Published in final edited form as:

J Meas Phys Behav. 2025 January ; 8(1): . doi:10.1123/jmpb.2024-0051.

Performance evaluation of algorithms to estimate daily sedentary time using wrist-worn sensors in free-living adults

Charles E. Matthews[1], **Pedro Saint-Maurice**[2], **Joshua R. Freeman**[1], **Hayden A. Hayes**[1], **Alaina H Shreves**[1,3,4], **Aiden Doherty**[3,4], **Eric T. Hyde**[1,5], **Katie Ylarregui**[6], **Rena R. Jones**[7], **Sarah K. Keadle**[6]

[1]Metabolic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

[2]Breast Cancer Research Unit, Champalimaud Foundation, Lisbon, Portugal

[3]Nuffield Department of Population Health, University of Oxford, UK.

[4]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK.

[5]Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, La Jolla, CA, USA.

[6]Department of Kinesiology and Public Health, California Polytechnic State University, San Luis Obispo, CA, USA

[7]Occupational & Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

Abstract

Purpose: Given the limited real-world testing of algorithms for wrist-worn sensors to estimate sedentary time, we examined the performance of 21 algorithms in free-living adults.

Methods: Seventy-one adults (35–65 years) wore a GENEActiv (wrist) and an activPAL (thigh) sensor for up to 10 days. activPAL was our reference measure. We estimated sedentary time (hours/day) using 21 classification algorithms, including cut point and machine-learning methods. Valid days from each monitor were matched by date and mean values were calculated. Equivalence testing (±10%) and linear regression were used to compare each algorithm's estimate to the reference, over all participants and by sex and age.

Results: activPAL recorded a mean of 9.4 hours/d sedentary. Five of 21 algorithms (24%) estimated sedentary time within 10% (±0.94 hours) of the reference. Two of these methods employed machine-learning algorithms (Trost Extended, OxWearables) and three employed cut points (GGIR ENMO 40mg; Bakrania ENMO 32.6mg; Fraysse ENMOa 62.5mg). Variance explained in linear regression was relatively high for the machine-learning (R^2=0.44–0.63) and cut point algorithms developed for younger (R^2=0.30–0.64) and older (R^2=0.45–0.66) adults. More

Corresponding author: Charles E. Matthews, PhD, Senior Investigator, Metabolic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, NCI Shady Grove, 9609 Medical Center Dr. Rm. 6E444, Bethesda, MD 20892-9776, charles.matthews2@nih.gov.

accurate performance was noted for algorithms developed in studies using posture-based ground truth measures and conducted in free-living settings.

Conclusion: Fifteen of 21 (71%) algorithms produced estimates of sedentary time that were moderate-strongly correlated with the reference measure, but only five (24%) were within 10% of the reference. Free-living benchmarking studies like this can identify more accurate and precise algorithms to estimate sedentary time and identify characteristics of algorithm development studies that yield better results.

Keywords

physical activity; public health; surveillance; epidemiology

INTRODUCTION

Adults in the United States (US) spend about 60% of their waking day, roughly 9.5 hours/day, sitting while in transit, at work, or during leisure time (Matthews et al., 2021). Greater amounts of sedentary behavior, or too much sitting, has been associated with a variety of adverse health outcomes (Katzmarzyk et al., 2019) and reducing sedentary time is a key part of the Physical Activity Guidelines for Americans, 2nd Edition (i.e., "…move more and sit less throughout the day")(Piercy et al., 2018). Accordingly, measurement methods that accurately distinguish between sedentary and physically active behaviors are critical for epidemiologic studies of physical activity and health, public health surveillance, and health promotion programs designed to reduce sedentary time and increase physical activity.

Epidemiologic studies utilizing sensors worn on the wrist, waist, and thigh have already provided novel insights into the health risks of sedentary time and health benefits of physical activity (Ahmadi et al., 2022; Chastin et al., 2015; Saint-Maurice et al., 2018; Walmsley et al., 2022) and newer cohorts are now in position to make further advances (American Cancer Society, 2024; "The German National Cohort: aims, study design and organization," 2014; Master et al., 2022). Many new studies have elected to have participants wear sensors on their wrist to simplify monitor administration, improve compliance (Doherty et al., 2017), and facilitate measurement of sleep, another important health behavior (e.g., (Fernandez-Mendoza, 2019)). Increased use of wrist-worn sensors has spawned a broad effort to develop classification algorithms that are applied to wrist movement and wrist orientation signals to estimate time spent in sedentary and physically active behaviors (Liu et al., 2022). Simple cut points and more complex algorithms to estimate sedentary time have been proposed for use in middle-aged (Bakrania et al., 2016; Dillon et al., 2016; Esliger et al., 2011; Hildebrand et al., 2017; Pavey et al., 2017; Rowlands et al., 2016) and older adults (Duncan et al., 2020; Fraysse et al., 2020; Migueles et al., 2021; Sanders et al., 2019), including cut-points for Euclidean norm minus one (ENMO) (Dillon et al., 2016; Esliger et al., 2011; Fraysse et al., 2020; Hildebrand et al., 2017; Migueles et al., 2021; Sanders et al., 2019) and mean amplitude deviation (MAD) acceleration metrics (Bakrania et al., 2016). Several cut-point independent algorithms have also been proposed, including those which utilize wrist orientation (Rowlands et al., 2016; Straczkiewicz et al., 2020) and machine-learning methods (Ellis et al., 2016; Pavey et al., 2017; Walmsley et al., 2022).

However, few of these methods have undergone rigorous (phase III) performance evaluation that includes: (1) testing in study populations independent from those employed to develop the methods; (2) evaluation during free-living activities in daily life; and (3) use of strong reference measures (Keadle et al., 2019). Furthermore, studies that have tested available algorithms have tested either one, or just a few, algorithms at a time. Concurrent evaluation of multiple algorithms against a common reference measurement can help identify methods that are more accurate and equivalent to one another (Dixon et al., 2018), and may provide insight into the elements of study designs employed to develop/calibrate the methods that yield better results (Fuller et al., 2022).

Thus, to better understand the performance of algorithms to estimate daily sedentary time using wrist-worn sensors, we evaluated a variety of methods in 71 adults in comparison to estimates from a thigh-worn monitor used as our reference measurement.

METHODS

Study Population.

Participants in the Physical Activity, Light, and Sleep (PALS) study were recruited into a sensor-based measurement study through word of mouth, social media advertisements, fliers and via Facebook between September 2022 and March 2023. Inclusion criteria were that participants were 30–65 years of age and had internet access via a smart phone, tablet, or computer. Exclusion criteria included orthopedic limitations that impair their ability to walk and major health conditions (e.g., cancer, heart disease, liver, or kidney disease) that require a restriction of physical activity at the time of recruitment. Recruitment was designed to enroll comparable numbers of adults above and below 50 years of age. All participants read and signed an informed consent document approved by the Institutional Review Board.

Each participant was asked to complete two in-person or virtual "laboratory visits". Participants completed a demographic questionnaire that asked to report their age (years), sex, race and ethnicity, employment status, education level. Height and weight were self-reported. They also wore an activPAL accelerometer on their thigh and a GENEActiv accelerometer for 10-days. The monitors were distributed in-person or by mail as needed using established protocols. Participants were also asked to complete an electronic sleep diary each morning of the study period, from which time out of and into bed was extracted and used to define the waking period for all measurements.

Reference Measure.

