Published in final edited form as:

Med Sci Sports Exerc. 2016 May ; 48(5): 951–957. doi:10.1249/MSS.0000000000000841.

Objective Assessment of Physical Activity: Classifiers for Public Health

Jacqueline Kerr[1], **Ruth E. Patterson**[1], **Katherine Ellis**[2], **Suneeta Godbole**[1], **Eileen Johnson**[1], **Gert Lanckriet**[2], and **John Staudenmayer**[3]

[1]Department of Family Medicine & Public Health, University of California, San Diego, CA

[2]Department of Computer Science & Engineering, University of California, San Diego, CA

[3]University of Massachusetts, Amherst, MA

Abstract

Purpose—Walking for health is recommended by health agencies, partly based on epidemiological studies of self-reported behaviors. Accelerometers are now replacing survey data but it is not clear that intensity based cut points reflect the behaviors previously reported. New computational techniques can help classify raw accelerometer data into behaviors meaningful for public health.

Methods—520 days of triaxial 30 hertz accelerometer data from 3 studies (n=78) were employed as training data. Study 1 included prescribed activities completed in natural settings. The other two studies included multiple days of free living data with SenseCam annotated ground truth. The two populations in the free living data sets were demographically and physical different. Random forest classifiers were trained on each data set and the classification accuracy on the training data set and applied to the other available data sets was assessed. Accelerometer cut points were also compared with the ground truth from the 3 datasets.

Results—The random forest classified all behaviors with over 80% accuracy. Classifiers developed on the prescribed data performed with higher accuracy than the free living data classifier, but did not perform as well on the free living datasets. Many of the observed behaviors occurred at different intensities than those identified by existing cut points.

Conclusions—New machine learned classifiers developed from prescribed activities (Study 1) were considerably less accurate when applied to free-living populations or to a functionally different population (Studies 2 & 3). These classifiers, developed on free living data, may have value when applied to large cohort studies with existing hip accelerometer data.

Keywords

Accelerometer; Measurement; Machine Learning; Walking; Sedentary Behavior

Corresponding Author: Jacqueline Kerr, PhD, Associate Professor, Center for Wireless & Population Health Systems, Department of Family Medicine & Public Health, 9500 Gilman Drive #0811, La Jolla, CA 92093-0811, Phone: (858) 534-9305, Fax (858) 534-9404, jkerr@ucsd.edu.

Conflicts of Interest

None of the authors have conflicts of interested and the results of the present study do not constitute endorsement by ACSM

Introduction

Over the past 20 years, there has been an exponential growth in knowledge regarding the role of physical activity (PA) in health promotion and disease prevention (3). For example, walking for health, is recommended in the national physical activity guidelines and by the Surgeon General, based on large epidemiological studies of self-reported behaviors. Accelerometers are now replacing survey data in large population studies due to concerns with inaccuracies and biases in self-report, but it is not clear that laboratory-derived intensity based cut points reflect the behaviors previously reported (32). For example, accelerometer based intensity cut points applied to the NHANES sample, indicate less than 10% of US adults, and only 3% of older adults, meet PA guidelines (31). Self-reported estimates, however, indicated that 30–50% of the population met guidelines (5). Further, accelerometer derived activities may not be related to health outcomes that are associated with self-reported activities (6). Accelerometer cut points are also known to underestimate certain behaviors, such as cycling, that may be related to the provision of safe built environments and could be key a behavior to target for population changes in active living (27). Large cohort studies are poised to analyze recently collected accelerometer data and impact public health guidelines. If intensity based cut points alone are applied, we will miss the opportunity to understand more about specific behaviors that can be communicated clearly to the public in health guidelines.

To develop and validate new methods for predicting PA behaviors, such as cycling or walking, the behaviors have to be known (i.e., ground truth is available). These data are then used to train the resulting classifier. The easiest way to observe a behavior is to prescribe and observe it in a laboratory setting. Prescribed behaviors have the advantage of ensuring that all behaviors of interest are captured and balanced, whereas free living data may not include behaviors of interest if they are infrequently performed in the general population. Although they are not performed, they may still be behaviors that would be targeted in an intervention. Further, laboratory trials and observational protocols have the advantage of using indirect calorimetry to estimate energy expenditure. However, it is doubtful that laboratory-based activities accurately reflect free-living behaviors across study samples that are older, obese, or have co-morbidities (1). Further, samples of prescribed activities even in free living do not reflect free living behaviors over multiple days in naturalistic settings.

