heal* 2015;**12**:178–83. doi:10.1123/jpah.2013-0437

87 Wahl Y, Düking P, Droszez A, *et al.* Criterion-validity of commercially available physical activity tracker to estimate step count, covered distance and energy expenditure during sports conditions. *Front Physiol* 2017;**8**:725. doi:10.3389/fphys.2017.00725

88 Wallen MP, Gomersall SR, Keating SE, *et al.* Accuracy of heart rate watches: Implications for weight management. *PLoS One* 2016;**11**:e0154420. doi:10.1371/journal.pone.0154420

89 Sanghvi A, Redman LM, Martin CK, *et al.* Validation of an inexpensive and accurate mathematical method to measure long-term changes in free-living energy intake. *Am J Clin Nutr* 2015;**102**:353–8. doi:10.3945/ajcn.115.111070

90 Pollard TM, Wagnild JM. Gender differences in walking (for leisure, transport and in total) across adult life: a systematic review. *BMC Public Health* 2017;**17**:341. doi:10.1186/s12889-017-4253-4

91 Schneller MB, Pedersen MT, Gupta N, *et al.* Validation of five minimally obstructive methods to estimate physical activity energy expenditure in young adults in semi-standardized settings. *Sensors (Basel)* 2015;**15**:6133–51. doi:10.3390/s150306133

92 Welk GJ, McClain JJ, Eisenmann JC, *et al.* Field Validation of the MTI Actigraph and BodyMedia Armband Monitor Using the IDEEA Monitor. *Obes* 2007;**15**:918–28.http://search.ebscohost.com/login.aspx?direct=true&db=sph&AN=24897397&site=ehost-live

93 A.H.K. M, J.M. P, L.M. M, *et al.* Comparison of Activity Type Classification Accuracy from Accelerometers Worn on the Hip, Wrists, and Thigh in Young, Apparently Healthy Adults. *Meas Phys Educ Exerc Sci* 2016;**20**:173–83. doi:http://dx.doi.org/10.1080/1091367X.2016.1192038

94 Plasqui G. Smart approaches for assessing free-living energy expenditure following identification of types of physical activity. *Obes Rev* 2017;**18**:50–5. doi:10.1111/obr.12506

95 Brage S, Ekelund U, Brage N, *et al.* Hierarchy of individual calibration levels for heart rate and accelerometry to measure physical activity. *J Appl Physiol* 2007;**103**:682–92. doi:10.1152/japplphysiol.00092.2006

96 Benedetto S, Caldato C, Bazzan E, *et al.* Assessment of the Fitbit Charge 2 for monitoring heart rate. *PLoS One* 2018;**13**:e0192691. doi:10.1371/journal.pone.0192691

97 Higgins JPT. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *Int J Epidemiol* 2008;**37**:1158–60. doi:10.1093/ije/dyn204

98 Kipp S, Byrnes WC, Kram R. Calculating metabolic energy expenditure across a wide range of exercise intensities: the equation matters. *Appl Physiol Nutr Metab* 2018;:1–4. doi:10.1139/apnm-2017-0781

Legends:

Table 1. Moderation analysis for level of sensors and grade of device by subgroup. Data are shown where at least 3 comparisons were included. *P*-value refers to a between subgroup comparison. *Significant effect size at the subgroup level ($p<.05$). Abbreviations: Accelerometry alone (ACC), accelerometry and heart rate (ACC+HR), accelerometry and heart rate and heat sensing (ACC+HR+HS) and accelerometry and heat sensing (ACC+HS). Activity energy expenditure (AEE), Total energy expenditure (TEE), Doubly labelled water (DLW).

PLEASE INSERT FIGURE 1 AROUND LINE 216
Figure 1. Flow diagram of study selection.

PLEASE INSERT FIGURE 2 AROUND LINE 254
Figure 2. Pooled Hedges' g and 95% confidence intervals (CI) 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.
Abbreviations: *Actical (ACT), Actigraph GT3X (AGT3X), Apple watch (AW), Apple Watch series 2 (AWS2), Beurer AS80 (BA), Bodymedia CORE armband (BMC), Basis Peak (BP), Epson Pulsense (EP), ePulse Personal Fitness Assistant (EPUL), Fitbit Blaze (FB), Fitbit Charge (FC), Fitbit Charge 2 (FC2), Fitbit Charge HR (FCHR), Fitbit Flex (FF), Garmin Forerunner 225 (GF225), Garmin Forerunner 920XT (GF920XT), Garmin Vivoactive (GVA), Garmin vivofit (GVF), Garmin vivosmart (GVS), Garmin Vivosmart HR (GVHR), Jawbone UP (JU), Jawbone UP24 (JU24), LifeChek calorie sensor (LC), Mio Alpha (MA), Microsoft band (MB), Misfit Shine (MS), Nike Fuel band (NF), Polar Loop (PL), Polar: AW200 (PO200), Polar: AW360 (PA360), Samsung Gear S (SG), SenseWear Armband (SWA), SenseWear Armband Pro 2 (SWA p2), SenseWear Armband Pro 3 (SWA p3), SenseWear Armband MINI (SWAM), TOMTOM Touch (TT), Vivago (V), Withings Pulse (WP), Withings Pulse O2 (WPO).*