The activPAL4 (PAL Technologies Inc., Glasgow, Scotland) is a thigh-worn triaxial accelerometer that samples at 20 Hz. Raw data were processed using activPAL software (CREA, v8.10.12.60), and the resulting event files were further processed by incorporating the start/stop times for each waking period using an R-package (activPAL processing, v0.4.1) (17). For analysis, the resulting waking day estimates were categorized as hours per day spent sedentary (sedentary, lying, traveling) and physically active (standing, stepping, cycling), and daily step counts were calculated. The strong performance of the activPAL for estimating sedentary time has been established in uncontrolled/free-living studies vs. direct observation (O'Brien et al., 2022). Seven studies using direct observation as criterion found mean differences or mean error for the device to range from < 1% to 5% with a median error of 1% (O'Brien et al., 2022). Three of these studies reported R^2 values at 0.94 or greater. Thus, the group level performance of activPAL appears to approach that of direct observation in free-living studies and was employed as the reference measure in our study.

Wrist-worn Accelerometer.

The GENEActiv monitor was initialized to collect raw 3-dimensional acceleration data at 40Hz and was worn on the non-dominant wrist. To develop analytic data for the sedentary behavior classification algorithms that used cut-points, raw acceleration data was processed using GGIR (v3.0.9), and GGIR parameters were set to exclude days with <19 hours of total valid wear time (i.e., sleep and waking periods), and the first and last non-protocol days on the file. Non-wear time was estimated from "invalid" epochs as defined by GGIR. Three acceleration metrics were calculated (ENMO, ENMOa, MAD) and output to a 60-second epoch data file to implement cut-point based algorithms within the waking periods defined by the sleep diary data.

Classification Algorithms Evaluated.

Algorithms to be evaluated were identified through a general search via documentation in available in the GGIR and Actimetric R packages used to score raw accelerometer data and by review of citations associated with the original publications proposing the methods.

We identified 12 studies in younger and older adults that proposed 21 different algorithms, 16 of which were cut-point based (Bakrania et al., 2016; Dillon et al., 2016; Duncan et al., 2020; Esliger et al., 2011; Fraysse et al., 2020; Hildebrand et al., 2017; Migueles et al., 2021; Sanders et al., 2019) and 5 of which employed machine learning- methodologies (Ellis et al., 2016; Pavey et al., 2017; Straczkiewicz et al., 2020; Walmsley et al., 2022). All methods were processed using raw acceleration data and thus we assume monitor independence and for descriptive purposes we also examined methods calibrated for both the non-dominant, dominant, left, and right hands when available. We implemented the machine-learning based methods using OxWearables code (Walmsley et al., 2022) and the Actimetric R package (version 0.1.2)(Migueles et al., 2019). We attempted to use the Sedentary Sphere method (Rowlands et al., 2016) but we were unable to generate reliable results, and the method was dropped from our analysis.

A more detailed description of the calibration studies used to develop each algorithm including the devices originally employed, study settings, training activities, reference measures and study results are presented in Supplementary Table 1. Briefly, the machine-learning based methods included Ellis (Ellis et al., 2016), OxWearables (Walmsley et al., 2022) and Trost Original (Pavey et al., 2017) plus two extensions of Trost Original implemented within the Actimetric R package that integrates sleep and wear time classifiers (Trost Extended) (Ahmadi et al., 2020) and one that integrates a body posture classifier using wrist orientation information (Trost SedUp) (Straczkiewicz et al., 2020). Cut point thresholds developed within younger adult populations included those from Bakrania (Bakrania et al., 2016), Hildebrand (Hildebrand et al., 2017), Dillon (Dillon et al., 2016) and Esliger (Esliger et al., 2011). We also included the default cut-point within GGIR of ENMO 40 mg, which appears to be a derivative of Hildebrand (Hildebrand et al., 2017). Cut point thresholds developed in older adult populations included those from Duncan (Duncan et al., 2020), Miqueles (Migueles et al., 2021), Sanders (Sanders et al., 2019), and Fraysse (Fraysse et al., 2020).

Statistical Methods.

Characteristics of the study population were described using means, standard deviations (SD), frequencies, and percentages. We evaluated the performance of each algorithm in comparison to days matched to the activPAL. Each valid day of observation for each monitor included at least 8 hours of observation in the waking day, at least 30 minutes of sedentary time, and monitor non-wear less than 5% of the waking day. For analysis, data from the date matched valid days from each device were averaged.

Using methods described by Dixon and colleagues (Dixon et al., 2018), we completed equivalence testing of the predicted values in comparison to the reference measure mean. We then evaluated an equivalence region of ±10% of the activPAL mean (0.94 hrs/d) and calculated the mean difference between predicted values and the activPAL values. Next, the 90% confidence intervals (90%CI) for the mean differences were calculated. Here the 90%CI of the mean differences that are fully inside the equivalence region are statistically equivalent to the reference measure at $\alpha = 0.05$ (Dixon et al., 2018). Predicted values with 90%CI mean difference values that overlap or are outside of the equivalence region are not equivalent to the reference. We also calculated p-values for two one-sided tests (TOST) for non-equivalence ($p < 0.05$ indicates equivalence) using the R package *equivalence* (version 0.7.2).

We also plotted the relationships between the predicted values from the algorithms (y-axis) and the activPAL (x-axis) and used linear regression to describe the amount of variance explained (R^2) and the overall accuracy and precision of each method (root mean square error [RMSE]). In sensitivity analysis, we examined equivalence and regression results by sex and age (< 50, ≥ 50 years).

RESULTS

Eighty-four participants consented, 78 of them contributed some valid monitor data, and 71 had at least one valid day of observation from the wrist- and thigh-worn monitors (Supplementary Figure 1). Mean age was 44.6 years and 79% of participants were women. The group had a mean BMI of 25.4 kg/m^2, were well-educated, and most were married (Table 1). Within the waking day (mean 15.1 hours/d) activPAL recorded means of 9.38 hours/d sedentary (62% of day), 5.68 hours/d active, and 8,280 steps/d. Participants wore devices for a mean of 7.5 days.

Equivalence Testing

Figure 1 reports equivalence test results of accuracy (±10%) of each of the algorithms compared to the thigh-worn reference measure (see Supplementary Table 1 for more detail). Five of the 21 methods were found to be equivalent to the reference method (i.e., the 90%CI of the mean difference fell within the ±10% equivalence region; see also Supplementary Table 2). Two of these methods employed machine-learning based estimates (Trost Extended, OxWearables) and three employed cut-point based estimates (GGIR ENMO 40mg; Bakrania ENMO 32.6mg; Fraysse ENMOa 62.5mg). Thus, 16 of the 21 algorithms (76%) produced estimates of sedentary time that were more than 10% different from the reference mean, and 9 of 21 methods (43%) were biased by ~1.85 hours/d or more.

Equivalence results among women were consistent with those reported in Figure 1, while among men only the Bakrania ENMO 32.6 mg method remained equivalent in this smaller group (n=15) (Supplementary Table 3). In general, bias was about 0.5 hours/d larger in men compared to women for the methods found to be equivalent in the overall sample. Also, the Trost + Sed UP method was found to be equivalent in men but not women (Supplementary Table 3).

Equivalence results by age are reported in Supplementary Table 4 and we found Trost Extended and Bakrania ENMO 32.6 mg to be equivalent in adults < 50 and ≥ 50 years of age, and results for OxWearables, GGIR default ENMO 40mg, and Fraysse ENMOa 62.5 mg to be equivalent only in adults < 50 years of age. The Trost + Sed UP method was found to be equivalent in adults ≥ 50 years of age. The magnitude and direction of bias observed in sedentary time estimates derived from algorithms originally developed among older adults remained consistent with our overall results reported in Figure 1 for adults < 50 and ≥ 50 years of age.