The research presented here describes a mobile technology – wearable cameras – that allows observation of behaviors across multiple days in free living populations. We used machine learning techniques to develop models (i.e., classifiers) that predict PA behaviors using raw data from tri-axial accelerometers. The aims is to compare the accuracy of a classifier developed on prescribed activities performed in free living setting with a classifier trained on multiple days of data from free living adults in naturalistic settings. Further, we will investigate the accuracies of a classifier trained on one population group applied to a functionally different population. Since current population studies employ accelerometer intensity cut points of single axis count data, we also compared our behavior classifications to these so that researchers can start to appreciate the differences in the two approaches and consider existing physical activity and sedentary behavior prevalence rates in this light.

METHODS

Overview

Data are from 3 studies in which prescribed activity or wearable cameras allowed us to capture PA behaviors in free living settings assessed by a tri-axial accelerometer recording data 30 times a second. A machine learning algorithm classified participants' daily activities into 6 types of behaviors: 1) sitting, 2) standing, 3) standing & moving, 4) walking/running, 5) sitting in a vehicle, and 6) cycling. Participants provided informed consent and all study procedures were approved by the research ethics board of the university. Detailed protocols, coding manuals, and procedures are available from the corresponding author and/or published (19).

Study design and sample

Study 1 – Prescribed trips—The aim of this study was to collect hip-worn GT3X+ ActiGraph accelerometer data on transportation modes across the city. Two trained research assistants in San Diego collected data under varying conditions (e.g., open space vs. urban, indoor vs. outdoor) for a variety of transportation modes such as walking and driving. A similar distance was travelled for each transportation mode and transitions between transportation modes were balanced (10). Trips were prescribed and the start and end time of each trip was noted. Over 500 trips were recorded.

Study 2 – Cyclist cohort—The aim of this study was to collect accelerometer data on cycling because it is known to be misclassified by accelerometers and travel diaries. Eligible participants were 18–70 year old university employees who routinely cycled for transportation. Participants agreed to wear the SenseCam and the hip-worn GT3X+ ActiGraph accelerometer during waking hours for 3–5 days. Approximately half of the sample wore devices during the weekend with at least one work day included, while the remainder wore the units on weekdays only (19).

Study 3 – Overweight women cohort—The aim of this study was to develop PA measurement algorithms for application in 2 weight loss trials that are part of the NCI-funded Transdisciplinary Research in Energetics and Cancer (TREC) Initiative (23). Briefly, 36 overweight or obese women wore the SenseCam and the hip-worn GT3X+ ActiGraph accelerometer during waking hours for 7 days. Almost half of these women were breast cancer survivors. Details are published regarding participant eligibility criteria and design of these randomized trials (24).

Devices and data processing used for physical activity assessment

Accelerometer data—Participants/researchers wore a GT3X+ ActiGraph accelerometer on a belt over the right hip. While wrist based accelerometers are being worn in some population studies e.g. NHANES, there are numerous large cohort studies with raw data from hip based accelerometers which could benefit from new data processing techniques. Further, it is not yet clear if the wrist location will provide equally accurate assessments as the hip in free living adults (12,29,33). Accelerometer non-wear time was defined as 90 minutes of consecutive zeros (7). Accelerometer data were processed in standard fashion

using ActiLife v6.2.1 software, with 30 Hertz data aggregated to 60 seconds. Accelerometer-based sedentary behavior (SB) was computed as the number of minutes spent below 100 counts per minute (cpm) from the vertical (y) axis data (20). MVPA was defined by the 1952 cut point. For the purposes of machine learning, the accelerometer assessed acceleration on all 3 axes at a rate of 30 Hertz.

Wearable camera (SenseCam) data—Participants in Study 2 and 3 wore the SenseCam on a lanyard around their neck with clothing-safe, adhesive tape attached to reduce movement. Details of the SenseCam system are described elsewhere (8,9). Briefly, it takes photos every 10–15 seconds when an onboard sensor is activated by a change in movement, light, temperature or presence of another person. If no photo is triggered by the sensors, a photo is taken every 20 seconds. Over 3000 wide angle low resolution images can be collected per day. Participants were trained on IRB-approved procedures for ensuring privacy and confidentially for themselves and others. These procedures are described in detail elsewhere (18). Briefly participants review their images and can delete any they do not wish to share, they are instructed to employ a privacy button when needed, to remove the camera in sensitive settings, and to ask permission to wear it when appropriate.