PLEASE INSERT FIGURE 3 AROUND LINE 284
Figure 3. Pooled Hedges' g and 95% confidence intervals (CI) 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.
Abbreviations: *Actigraph GT3X (AGT3X), Apple watch (AW), Beurer AS80 (BA), Bodymedia CORE armband (BMC), Basis Peak (BP), ePulse Personal Fitness Assistant (EPUL), Fitbit Charge (FC), Fitbit Charge HR (FCHR), Fitbit Flex (FF), Garmin Forerunner 225 (GF225), Garmin Forerunner 920XT (GF920XT), Garmin Vivoactive (GVA), Garmin vivofit (GVF), Garmin vivosmart (GVS), Jawbone UP (JU), Jawbone UP24 (JU24), Microsoft band (MB), Nike Fuel band (NF), Polar Loop (PL), Polar: AW200 (PO200), SenseWear Armband (SWA), SenseWear Armband Pro 2 (SWA p2), SenseWear Armband Pro 3 (SWA p3), SenseWear Armband MINI (SWAM), Vivago (V), Withings Pulse (WP), Withings Pulse O2 (WPO).*

PLEASE INSERT FIGURE 4 AROUND LINE 313
Figure 4. Pooled Hedges' g and 95% confidence intervals (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.
Abbreviations: *Apple watch (AW), Bodymedia CORE armband (BMC), Basis Peak (BP), ePulse Personal Fitness Assistant (EPUL), Fitbit Charge HR (FCHR), Fitbit Flex (FF), Garmin Forerunner 225 (GF225), Garmin vivofit (GVF), Jawbone UP (JU), Jawbone UP24 (JU24), Microsoft band (MB), SenseWear Armband Pro 2 (SWA p2), SenseWear Armband Pro 3 (SWA p3), SenseWear Armband MINI (SWAM), Vivago (V), Withings Pulse (WP).*

PLEASE INSERT FIGURE 5 AROUND LINE 320

Figure 5. Pooled Hedges' g and 95% confidence intervals (CI) for estimates of energy expenditure relative to criterion measures per device for total energy expenditure (TEE).
Total refers to number of effect sizes. A negative Hedges' g statistic represents an underestimation and a positive Hedges' g represents an overestimation.
Abbreviations: *Epson Pulsense (EP), Fitbit Flex (FF), Garmin vivofit (GVF), Jawbone UP24 (JU24), Misfit Shine (MS), SenseWear Armband (SWA), SenseWear Armband Pro 2 (SWA p2), SenseWear Armband Pro 3 (SWA p3), SenseWear Armband MINI (SWAM), Withings Pulse O2 (WPO).*