Descriptive Results

Figures 2, 3 and 4 describe the relationships and linear regression results for the different algorithms (y-axis) relative to the reference measure (x-axis). Values above the line of identity indicate underestimates of sedentary time for an algorithm and values below the line indicate overestimates. For the machine-learning based algorithms (Figure 2) the variance explained (R^2) ranged from 0.44 to 0.63 across methods and the root mean square error (RMSE) values ranged from 1.09 to 1.33 hrs/d. The OxWearables method had the highest R^2 (0.63) and lowest RMSE, 1.09 hrs/d among these methods. Figure 3 describes results for the algorithms from studies of younger adults. R^2 values ranged from 0.30 to 0.64 and RMSE values ranged from 1.07 to 1.49 hrs/d. The Bakrania ENMO 32.6 mg and GGIR default ENMO 40 mg values were more centered around line of identity, and the Bakrania MAD 42.4 mg and Esliger 45.0 mg cut-points had the higher R^2 (0.64 and 0.60) and lower RMSE values (1.07 and 1.12 hrs/d) among these methods. Figure 4 describes results for the algorithms developed among older adults. R^2 values ranged from 0.45 to 0.60 and RMSE values ranged from 1.12 to 1.32. The Fraysse ENMOa 62.5 mg cut-point had an R^2 of 0.58 and an RMSE of 1.15 hrs/d.

Table 2 reports mean values and streamlined results from linear regression by sex and age (< 50 and ≥ 50 years). In general, R^2 values tended to be higher and RMSE values lower for men compared to women, although the small number of men (n=15) suggests that this comparison should be interpreted cautiously. Notably, in detailed analysis our results for the Ellis machine-learning method that was developed in overweight women were no better when examined in women with BMI < 25 (mean bias +1.87 hrs/d; R^2=0.53; RMSE=1.37) or BMI ≥ 25 (mean bias +1.70 hrs/d; R^2=0.33; RMSE=1.23). In terms of age, R^2 values tended to be higher and RMSE values lower for participants < 50 compared to those who were ≥ 50 years of age. This was true for methods employing machine-learning and cut-point based methods and for methods developed among younger and older adults (Table 2). We also examined results stratified by BMI (<25, >=25 kg/m^2) and by activPAL step counts (median split). We found broadly comparable results for differences in mean sedentary time and RMSE values in the different strata but slightly lower R^2 values for the higher BMI and lower step count groups (Supplementary Table 5).

Study Design Elements in the Algorithm Development Studies

We next examined characteristics of the study designs used to develop the included classification algorithms. Ten algorithms were developed using oxygen consumption as the reference measure and 11 used behavioral information related to body position (Supplementary Table 1). Of the methods found to be equivalent in our field-based testing (Figure 1), four of five (80%) were originally calibrated to behavioral/body posture measures (Trost Extended, OxWearables, GGIR default ENMO 40mg, Bakrania ENMO 32.6mg). If we combine equivalent and borderline equivalent methods (i.e., 90%CI for mean difference overlapped the equivalence interval), six of 11 (55%) of the methods using behavioral metrics were this accurate, while only two in 10 (20%) of the methods using oxygen consumption were in this range (Supplementary Figure 2). In terms of the settings in which the activities were performed in the calibration studies, four of five (80%) of the machine-learning methods used uncontrolled free-living information to train the models and three (60%) of the methods (Trost Extended, OxWearables, Trost + SedUP) were at least borderline equivalent to our field-based reference measure (Supplementary Figure 3). All the studies which developed cut points employed either Laboratory or Controlled free-living activities, and five of 16 (31%) of the methods were at least borderline equivalent (Hildebrand ENMO 45.8mg; GGIR default ENMO 40mg; Bakrania ENMO 32.6mg; Esliger ENMOa 45mg; Fraysse ENMOa 62,5mg) in our evaluation.

DISCUSSION

We evaluated 21 different algorithms to estimate daily sedentary time from wrist-worn sensors in 71 adults over a mean 7.5 days of free-living activity. While 15 of the 21 algorithms (71%) were moderate-to-strongly correlated with our reference measure (R^2 > 0.5), only five (24%) were found to provide accurate estimates of mean sedentary time. This suggests that most algorithms tested, given their relatively high correlations, can provide useful insights in association studies of disease risk and correlates of behavior given to their ability to identify those with lower and higher levels of sedentary time. However, the large mean differences from one algorithm to the next suggests that quantification the amount of sedentary time associated with poor health could vary substantially depending on the algorithm used. Accurately quantifying the amount of behavior on the x-axis of the dose-response curve for physical activity and health associations is essential for public health translation. By rigorously evaluating a broad range of algorithms in an independent free-living study population, our findings point to several methods that are more accurate and precise, and possibly several study design elements used for algorithm development that may yield more robust estimates of sedentary time.

Even though at least 21 algorithms to estimate sedentary time have appeared in the literature, it is surprising and worrisome that so few have been tested in free-living studies. The Hildebrand cut point (ENMO 45.8mg) is one of the older methods and it has been evaluated in daily life. In the original development study, it was found to overestimate sedentary time by 34% vs. activPAL in free-living conditions (Hildebrand et al., 2017). In a short-duration free-living direct observation study (Marcotte et al., 2020) the method was found to overestimate sedentary time by ~20%. We found similar results, with a smaller overestimate of ~12%. The consistency of these three results suggests that the Hildebrand ENMO 45.8 mg overestimates sedentary time by 12 to 34%. It may be instructive that just three free-living studies using strong reference measures begins to outline the performance of this widely used algorithm. This information should inform dose-response results from studies that use this method. The origin of the GGIR default ENMO 40 mg threshold is not immediately clear, but if it was an adaptation of the published Hildebrand cut point the lowering of the threshold resulted in more accurate estimates in our study.

We are the first to evaluate the Bakrania ENMO 32.6 mg threshold found to be equivalent here, but Marcotte and colleagues (Marcotte et al., 2020) examined the ENMO 27.9 mg threshold that was identified for differentiating sitting from dusting activity in Bakrania (Bakrania et al., 2016) and found it to be ~10% lower than directly observed sedentary time. Suorsa and colleagues (Suorsa et al., 2020) examined an ENMO 30 mg threshold and found it to be 9% lower than estimates from a thigh worn device, with a Pearson r of 0.62 (R^2 = 0.38). We found the slightly higher ENMO 32.6 mg threshold calibrated to two common daily activities (vs. sitting) to be within 3% (−0.29 hrs/d) of our reference with a relatively high R^2 of 0.57. This threshold may be promising as simple approach to estimating sedentary time, but additional work is needed to confirm our findings.

Outside of the original development studies for the OxWearables (Walmsley et al., 2022) and Trost Original (Pavey et al., 2017) algorithms we are unaware of other studies that evaluated these methods. OxWearables was initially evaluated using leave one out cross-validation and camera images as a reference measure and found estimates of sedentary time to be 86% accurate (Walmsley et al., 2022) and these results are broadly consistent with the present study. We found the Trost Extended method, which adds sleep and nonwear detection information (Ahmadi et al., 2020) to the Trost Original algorithm to be more accurate than Trost Original (Pavey et al., 2017) alone. We believe that this is the first independent test of the Fraysse cut points (Fraysse et al., 2020) developed in older adults.

Implications for Future Algorithm Development and Testing

There were a wide range of study design/protocols used to develop of the 21 classification algorithms examined here which enabled a qualitative evaluation of study design elements associated with better algorithm performance. Of the five algorithms we found to be equivalent, four were developed using behavioral/body posture-based measures as a reference rather than oxygen consumption. Fifty five percent of algorithms that were developed using behavior/posture-based reference measures were found to be more accurate (borderline equivalent or greater), while only 20% of the algorithms developed using oxygen consumption alone were similarly accurate. Given the relevance of body posture in classifying sedentary behavior, body posture should be incorporated into future algorithm development studies. Similarly, 60% of the algorithms developed in the setting of free-living activities were at least borderline accurate compared to 31% of algorithms developed in laboratory and controlled free-living settings, highlighting the importance of algorithm development in naturalistic environments (Keadle et al., 2019).