SenseCam image annotation—SenseCam image data were downloaded and imported into the Clarity SenseCam Browser (9). A standardized coding protocol was developed based on existing behavioral taxonomies (e.g., SOFIT (21)) and refined using principles of nominal group technique (26). Inter-rater reliability of image coding was established using an iterative cycle of blind-coding (relative to other coders) followed by discussion, with all disagreements resolved by group consensus. Subsequent coding was done by 3 research assistants who demonstrated >80% agreement with criterion-coded images. Once certified, ~10% of all subsequent images were checked to minimize observer drift.

Coding procedures—A series of at least 5 consecutive images (approximately 2 minutes) in the same behavior were grouped as an ‘Event’ and assigned a corresponding behavior code. First, ‘Sedentary Posture’ was determined as sitting, lying or reclining. Second, ‘Standing’ was defined as standing, moving in place, and moving towards an object. When objects in the image were in the same place from one image to the next, ‘Standing Still’ was coded. If movement was observed, it was coded as ‘Standing & Moving’. If progress towards a distant point was observed, it was coded as ‘Walking/Running’. Street and stationary “Cycling” was coded when handle-bars were present in the image. “Riding in a vehicle” was coded when a steering wheel or dashboard was observed.

Machine learning algorithms—The machine learning process is comprised of three steps: feature extraction, minute-level classification, and time-smoothing. The feature extraction step is the process of transforming raw accelerometer data streams into vectors of a consistent length that capture predictive information. We broke the data stream into 1-minute windows of accelerometer data, each with a corresponding PA behavior label. A window of acceleration measurements contains $T = 60\text{s} \times 30\text{ Hz} = 1800$ time samples of acceleration measurements along the x, y and z axes, which we represent as a matrix,

$$A = \begin{bmatrix} a_{1,x} & a_{2,x} & \dots & a_{T,x} \\ a_{1,y} & a_{2,y} & \dots & a_{T,y} \\ a_{1,z} & a_{2,z} & \dots & a_{T,z} \end{bmatrix}.$$

Most features were computed from the vector magnitude of the 3-axis acceleration.

$$a_t = \sqrt{a_{t,x}^2 + a_{t,y}^2 + a_{t,z}^2}, .$$

We computed 43 features from each 1-minute window of acceleration data, which included basic descriptive statistics as well as entropy, angular features (e.g., roll, pitch and yaw), principle direction of motion, autoregressive coefficients, Fast Fourier Transform coefficients, total power and dominant frequency. We normalized the features to have mean zero and standard deviation one to account for the scale difference between features. The list of features and their importance are described elsewhere (12). Using a 1-minute epoch resulted in 198,622 minutes of data over 520 days which had corresponding ground truth labels.

Minute-level Classification—We tested several standard machine learning algorithms to classify PA behaviors: *k*-nearest neighbor, support vector machines, naive Bayes, decision trees, and random forests. Of these algorithms, the random forest algorithm, which is an ensemble method based on decision trees, produced the highest accuracy. It is notable that Shotton et al. also used random forests to recognize human poses for the Xbox Kinect sensor (28). The training phase of the algorithm consists of building the decision trees, i.e. learning the branches that lead to a tree that correctly classifies as many examples in the training data set as possible. A random forest combines the outputs of multiple randomized decision trees. To learn each decision tree, we chose a stratified random sample of 2000 training examples per behavior class (at 1-min epoch) and a random subset of 25 features. We learned 500 of these randomized decision trees. To classify a given test example, the random forest traverses each tree until it arrives at a leaf node. Each leaf node has a probability score for each behavior, according to the ratio of training examples of each behavior that land in that node. The random forest sums these probability scores in the final leaf node over the trees, and chooses the PA behavior with highest probability for the given test example. We chose the parameters for our classification algorithms (i.e., number of trees to use) using a held-out day of data that were not included in the final cross-validation results.