Figure 1

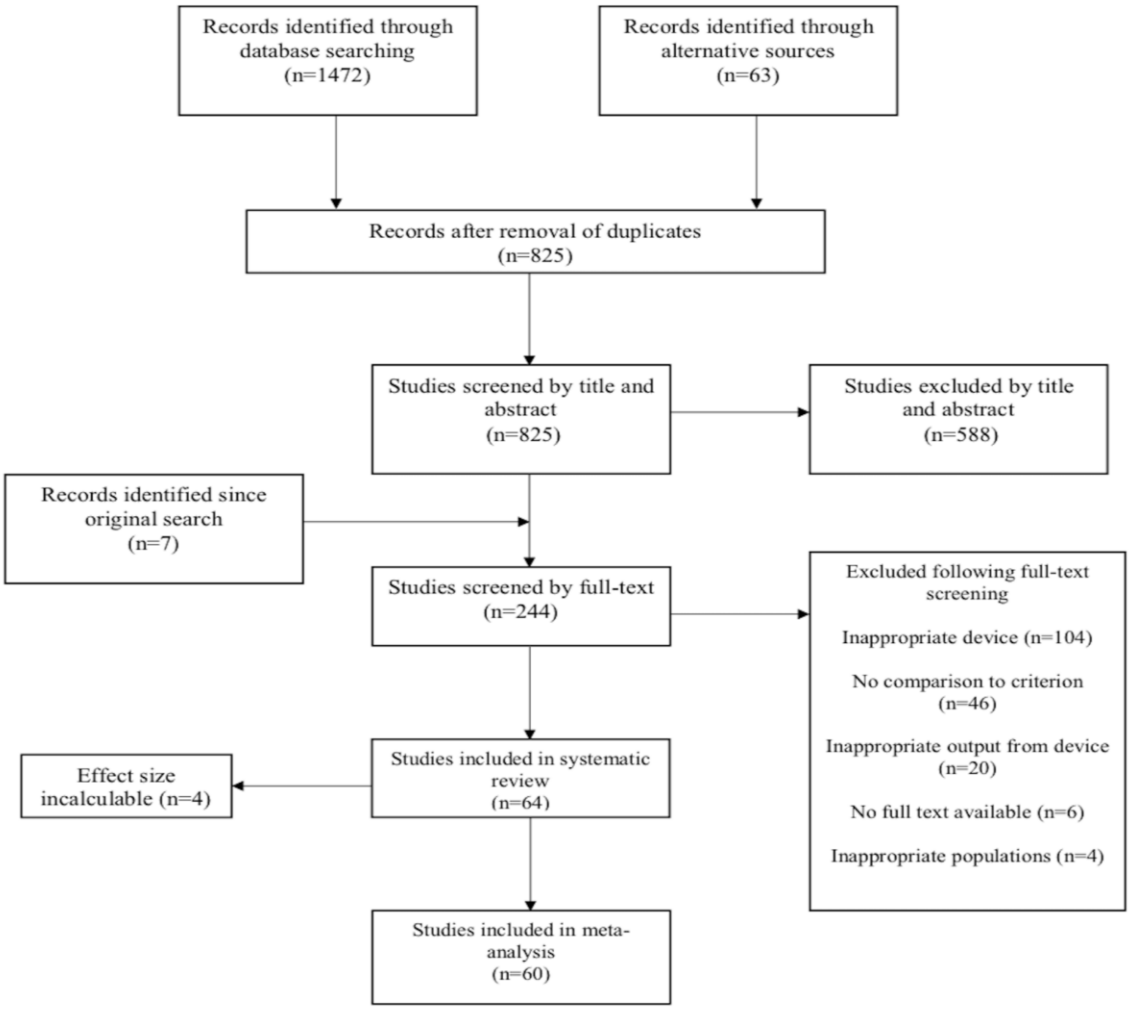

Records identified through database searching (n=1472)
Records identified through alternative sources (n=63)
Records after removal of duplicates (n=825)
Studies screened by title and abstract (n=825)
Studies excluded by title and abstract (n=588)
Records identified since original search (n=7)
Studies screened by full-text (n=244)
Excluded following full-text screening
Inappropriate device (n=104)
No comparison to criterion (n=46)
Inappropriate output from device (n=20)
No full text available (n=6)
Inappropriate populations (n=4)
Studies included in systematic review (n=64)
Effect size incalculable (n=4)
Studies included in meta-analysis (n=60)