Our qualitative examination also demonstrated the importance of verifying the performance of classification algorithms to ensure that they work as expected in new study populations. The procedures that have been developed for standardizing/harmonizing clinical assays (e.g., blood cholesterol, hormones) involves three basic steps (Vesper et al., 2016), the last of which is to verify the performance of new assays in target patient populations. This discipline has found that even state-of-the-art assay development methods can fail when applied to target clinical populations (Vesper et al., 2016). Our study revealed at least three unexpected results that would not have been identified without testing. First, two of the machine-learning algorithms (OxWearables, Ellis) were both developed in free-living populations using the same reference measure (i.e., camera images), but the OxWearables algorithm performed better in our testing. Second, we found that two cut points derived from the exact same study protocol but using different acceleration metrics (Bakrania ENMO 32.6 mg, Bakrania MAD 42.4 mg) produced different results in our equivalence testing. Third, we employed the non-dominant wrist location to test some algorithms developed for the dominant and right wrists. Our qualitative review showed large differences in cut-points for each wrist location for some study protocols (Esliger, Dillon, Fraysse) but not others (Duncan, Miqueles), suggesting the potential for a protocol effect vs. a wrist location effect. We also found the Fraysse 62.5 mg ENMOa cut point developed for the dominant wrist to be equivalent to the activPAL when applied using the non-dominant wrist location. It is not clear to us what may explain these unexpected findings, but our independent testing using a common benchmark reference measure in free-living adults points to important areas for future investigation.

The limited amount of independent testing of the many algorithms proposed to estimate sedentary time using wrist-worn sensors is consistent with the more general problem of lack of rigorous testing of methods in the field of ambulatory monitoring. A recent review of free-living validation studies of 163 wearable sensors found that about 60% of the methods proposed for use were only tested once, and virtually all the free-living validation testing that was done was of lower quality (Giurgiu et al., 2022). To more completely understand the accuracy and precision of sensor-based measure of physically active and sedentary behaviors there is a need for additional large benchmarking studies like this one using a common evaluation framework (Keadle et al., 2019) to build an evidence base that identifies accurate and precise algorithms for estimating sedentary time and overall physical activity (Giurgiu et al., 2022). Data from this study is available on request.

This performance evaluation study has several strengths including a relatively large number of middle-aged adults studied in a 10-day free-living protocol. The reference measure (activPAL) has been shown to be accurate and precise for estimating sedentary time in comparison to direct observation, a gold standard measure, in free-living populations. Use of this reference measure assumes that sitting and reclining behaviors involving upper body activity that substantially increase energy expenditure are relatively rare (e.g., weightlifting, arm ergometry), with only small impacts on average sedentary time in the overall population. We also examined many algorithms that enabled both the identification of more accurate methods and provided some insight into the study design elements associated with algorithms that produced more accurate daily estimates. Weaknesses of our study included enrollment of generally healthy adults and only a small number of men and older adults. This may reduce the generalizability of our findings, particularly with respect to understanding sex differences and the performance of algorithms that were developed for adults over 70 years of age. It would be interesting to begin to test the hypothesis that age-specific algorithms to estimate sedentary time are needed utilizing free-living validation studies such as this but with a wider age distribution. There were several algorithms for wrist data that we did not evaluate. When this project was initiated, code for generating activity counts that would allow evaluation of ActiGraph-based algorithms (Montoye et al., 2020) was less readily available, limiting the inclusion of these methods in this report. Also, we were unable to reliably implement the Sedentary Sphere algorithm which has shown strong performance previously vs. activPAL (Pavey et al., 2016; Rowlands et al., 2016). However, our evaluation of the Trost + SedUp algorithm provides some insight into the ability of wrist orientation to estimate sedentary time. Also, we focused only on algorithms to estimate sedentary time here because we did not have a strong reference measure for more refined estimates of physical activity (e.g., moderate-vigorous intensity activity). Future studies are needed to confirm our results by sex and in older adults. Last, we studied adults without major mobility impairments or significant comorbidities, and this should be taken into consideration when interpreting our findings.

CONCLUSION

Wrist-worn sensors are increasingly being used in epidemiologic studies the results of which will inform future public health recommendations, worldwide surveillance efforts, and health promotion programs in community and clinical settings. To date, there has been limited information about the accuracy of sedentary time estimates derived from wrist worn sensors, which is surprising because understanding the accuracy of the algorithms is essential for translating findings from etiologic health studies regarding the amount of sedentary time associated with poor health. Estimating sedentary time accurately is also essential for the accurate estimation of light intensity physical activity. To move beyond general advice to “sit less” toward providing specific recommendations for the amounts of sitting time associated with increased risk, measurement methods that produce consistently accurate measurements are needed. Last, we hope that this large benchmarking study informed by a strong framework for evaluation (Keadle et al., 2019) will serve as a model for future more comprehensive efforts that will accelerate our ability to identify more accurate and precise measures of sedentary time and physical activity using wrist-worn sensors (Giurgiu et al., 2022).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

Results from this study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation. The results of this research represent the views of the authors alone and not the institutions within which they are employed.

REFERENCES

Ahmadi MN, Clare PJ, Katzmarzyk PT, Del Pozo Cruz B, Lee IM, & Stamatakis E (2022). Vigorous physical activity, incident heart disease, and cancer: how little is enough? European Heart Journal, 43(46), 4801–4814. 10.1093/eurheartj/ehac572 [PubMed: 36302460]

Ahmadi MN, Nathan N, Sutherland R, Wolfenden L, & Trost SG (2020). Non-wear or sleep? Evaluation of five non-wear detection algorithms for raw accelerometer data. Journal of Sports Sciences, 38(4), 399–404. 10.1080/02640414.2019.1703301 [PubMed: 31826746]

American Cancer Society. (2024). Cancer Prevention Study-3: Accelerometry Substudy. Retrieved June 4, 2024 from https://www.cancer.org/research/cps3-cancer-prevention-study-3.html

Bakrania K, Yates T, Rowlands AV, Esliger DW, Bunnewell S, Sanders J, Davies M, Khunti K, & Edwardson CL (2016). Intensity Thresholds on Raw Acceleration Data: Euclidean Norm Minus One (ENMO) and Mean Amplitude Deviation (MAD) Approaches. PloS One, 11(10), e0164045. 10.1371/journal.pone.0164045 [PubMed: 27706241]

Chastin SF, Egerton T, Leask C, & Stamatakis E (2015). Meta-analysis of the relationship between breaks in sedentary behavior and cardiometabolic health. Obesity (Silver Spring), 23(9), 1800–1810. 10.1002/oby.21180 [PubMed: 26308477]

Dillon CB, Fitzgerald AP, Kearney PM, Perry IJ, Rennie KL, Kozarski R, & Phillips CM (2016). Number of Days Required to Estimate Habitual Activity Using Wrist-Worn GENEActiv Accelerometer: A Cross-Sectional Study. PloS One, 11(5), e0109913. 10.1371/journal.pone.0109913 [PubMed: 27149674]

Dixon PM, Saint-Maurice PF, Kim Y, Hibbing P, Bai Y, & Welk GJ (2018). A Primer on the Use of Equivalence Testing for Evaluating Measurement Agreement. Medicine and Science in Sports and Exercise, 50(4), 837–845. 10.1249/mss.0000000000001481 [PubMed: 29135817]

Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, White T, van Hees VT, Trenell MI, Owen CG, Preece SJ, Gillions R, Sheard S, Peakman T, Brage S, & Wareham NJ (2017). Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. PloS One, 12(2), e0169649. 10.1371/journal.pone.0169649 [PubMed: 28146576]

Duncan MJ, Rowlands A, Lawson C, Leddington Wright S, Hill M, Morris M, Eyre E, & Tallis J (2020). Using accelerometry to classify physical activity intensity in older adults: What is the optimal wear-site? Eur J Sport Sci, 20(8), 1131–1139. 10.1080/17461391.2019.1694078 [PubMed: 31726952]

Ellis K, Kerr J, Godbole S, Staudenmayer J, & Lanckriet G (2016). Hip and Wrist Accelerometer Algorithms for Free-Living Behavior Classification. Medicine and Science in Sports and Exercise, 48(5), 933–940. 10.1249/mss.0000000000000840 [PubMed: 26673126]

Esliger DW, Rowlands AV, Hurst TL, Catt M, Murray P, & Eston RG (2011). Validation of the GENEA Accelerometer. Medicine and Science in Sports and Exercise, 43(6), 1085–1093. 10.1249/MSS.0b013e31820513be [PubMed: 21088628]

Fernandez-Mendoza J (2019). Insomnia and cardiometabolic disease risk. In Grandner MA (Ed.), Sleep and Health (pp. 391–407). Academic Press.