Time-smoothing—The final stage of the machine learning process uses information about neighboring minutes to improve the minute-level predictions output by the random forest classifier. We used a hidden Markov model (HMM) with one observation per minute to do time-smoothing. The hidden states in the HMM are the true behaviors, and the observed states are the behaviors predicted by the random forest classifier. Transition probabilities between hidden states model the probabilities of transitioning between different behaviors from minute to minute. Observation probabilities between hidden and observed states model the probability of the random forest classifier correctly classifying behaviors. The training

stage of the HMM consists of learning these transition and observation probabilities from the examples in the training dataset. To smooth test data, the HMM learns the most probable sequence of hidden states using the Viterbi algorithm. The ML algorithms presented in this paper are available in an R package (https://cran.r-project.org/web/packages/TLBC/index.html).

Estimating accuracy—Classification accuracy is estimated within each dataset using leave-one-participant-out cross-validation. In this procedure each participant is held out in turn. A classifier is trained on the remaining participants and applied to the held-out participant. Overall accuracy is averaged across all participants. Classification accuracy across datasets is estimated by training a classifier on all the data from a given dataset and applying the trained classifier to each participant in another dataset. For each behavior we report the balanced accuracy, which is the mean of sensitivity and specificity.

Accumulated minutes—Finally, we create day-level variables to represent minutes/day in each behavior. Mixed effects linear regression, adjusted for nesting of days within participants, was used to compare minutes/day as indicated by each classifier.

RESULTS

As shown in Table 1, Study 1 (Prescribed trips) included 2 female research assistants who were under 30 years old. Participants in Study 2 (40 cyclists) had a mean (SD) age of 36 (12) years, 30% were women, and 25% were overweight or obese (BMI ≥ 25 kg/m^2). Participants in Study 3 (36 overweight women) had a mean age of 56 (16) years and all were overweight or obese. Table 1 also shows the proportion of time spent in behaviors when wearing the PA measurement devices. In the free living samples (Studies 2 and 3), participants spent the most time in SB (~50%). In the cyclist cohort, over 6000 minutes of bicycling was observed (6.3%) compared to almost no cycling in the overweight women group (0.1%). Only 12.4% of the total data did not have a corresponding annotation.

Figure 1 compares the current accelerometer thresholds for MVPA and SB to known PA behaviors with box and whisker plots. It is notable that the intensity ranges with the prescribed trips performed by research staff (Study 1) were different in comparison to the free living cohorts (Studies 2 and 3). Panel A shows that when the 1952 cut point was used to define MVPA, approximately 50% of Walking/Running was not at moderate level intensities in the cyclist cohort (Study 2) and almost 80% was not at moderate level intensities in the cohort of overweight women (Study 3). In addition, this cut point did not capture any cycling behavior. Panel B shows that when using the 100 cpm accelerometer cut point, most of Standing Still, and even some Standing & Moving, was incorrectly classified as SB. Panel C shows that all of Sitting, and two thirds of Sitting in a Vehicle, was correctly classified as SB.

Table 2 gives data regarding the accuracy of traditional cut points when compared to annotated PA behaviors. Overall, these cut points classified 43% of walking and 9% of cycling as MVPA; 68% of riding in a vehicle, 84% of sitting, and 66% of standing as sitting. These data also show the accuracy of machine learning algorithms when trained on different

populations. Overall, the best performing algorithm was from the Prescribed Trips (Study 1) when applied to itself (93% accuracy). The Study 1 algorithms performed less accurately in the free living samples (Study 2: 86%, Study 3: 80%). The algorithms developed on the free living cohorts (Studies 2 and 3) performed with 89% accuracy when applied to themselves. However, classifiers trained on one free living population (cyclists in Study 2) showed approximately 6% lower accuracy when applied to a functionally different free living population (overweight women in Study 3).

Table 3 presents comparisons of predicted minutes/day of each behavior in Study 3, according to classifiers trained on Study 2 and Study 3. On average, Study 2 overestimated participants' minutes/day of cycling and standing and underestimated participants' minutes of sitting, riding in a vehicle and walking, when compared to estimates according to Study 3.