Figure 2

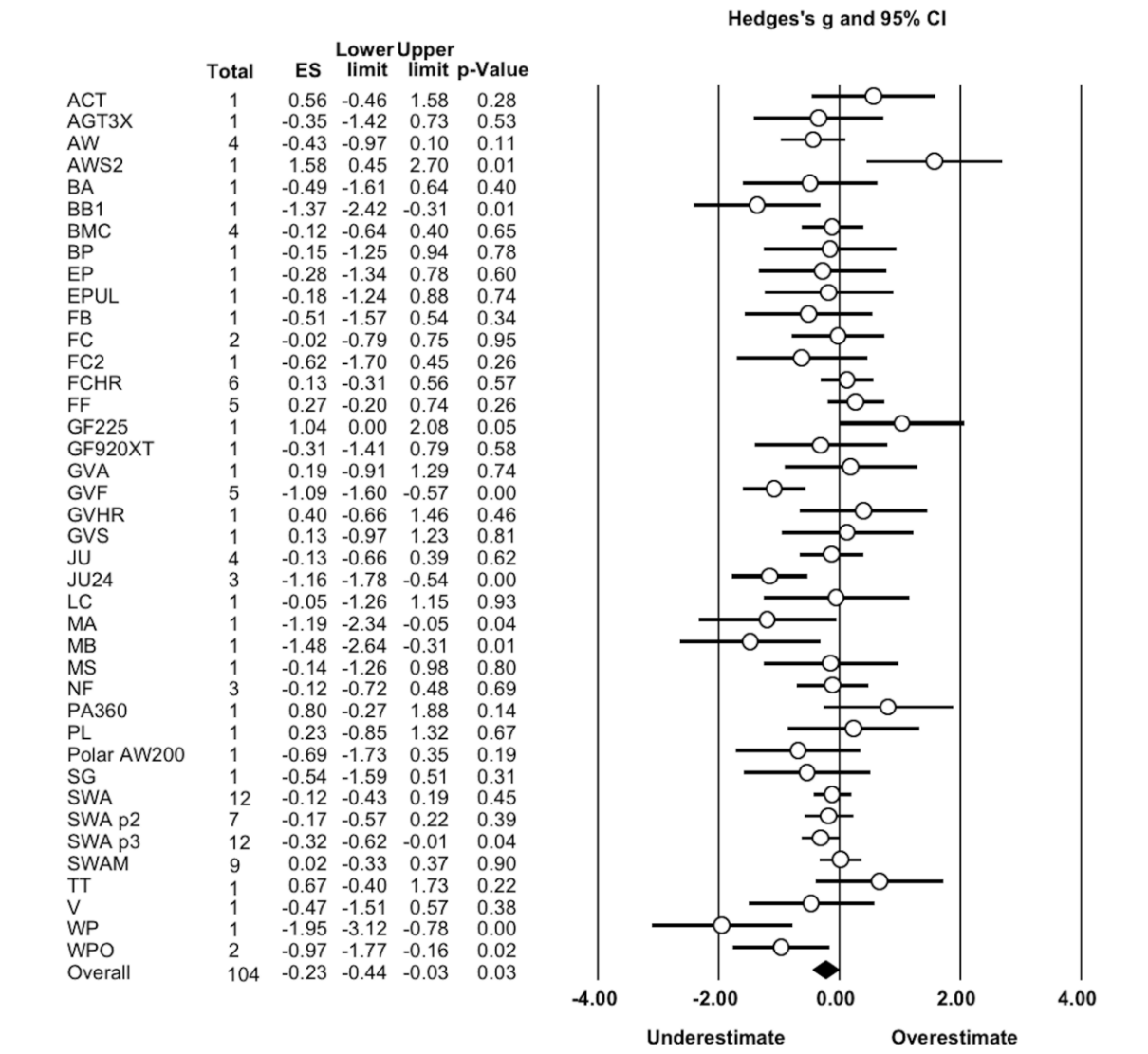
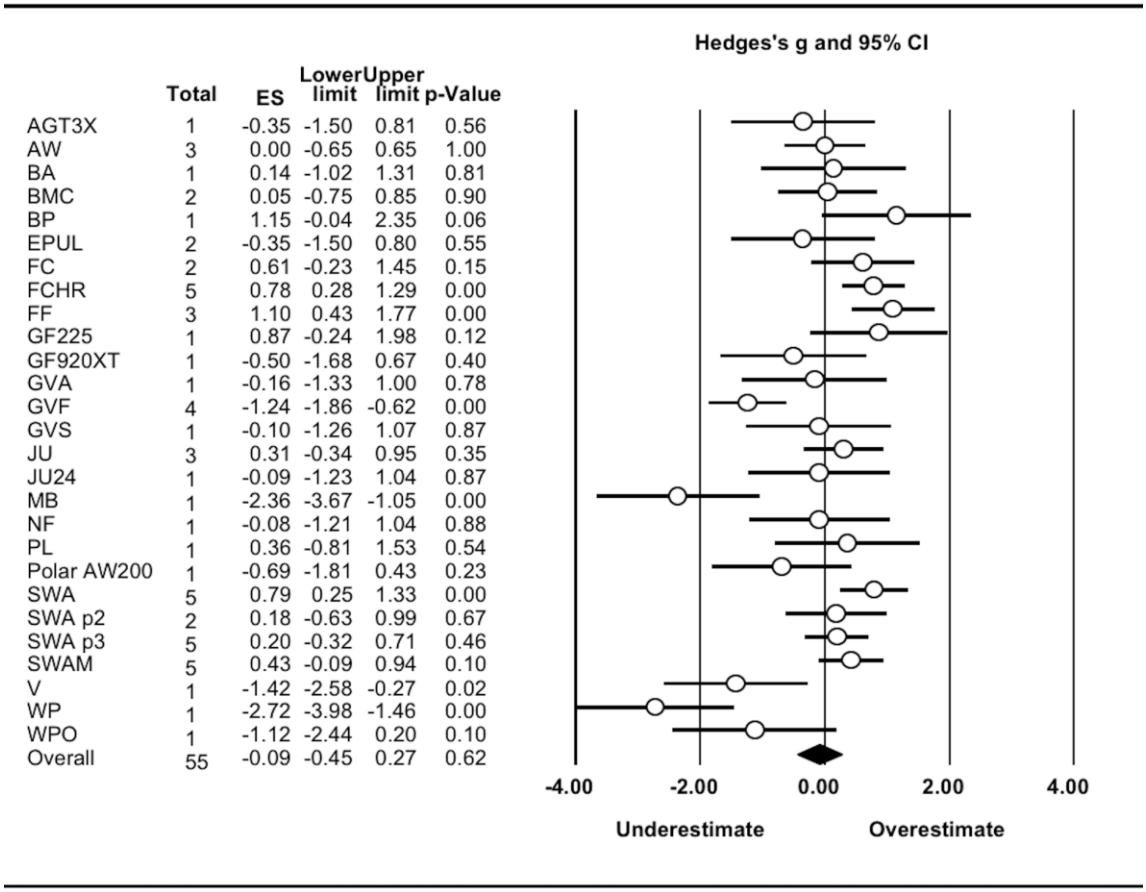
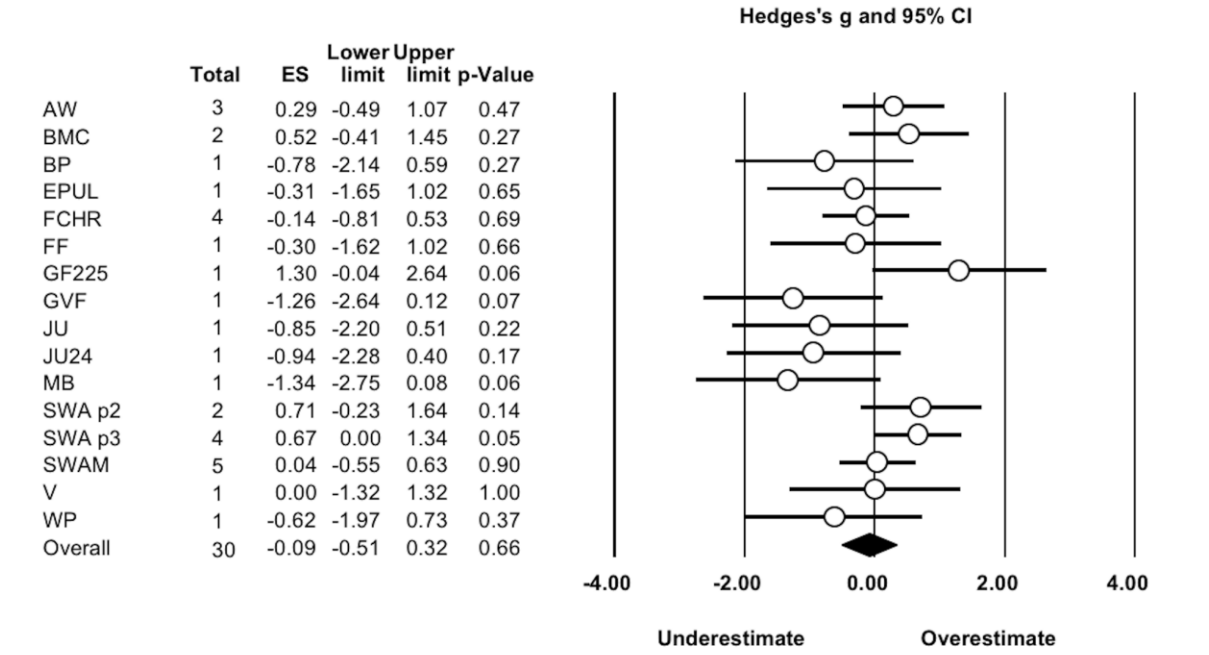