Fraysse F, Post D, Eston R, Kasai D, Rowlands AV, & Parfitt G (2020). Physical Activity Intensity Cut-Points for Wrist-Worn GENEActiv in Older Adults. Front Sports Act Living, 2, 579278. 10.3389/fspor.2020.579278 [PubMed: 33521631]

Fuller D, Ferber R, & Stanley K (2022). Why machine learning (ML) has failed physical activity research and how we can improve. BMJ Open Sport & Exercise Medicine, 8(1), e001259. 10.1136/bmjsem-2021-001259

The German National Cohort: aims, study design and organization. (2014). European Journal of Epidemiology, 29(5), 371–382. 10.1007/s10654-014-9890-7 [PubMed: 24840228]

Giurgiu M, Timm I, Becker M, Schmidt S, Wunsch K, Nissen R, Davidovski D, Bussmann JBJ, Nigg CR, Reichert M, Ebner-Priemer UW, Woll A, & von Haaren-Mack B (2022). Quality Evaluation of Free-living Validation Studies for the Assessment of 24-Hour Physical Behavior in Adults via Wearables: Systematic Review. JMIR Mhealth Uhealth, 10(6), e36377. 10.2196/36377 [PubMed: 35679106]

Hildebrand M, Hansen BH, van Hees VT, & Ekelund U (2017). Evaluation of raw acceleration sedentary thresholds in children and adults. Scandinavian Journal of Medicine and Science in Sports, 27(12), 1814–1823. 10.1111/sms.12795 [PubMed: 27878845]

Katzmarzyk PT, Powell KE, Jakicic JM, Troiano RP, Piercy K, & Tennant B (2019). Sedentary Behavior and Health: Update from the 2018 Physical Activity Guidelines Advisory Committee. Medicine and Science in Sports and Exercise, 51(6), 1227–1241. 10.1249/mss.0000000000001935 [PubMed: 31095080]

Keadle SK, Lyden K, Strath S, Staudenmayer JW, & Freedson PS (2019). Perspectives for Progress - A Framework to Evaluate Devices that Assess Physical Behavior. Exercise and Sport Sciences Reviews, *Publish Ahead of Print*. 10.1249/jes.0000000000000206

Liu F, Wanigatunga AA, & Schrack JA (2022). Assessment of Physical Activity in Adults Using Wrist Accelerometers. Epidemiologic Reviews, 43(1), 65–93. 10.1093/epirev/mxab004 [PubMed: 34215874]

Marcotte RT, Petrucci GJ Jr., Cox MF, Freedson PS, Staudenmayer JW, & Sirard JR (2020). Estimating Sedentary Time from a Hip- and Wrist-Worn Accelerometer. Medicine and Science in Sports and Exercise, 52(1), 225–232. 10.1249/mss.0000000000002099 [PubMed: 31343523]

Master H, Annis J, Huang S, Beckman JA, Ratsimbazafy F, Marginean K, Carroll R, Natarajan K, Harrell FE, Roden DM, Harris P, & Brittain EL (2022). Association of step counts over time with the risk of chronic disease in the All of Us Research Program. Nature Medicine, 28(11), 2301–2308. 10.1038/s41591-022-02012-w

Matthews CE, Carlson SA, Saint-Maurice PF, Patel S, Salerno EA, Loftfield E, Troiano RP, Fulton JE, Sampson JN, Tribby C, Keadle SK, & Berrigan D (2021). Sedentary Behavior in U.S. Adults: Fall 2019. Medicine and Science in Sports and Exercise, 53(12), 2512–2519. 10.1249/mss.0000000000002751 [PubMed: 34310489]

Migueles JH, Cadenas-Sanchez C, Alcantara JMA, Leal-Martín J, Mañas A, Ara I, Glynn NW, & Shiroma EJ (2021). Calibration and Cross-Validation of Accelerometer Cut-Points to Classify Sedentary Time and Physical Activity from Hip and Non-Dominant and Dominant Wrists in Older Adults. Sensors (Basel, Switzerland), 21(10). 10.3390/s21103326

Migueles JH, Rowlands AV, Huber F, Sabia S, & van Hees VT (2019). GGIR: A Research Community–Driven Open Source R Package for Generating Physical Activity and Sleep Outcomes From Multi-Day Raw Accelerometer Data. Journal for the Measurement of Physical Behaviour, 2(3), 188–196. 10.1123/jmpb.2018-0063

Montoye AHK, Clevenger KA, Pfeiffer KA, Nelson MB, Bock JM, Imboden MT, & Kaminsky LA (2020). Development of cut-points for determining activity intensity from a wrist-worn ActiGraph accelerometer in free-living adults. Journal of Sports Sciences, 38(22), 2569–2578. 10.1080/02640414.2020.1794244 [PubMed: 32677510]

O'Brien MW, Wu Y, Petterson JL, Bray NW, & Kimmerly DS (2022). Validity of the ActivPAL monitor to distinguish postures: A systematic review. Gait and Posture, 94, 107–113. 10.1016/j.gaitpost.2022.03.002 [PubMed: 35276456]

Pavey TG, Gilson ND, Gomersall SR, Clark B, & Trost SG (2017). Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. Journal of Science and Medicine in Sport, 20(1), 75–80. 10.1016/j.jsams.2016.06.003 [PubMed: 27372275]

Pavey TG, Gomersall SR, Clark BK, & Brown WJ (2016). The validity of the GENEActiv wrist-worn accelerometer for measuring adult sedentary time in free living. Journal of Science and Medicine in Sport, 19(5), 395–399. 10.1016/j.jsams.2015.04.007 [PubMed: 25956687]

Piercy KL, Troiano RP, Ballard RM, & et al. (2018). The physical activity guidelines for americans. JAMA, 320(19), 2020–2028. 10.1001/jama.2018.14854 [PubMed: 30418471]

Rowlands AV, Yates T, Olds TS, Davies M, Khunti K, & Edwardson CL (2016). Wrist-Worn Accelerometer-Brand Independent Posture Classification. Medicine and Science in Sports and Exercise, 48(4), 748–754. 10.1249/mss.0000000000000813 [PubMed: 26559451]

Saint-Maurice P, Troiano RP, Matthews CE, & Kraus WE (2018). Moderate-Vigorous Intensity Physical Activity and All-Cause Mortality: Do Bouts Matter?. Journal of the American Heart Association, 7(6), e007678. [PubMed: 29567764]

Sanders GJ, Boddy LM, Sparks SA, Curry WB, Roe B, Kaehne A, & Fairclough SJ (2019). Evaluation of wrist and hip sedentary behaviour and moderate-to-vigorous physical activity raw acceleration cutpoints in older adults. Journal of Sports Sciences, 37(11), 1270–1279. 10.1080/02640414.2018.1555904 [PubMed: 30558487]