DISCUSSION

These findings demonstrate that multiple behaviors can be correctly classified with new machine learning approaches and that these behaviors provide information not captured by intensity cut points alone. For example, our data confirm that 50% of walking, which is the most common and modifiable form of PA in the US (3) does not often occur at intensities identified as moderate in laboratory trials in young people. The new behavior classification may have numerous benefits for research and policy. Research that is focused on specific PA behaviors is needed so that public health recommendations for PA can be specific and interpretable. In particular, a robust measure of walking is needed to design and evaluate studies of walking as a health-related exposure. This type of PA measurement can address public health questions such as whether longer bouts of slow walking have the same health impacts as shorter bouts of fast walking. In addition, it is likely that the intensity or speed of walking needed to improve health will vary depending on the population of interest (e.g., young fit males versus older, obese women). Improving cycling facilities in the US has also been proposed as an impactful policy to change population level activity levels, and our findings show that machine learning techniques applied to hip worn accelerometers predict cycling with accuracies up to 99%. Finally, given that commuting in a vehicle comprises a meaningful proportion of many adults' day, it is important to be able to classify this behavior for epidemiologists studying the health effects of driving or environmental exposures experienced during driving.

While SB research is in its infancy, there is considerable interest in determining the degree to which interventions to reduce SB can improve health (22). However, using an accelerometer cut point to assess SB will lead to considerable misclassification. While the 100 count captured most sitting outside of a vehicle (90%), almost 70% of standing time also occurred under the 100 cpm threshold for SB. This is problematic for detecting breaks in sitting and prolonged sitting that may have stronger relations with outcomes than total sitting (2). Our machine learning algorithms appeared to increase the accuracy of identifying sitting behavior (both in and outside a vehicle) and decreased inaccurate classifications of standing as being sedentary. Additional comparisons of sedentary vs standing using machine learned techniques can be found elsewhere (33). ActivPAL devices also assess sitting and standing with high accuracy.

Our algorithms performed similarly to algorithms developed using other machine learning or statistical techniques (11,13,17,25,30,33). Most previous studies have been conducted in laboratory settings (13), or with some observed data in outdoor locations (30). No previous study has employed automatically captured observations over multiple days in free living or included driving and cycling. Our findings suggest caution when applying accelerometer algorithms developed in restricted settings, such as laboratory trials among young adults. For example, algorithms developed on treadmill walking are unlikely to represent free living walking in diverse study samples. Our analyses show that algorithms developed in more controlled conditions performed with 13% less accuracy on data obtained from free living populations. Bastian et al. (1) found a 20% difference in behavior predictions between laboratory and prescribed activities. In our prescribed data, the research assistants walked at higher intensities than participants in free living.

In addition to the biomechanical differences that are reflected in the raw acceleration patterns, the structure of the training data is important. The performance of the cycling algorithm, for example, was particularly affected by the very low minutes of cycling in study 3. While balancing the amount of data collected on each behavior provides appropriate examples to train the algorithm, information about the prevalence of certain behaviors in free-living can be used to improve prediction performance (e.g. sitting is the most prevalent behavior in free living but not in laboratory trials). Further, training the algorithm on real-length bouts of behaviors can improve the accuracy of predicted bout length (e.g. to predict long bouts of sitting, the training data should contain long bouts of sitting).

There are a number of limitations to these analyses. It is possible that the “ground truth” annotated image data employed to train the machine learning classifiers contained error. Nonetheless, compared to participant diaries or records, the wearable camera is a considerably less burdensome and reactive technique for capturing free-living behaviors over multiple days. In addition, the thorough training and quality control procedures should have minimized coding errors. The sample sizes of the study cohorts were modest, although comparable to most studies of this sort (11,13,17,25,30,33). The total minutes and days of data collected, however, were much greater than previous studies. In addition, we only predicted a small set of behaviors and studies with more behaviors are likely to show lower prediction rates. A strength of this research is that it is the first conducted in free living individuals over multiple days, includes overweight and obese adults, and focuses on transportation behaviors such as walking, cycling, and sitting in a vehicle.

Conclusion

Our findings indicate that machine learning algorithms to classify PA and SB can have high accuracy across markedly different, free-living cohorts over multiple days. We used this protocol to identify 6 behaviors. However this same protocol can be used to develop machine learning algorithms to predict many other behaviors, such as housework, gardening, TV watching, and specific sports.