Hedges's g and 95% CI
Total ES Lower limit Upper limit p-Value
ACT 1 0.56 -0.46 1.58 0.28
AGT3X 1 -0.35 -1.42 0.73 0.53
AW 4 -0.43 -0.97 0.10 0.11
AWS2 1 1.58 0.45 2.70 0.01
BA 1 -0.49 -1.61 0.64 0.40
BB1 1 -1.37 -2.42 -0.31 0.01
BMC 4 -0.12 -0.64 0.40 0.65
BP 1 -0.15 -1.25 0.94 0.78
EP 1 -0.28 -1.34 0.78 0.60
EPUL 1 -0.18 -1.24 0.88 0.74
FB 1 -0.51 -1.57 0.54 0.34
FC 2 -0.02 -0.79 0.75 0.95
FC2 1 -0.62 -1.70 0.45 0.26
FCHR 6 0.13 -0.31 0.56 0.57
FF 5 0.27 -0.20 0.74 0.26
GF225 1 1.04 0.00 2.08 0.05
GF920XT 1 -0.31 -1.41 0.79 0.58
GVA 1 0.19 -0.91 1.29 0.74
GVF 5 -1.09 -1.60 -0.57 0.00
GVHR 1 0.40 -0.66 1.46 0.46
GVS 1 0.13 -0.97 1.23 0.81
JU 4 -0.13 -0.66 0.39 0.62
JU24 3 -1.16 -1.78 -0.54 0.00
LC 1 -0.05 -1.26 1.15 0.93
MA 1 -1.19 -2.34 -0.05 0.04
MB 1 -1.48 -2.64 -0.31 0.01
MS 1 -0.14 -1.26 0.98 0.80
NF 3 -0.12 -0.72 0.48 0.69
PA360 1 0.80 -0.27 1.88 0.14
PL 1 0.23 -0.85 1.32 0.67
Polar AW200 1 -0.69 -1.73 0.35 0.19
SG 1 -0.54 -1.59 0.51 0.31
SWA 12 -0.12 -0.43 0.19 0.45
SWA p2 7 -0.17 -0.57 0.22 0.39
SWA p3 12 -0.32 -0.62 -0.01 0.04
SWAM 9 0.02 -0.33 0.37 0.90
TT 1 0.67 -0.40 1.73 0.22
V 1 -0.47 -1.51 0.57 0.38
WP 1 -1.95 -3.12 -0.78 0.00
WPO 2 -0.97 -1.77 -0.16 0.02
Overall 104 -0.23 -0.44 -0.03 0.03
-4.00
-2.00
0.00
2.00
4.00
Underestimate
Overestimate
Meta Analysis Overall