Straczkiewicz M, Glynn NW, Zipunnikov V, & Harezlak J (2020). Fast and Robust Algorithm for Detecting Body Posture Using Wrist-Worn Accelerometers. Journal for the Measurement of Physical Behaviour, 3(4), 285–293. 10.1123/jmpb.2019-0067

Suorsa K, Pulakka A, Leskinen T, Pentti J, Holtermann A, Heinonen OJ, Sunikka J, Vahtera J, & Stenholm S (2020). Comparison of Sedentary Time Between Thigh-Worn and Wrist-Worn Accelerometers. Journal for the Measurement of Physical Behaviour, 3(3), 234–243. 10.1123/jmpb.2019-0052

Vesper HW, Myers GL, & Miller WG (2016). Current practices and challenges in the standardization and harmonization of clinical laboratory tests. American Journal of Clinical Nutrition, 104 Suppl 3(Suppl 3), 907s–912s. 10.3945/ajcn.115.110387 [PubMed: 27534625]

Walmsley R, Chan S, Smith-Byrne K, Ramakrishnan R, Woodward M, Rahimi K, Dwyer T, Bennett D, & Doherty A (2022). Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease. British Journal of Sports Medicine, 56(18), 1008–1017. 10.1136/bjsports-2021-104050

Author Manuscript

Author Manuscript

Author Manuscript

HIGHLIGHTS

1. Real-world testing of classification algorithms for wrist-worn sensors to estimate sedentary time against strong reference measures in adults is limited.
2. Although all the wrist-based algorithms evaluated were moderate to strongly correlated with the reference measure of sedentary time, only five of 21 (24%) algorithms were equivalent to the reference measure (±10%).
3. Algorithm development studies conducted in uncontrolled/free-living settings or that employed posture-based criterion measures tended to produce more accurate classification algorithms.

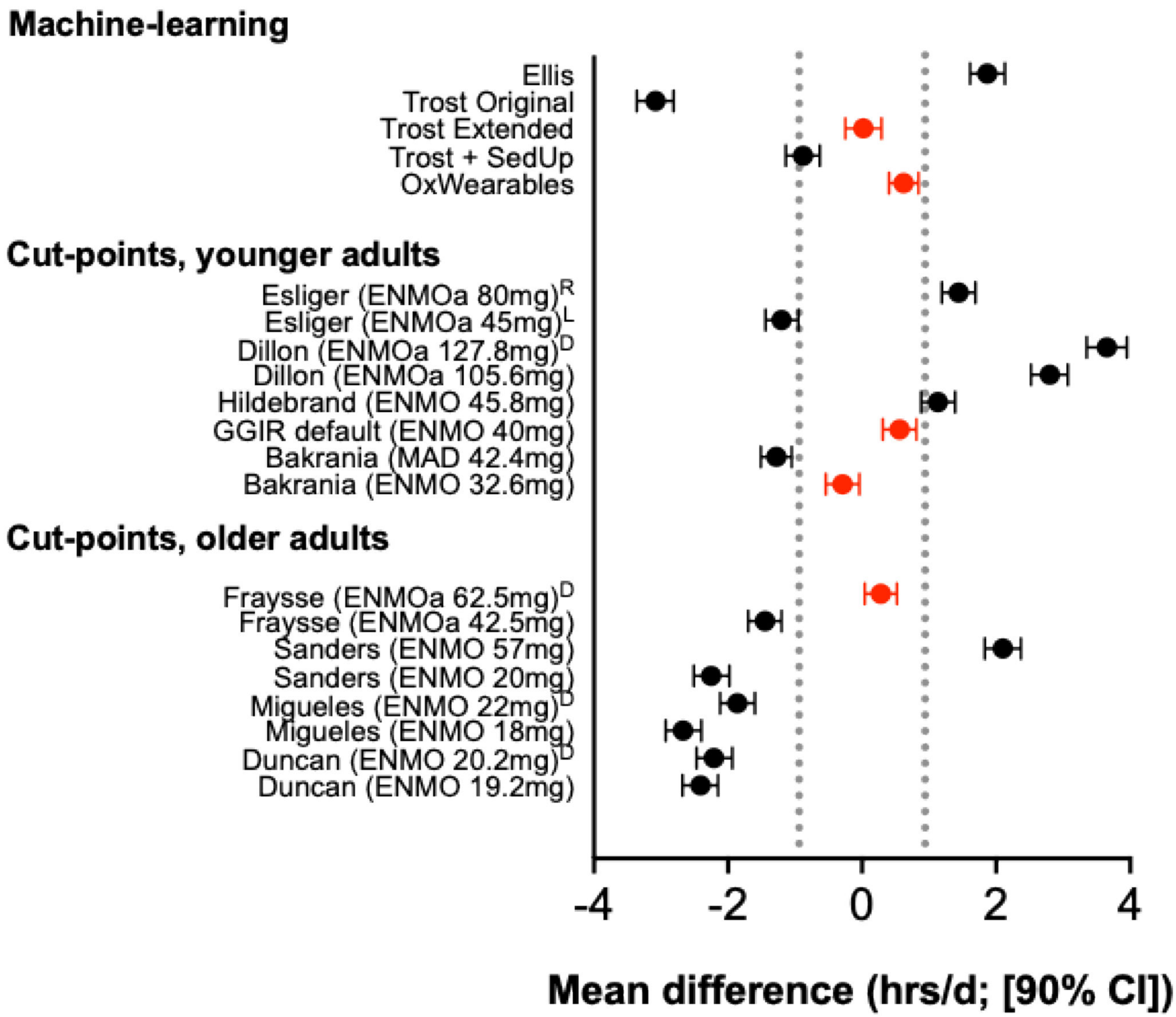

Figure 1.

Equivalence testing for each method vs. activPAL (±10% of mean).

Values within dotted line (in red) are equivalent to reference method. Wear location was non-dominant wrist unless indicated (L=left, R=right; D=dominant).