Classifiers predict behavior most accurately when they are specifically trained, for example in voice recognition. Therefore, ideally, researchers would collect training data on their population of interest and include a calibration phase. For existing hip accelerometer data in

large cohorts, this step may not be possible, so it is important for researchers to understand that classification accuracy will depend on how similar their cohort is to the data on which the classifier was trained. In future, it may be possible to develop a robust classifier that is trained on a large, diverse cohort and therefore can be reliably applied to multiple population groups and behaviors with sufficient levels of accuracy, even if it is not the best performance for any one group. This approach might help with future standardization of processing (34). Researchers who may have been reluctant to use new techniques due to their lack of real world validation may now consider applying these algorithms to their population data with more confidence, if the samples are similar. While new approaches to accelerometer data processing may temporarily hamper calls for standardization of techniques (34), the public health benefits of incorporating a behavioral framework into analyses seem worth it. Granted, our machine learned approach is complex. However, it seems the most appropriate starting point for new data processing techniques; first demonstrate what levels of accuracy can be achieved and then refine algorithms to me simpler and assess what sacrifices to accuracy researchers are willing to make for feasibility. Although complex, the algorithms run smoothly in R (package available https://cran.r-project.org/web/packages/TLBC/index.html) and do not consume much computer power or time. To date, no study has shown whether simpler approaches (29) can deliver similar accuracy levels in totally free living data.

Future studies should also compare an intensity-based approach to PA measurement (i.e., accelerometer cut points) with a behavior-based approach and assess which paradigm correlates most strongly with markers of PA activity and disease risk. Finally, future research should be conducted on wrist worn accelerometer data in free-living populations (not laboratory settings (14)), as these accelerometers are increasingly common. Nonetheless, hip based accelerometers are still employed in multiple large studies. Some researchers have employed 24 hour protocols for hip based accelerometers providing similar compliance benefits to the wrist worn devices (16). If the hip location provides more accurate estimates than the wrist for assessing PA in free living (33), some researchers may choose to stay with the hip location.

These data demonstrate how differences in the training setting, behavioral prevalence, and population can affect algorithm performance. The free living classifier we have developed in middle aged and older adults may have value when applied to existing large cohort studies with hip accelerometer data. Further validation will strengthen the evidence from our current findings.

Acknowledgments

This work was supported by the National Cancer Institute at the National Institutes of Health (Grants U01CA130771, U54CA155435, R01CA164993).

References

1. Bastian T, Maire A, Dugas J, et al. Automatic identification of physical activity types and sedentary behaviors from triaxial accelerometer: laboratory-based calibrations are not enough. J Appl Physiol (1985). 2015; 118(6):716–22. [PubMed: 25593289]