Figure 3



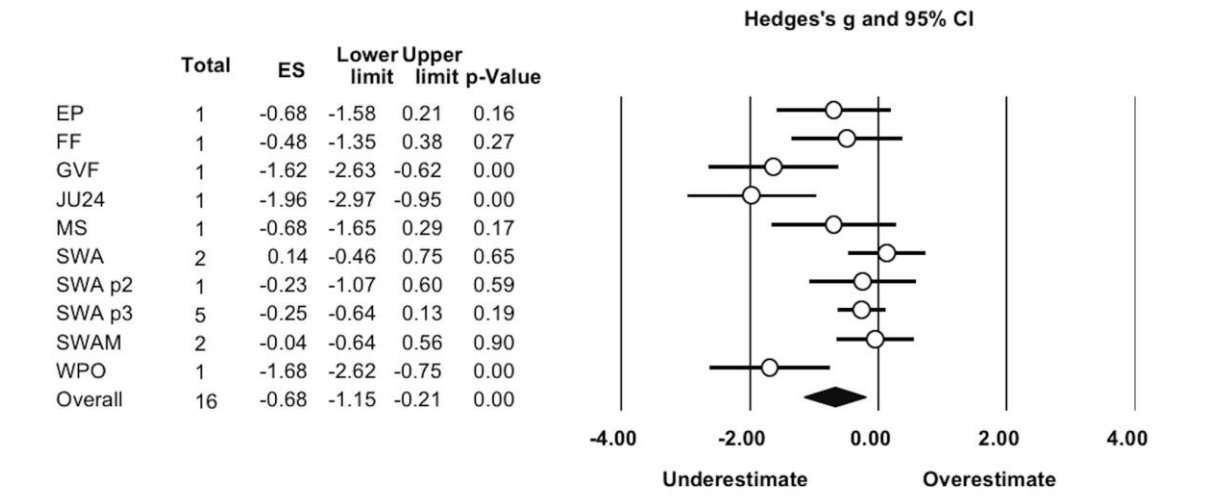
Meta Analysis Ambulation and Stairs

Figure 4



Meta Analysis Sedentary and Household

Figure 5



Meta Analysis TEE (DLW)