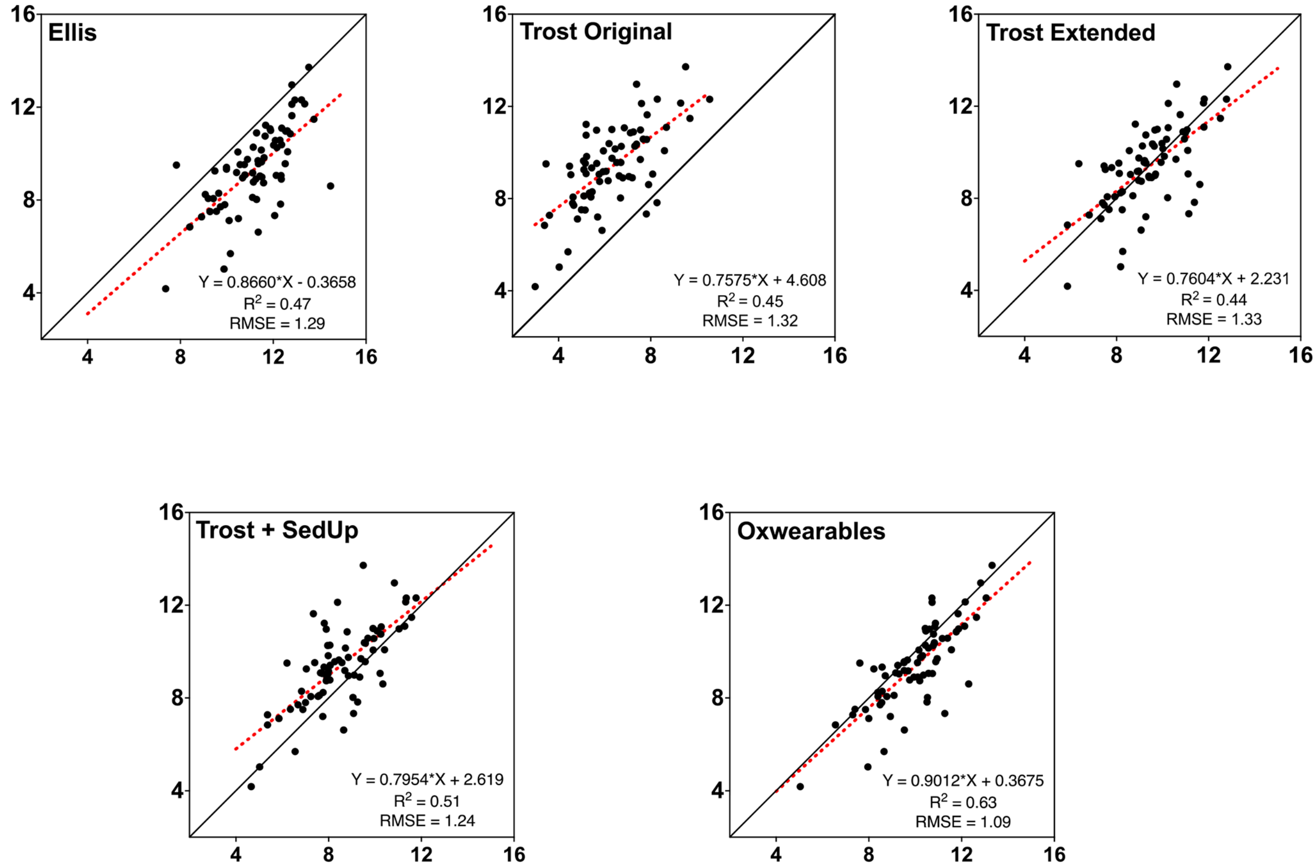

Figure 2.

Machine-learning based methods (y-axis) vs. activPAL (x-axis), including linear regression results (N=71).

Values are hrs/d.

Solid line = line of identity; red dotted line = regression prediction line. Each method used the non-dominant wrist.

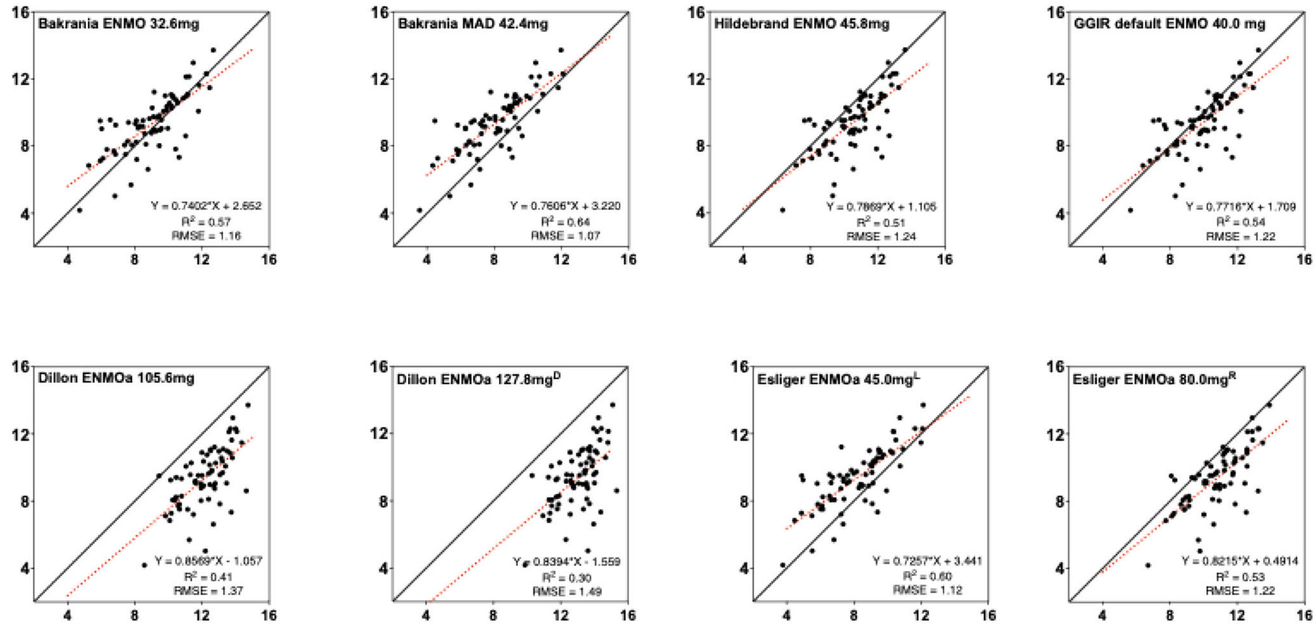

Figure 3.

Cut point-based methods developed with younger adults (y-axis) vs. activPAL (x-axis), including linear regression results (N=71). Values are hrs/d.

Solid line = line of identity; red dotted line = regression prediction line. Wear location other than non-dominant wrist indicated (D=dominant; L=left, R=right wrist).

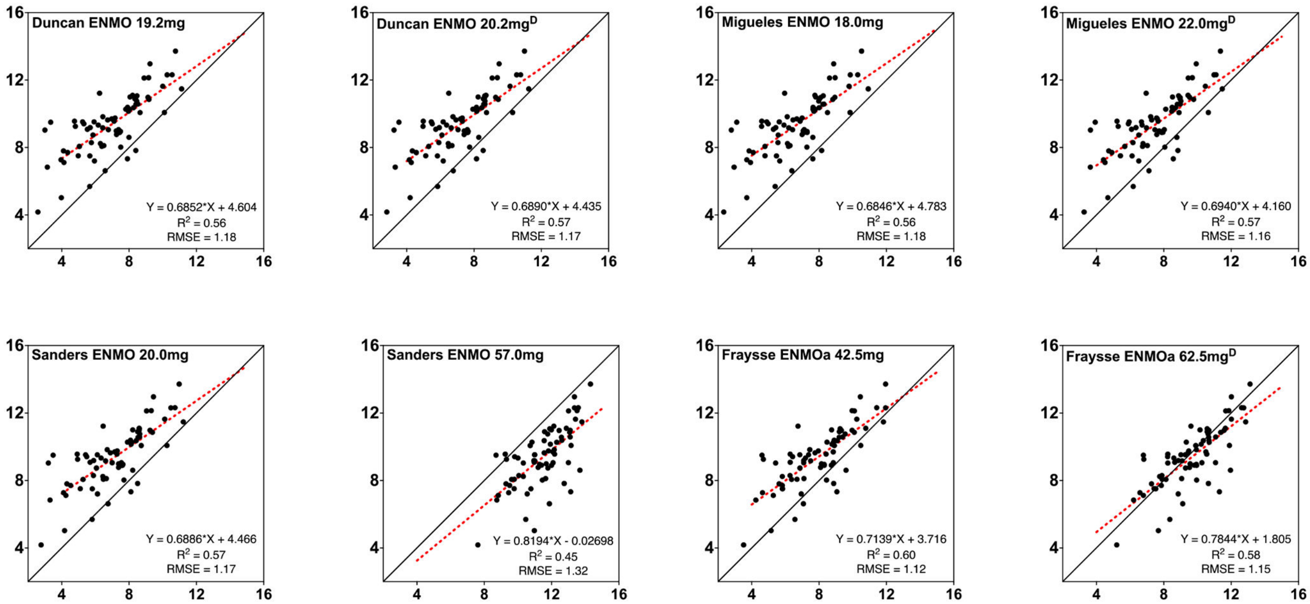

Figure 4.

Cut point-based methods developed with older adults (y-axis) vs. activPAL (x-axis), including linear regression results (N=71).

Values are hrs/d.

Solid line = line of identity; red dotted line = regression prediction line. Wear location was non-dominant wrist unless indicated (D=dominant).

Table 1.

Description of study population

Variable	All (n=71)	Women (n=56)	Men (n=15)
Age (mean, SD)	44.6 (10.6)	44.9 (10.9)	43.6 (9.4)
Age Group			
Less than 40	28 (39%)	21 (38%)	7 (47%)
40 – 50	15 (21%)	12 (21%)	3 (20%)
50 – 60	22 (31%)	18 (32%)	4 (26%)
60 and older	6 (9%)	5 (9%)	1 (7%)
Body mass index (mean, SD)	25.4 (3.9)	25.0 (3.9)	26.8 (3.7)
Body mass index 30+ (n, percent)	11 (15%)	7 (13%)	4 (27%)
Comorbidities, 1 or more (n, percent)	19 (27%)	15 (27%)	4 (27%)
Education			
1–3 years college/Associate’s	11 (16%)	10 (18%)	1 (7%)
College graduate	35 (49%)	26 (46%)	9 (60%)
Master’s degree	19 (27%)	16 (29%)	3 (20%)
PhD or equivalent	6 (8%)	4 (7%)	2 (13%)
Marital Status			
Married	48 (68%)	37 (66%)	11 (74%)
Partnered	9 (13%)	7 (12%)	2 (13%)
Single	8 (11%)	6 (11%)	2 (13%)
Divorced/Separated	6 (8%)	6 (11%)	0
Race			
White	54 (76%)	44 (78%)	10 (67%)
Black/African American	0	0	0
American Indian/Alaskan Native	3 (4%)	1 (2%)	2 (13%)
Asian	3 (4%)	2 (4%)	1 (7%)
Native Hawaiian/Pacific Islander	1 (2%)	1 (2%)	0
Mixed race	8 (11%)	6 (11%)	2 (13%)
Other/Declined to answer	2 (3%)	2 (4%)	0
Ethnicity			
Hispanic or Latino	17 (24%)	13 (23%)	4 (27%)
Not Hispanic or Latino	54 (76%)	43 (77%)	11 (73%)
Reference Measure (activPAL)			
Sedentary time (hours/d)	9.38 (1.8)	9.38 (1.8)	9.36 (1.7)
Active time (hours/d)	5.68 (1.9)	5.63 (1.9)	5.85 (1.8)
Steps (per day)	8,280 (3,708)	8,188 (3,911)	8,626 (2,912)
Waking day (hours/d)	15.1 (0.8)	15.0 (0.8)	15.2 (0.5)
Valid days (number)	7.5 (2.2)	7.5 (2.1)	7.5 (2.5)

Table 2.

Sedentary time (hrs/d) for each algorithm and summary regression results, by sex and age

		Women (n=56)			Men (n=15)			< 50 years (n=43)			≥ 50 years (n=28)		
Method (Published)	**Wrist Position**	**Mean (SD)**	**R^2**	**RMSE**	**Mean (SD)**	**R^2**	**RMSE**	**Mean (SD)**	**R^2**	**RMSE**	**Mean (SD)**	**R^2**	**RMSE**
activPAL (ref)	ND	9.38 (1.82)	-	-	9.36 (1.73)	-	-	9.49 (1.68)	-	-	9.20 (1.96)	-	-
Machine Learning Methods													
Ellis	ND	11.17 (1.49)	0.46	1.32	11.53 (1.16)	0.59	1.07	11.22 (1.36)	0.50	1.18	11.29 (1.55)	0.46	1.41
Trost Original	ND	6.18 (1.60)	0.45	1.34	6.73 (1.45)	0.50	1.18	6.28 (1.49)	0.50	1.17	6.31 (1.74)	0.39	1.50
Trost Extended	ND	9.26 (1.59)	0.44	1.35	9.89 (1.43)	0.55	1.11	9.26 (1.50)	0.50	1.18	9.60 (1.67)	0.43	1.46
Trost + SedUp	ND	8.32 (1.58)	0.51	1.26	9.13 (1.65)	0.66	0.97	8.40 (1.45)	0.37	1.32	8.64 (1.85)	0.72	1.01
OxWearables	ND	9.90 (1.60)	0.62	1.12	10.34 (1.46)	0.72	0.88	9.95 (1.46)	0.63	1.02	10.07 (1.75)	0.65	1.15
Younger Adults													
Bakrania (ENMO 32.6mg)	ND	8.98 (1.87)	0.56	1.20	9.49 (1.66)	0.70	0.92	8.99 (1.73)	0.64	1.00	9.22 (2.00)	0.53	1.33
Bakrania (MAD, 42.4mg)	ND	7.97 (1.90)	0.63	1.10	8.54 (1.79)	0.73	0.87	8.05 (1.78)	0.69	0.93	8.16 (2.04)	0.59	1.23
Hildebrand (ENMO 45.8mg)	ND	10.42 (1.66)	0.50	1.28	10.85 (1.49)	0.64	1.00	10.41 (1.60)	0.58	1.08	10.66 (1.69)	0.47	1.41
GGIR Default (ENMO 40mg)	ND	9.84 (1.74)	0.53	1.24	10.29 (1.57)	0.66	0.97	9.84 (1.66)	0.61	1.04	10.08 (1.80)	0.50	1.37
Dillon (ENMOa 105.6mg)	ND	12.12 (1.37)	0.39	1.41	12.37 (1.23)	0.52	1.16	12.10 (1.31)	0.49	1.19	12.30 (1.39)	0.35	1.56
Dillon (ENMOa 127.8mg)	D	12.99 (1.20)	0.28	1.53	13.15 (1.02)	0.38	1.31	12.96 (1.13)	0.39	1.30	13.14 (1.23)	0.22	1.70
Esliger (ENMOa 45mg)	Left	8.05 (1.93)	0.59	1.15	8.65 (1.84)	0.72	0.88	8.09 (1.82)	0.64	0.99	8.32 (2.09)	0.59	1.23
Esliger (ENMOa 80mg)	Right	10.74 (1.60)	0.51	1.26	11.10 (1.50)	0.62	1.02	10.72 (1.55)	0.59	1.07	10.95 (1.63)	0.49	1.38
Older Adults (55+)													
Duncan (ENMO 19.2mg)	ND	6.82 (2.00)	0.55	1.21	7.50 (1.73)	0.71	0.91	6.90 (1.78)	0.63	1.01	7.05 (2.24)	0.51	1.35
Duncan (ENMO 20.2mg)	D	7.02 (2.00)	0.55	1.21	7.72 (1.74)	0.72	0.89	7.11 (1.78)	0.64	1.00	7.27 (2.23)	0.52	1.34
Migueles (ENMO 18mg)	ND	6.56 (1.99)	0.55	1.22	7.25 (1.73)	0.69	0.93	6.66 (1.77)	0.63	1.02	6.79 (2.23)	0.50	1.36
Migueles (ENMO 22mg)	D	7.37 (1.99)	0.56	1.20	8.05 (1.76)	0.72	0.88	7.44 (1.79)	0.65	0.99	7.62 (2.21)	0.52	1.34
Sanders (ENMO 20mg)	ND	6.99 (2.00)	0.55	1.21	7.67 (1.74)	0.71	0.89	7.07 (1.78)	0.64	1.00	7.23 (2.23)	0.51	1.34
Sanders (ENMO 57mg)	ND	11.40 (1.5)	0.43	1.36	11.77 (1.31)	0.58	1.09	11.38 (1.44)	0.53	1.14	11.62 (1.52)	0.39	1.50
Fraysse (ENMOa 42.5mg)	ND	7.79 (1.95)	0.59	1.16	8.43 (1.87)	0.73	0.86	7.84 (1.83)	0.63	1.01	8.06 (2.12)	0.59	1.24
Fraysse (ENMOa 62.5mg)	D	9.56 (1.77)	0.57	1.18	9.99 (1.65)	0.66	0.97	9.56 (1.70)	0.63	1.01	9.80 (1.84)	0.57	1.27