2. Barreira TV, Zderic TW, Schuna JM Jr, Hamilton MT, Tudor-Locke C. Free-living activity counts-derived breaks in sedentary time: Are they real transitions from sitting to standing? Gait Posture. 2015 Jun; 42(1):70–2. [PubMed: 25953504]
3. Berrigan D, Carroll D. Vital signs: walking among adults-United States, 2005 and 2010. Morb Mortal Wkly Rep. 2012; 61(31):595–601.
4. Blair S. The evolution of physical activity recommendations: how much is enough? Am J Clin Nutr. 2004; 79:913–920.
5. Bowles HR, FitzGerald SJ, Morrow JR, et al. Construct validity of self-reported historical physical activity. Am J Epidemiol. 2004; 160(3):279–86. [PubMed: 15258001]
6. Celis-Morales, Ca; Perez-Bravo, F.; Ibañez, L. Objective vs. self-reported physical activity and sedentary time: effects of measurement method on relationships with risk biomarkers. PLoS One. 2012; 7(5):e36345. [PubMed: 22590532]
7. Choi L, Liu Z. Validation of accelerometer wear and nonwear time classification algorithm. Med Sci Sport Exerc. 2011; 43(2):357–364.
8. Doherty AR, Hodges SE, King AC, et al. Wearable Cameras in Health. Am J Prev Med. 2013; 44(3):320–323. [PubMed: 23415132]
9. Doherty AR, Moulin CJa, Smeaton AF. Automatically assisting human memory: a SenseCam browser. Memory. 2011; 19(7):785–95. [PubMed: 20845223]
10. Ellis K, Godbole S, Marshall S, et al. Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms. Front Public Heal. 2014 Apr.2:36.
11. Ellis K, Kerr J, Godbole S, et al. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Phys Meas. 2014; 35(11):2191–203.
12. Ellis K, Kerr J, Godbole S, Staudenmayer J, Lanckriet G. A comparison of wrist and hip accelerometer algorithms for free-living behavior classification. MSSE. 2015 in review.
13. He B, Bai J, Koster A, et al. Predicting human movement type based on multiple accelerometers using movelets. Med Sci Sport Exerc. 2014; 46(9):1859–1866.
14. Hildebrand M, VAN Hees VT, Hansen BH, et al. Age group comparability of raw accelerometer output from wrist- and hip-worn monitors. Med Sci Sport Exerc. 2014; 46(9):1816–24.
15. Hodges S, Williams L, Berry E. SenseCam: A retrospective memory aid. UbiComp 2006. 2006:177–193.
16. Huberty J, Ehlers DK, Kurka J, Ainsworth B, Buman M. Feasibility of three wearable sensors for 24 hour monitoring in middle-aged women. BMC Womens Health. 2015 Jul 30.15:55. [PubMed: 26223521]
17. John D, Liu S, Sasaki JE, et al. Calibrating a novel multi-sensor physical activity measurement system. Phys Meas. 2011; 32(9):1473–1489.
18. Kelly P, Marshall SJ, Badland H, Kerr J, Oliver M, Doherty AR, Foster C. An Ethical Framework for Automated, Wearable Cameras in Health Behavior Research. American journal of preventive medicine. 2013; 44(3):314–319. [PubMed: 23415131]
19. Kerr J, Marshall S, Godbole S, et al. Using the SenseCam to improve classifications of sedentary behavior in free-living settings. Am J Prev Med. 2013; 44(3):290–296. [PubMed: 23415127]
20. Matthews CE, Chen KY, Freedson PS, et al. Amount of time spent in sedentary behaviors in the United States, 2003–2004. Am J Epidemiol. 2008; 167(7):875–81. [PubMed: 18303006]
21. McKenzie T. Observational measures of children's physical activity. J Sch Heal. 1991; 61(5):224–227.
22. Owen N, Healy G. Too much sitting: the population-health science of sedentary behavior. Exerc Sport. 2010; 38(3):105–113.
23. Patterson RE, Colditz Ga, Hu FB, et al. The 2011–2016 Transdisciplinary Research on Energetics and Cancer (TREC) initiative: rationale and design. Cancer Causes Control. 2013; 24(4):695–704. [PubMed: 23378138]
24. Patterson RE, Rock CL, Kerr J, et al. Metabolism and breast cancer risk: frontiers in research and practice. J Acad Nutr Diet. 2013; 113(2):288–96. [PubMed: 23127511]

25. Pober DM, Staudenmayer J, Raphael C, et al. Development of novel techniques to classify physical activity mode using accelerometers. Med Sci Sport Exerc. 2006; 38(9):1626–34.

26. Potter M, Gordon S, Hamer P. The Nominal Group Technique: A useful consensus methodology in physiotherapy research. New Zeal J Physiother. 2004; 32(3):126–130.

27. Sallis JF, Cervero RB, Ascher W, et al. An ecological approach to creating active living communities. Annu Rev Public Heal. 2006; 27:297–322.

28. Shotton J, Sharp T, Kipman A. Real-time human pose recognition in parts from single depth images. Commun ACM. 2013; 56(1):116–124.

29. Staudenmayer J, He S, Hickey A, Sasaki J, Freedson P. Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. J Appl Physiol (1985). 2015 Aug 15; 119(4):396–403. [PubMed: 26112238]

30. Staudenmayer J, Pober D, Crouter S, et al. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. J Appl Physiol. 2009; 107(4):1300–7. [PubMed: 19644028]

31. Troiano R, Berrigan D. Physical activity in the United States measured by accelerometer. Med Sci Sport Exerc. 2008; 40(1):181–188.

32. Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. Br J Sports Med. 2014; 48(13):1019–23. [PubMed: 24782483]

33. Trost SG, Zheng Y, Wong W-K. Machine learning for activity recognition: hip versus wrist data. Phys Meas. 2014; 35(11):2183–2189.

34. Wijndaele K, Westgate K, Stephens SK, Blair SN, Bull FC, Chastin SF, Dunstan DW, Ekelund U, Esliger DW, Freedson PS, Granat MH, Matthews CE, Owen N, Rowlands AV, Sherar LB, Tremblay MS, Troiano RP, Brage S, Healy GN. Utilization and Harmonization of Adult Accelerometry Data: Review and Expert Consensus. Med Sci Sports Exerc. 2015 Oct; 47(10): 2129–39. [PubMed: 25785929]

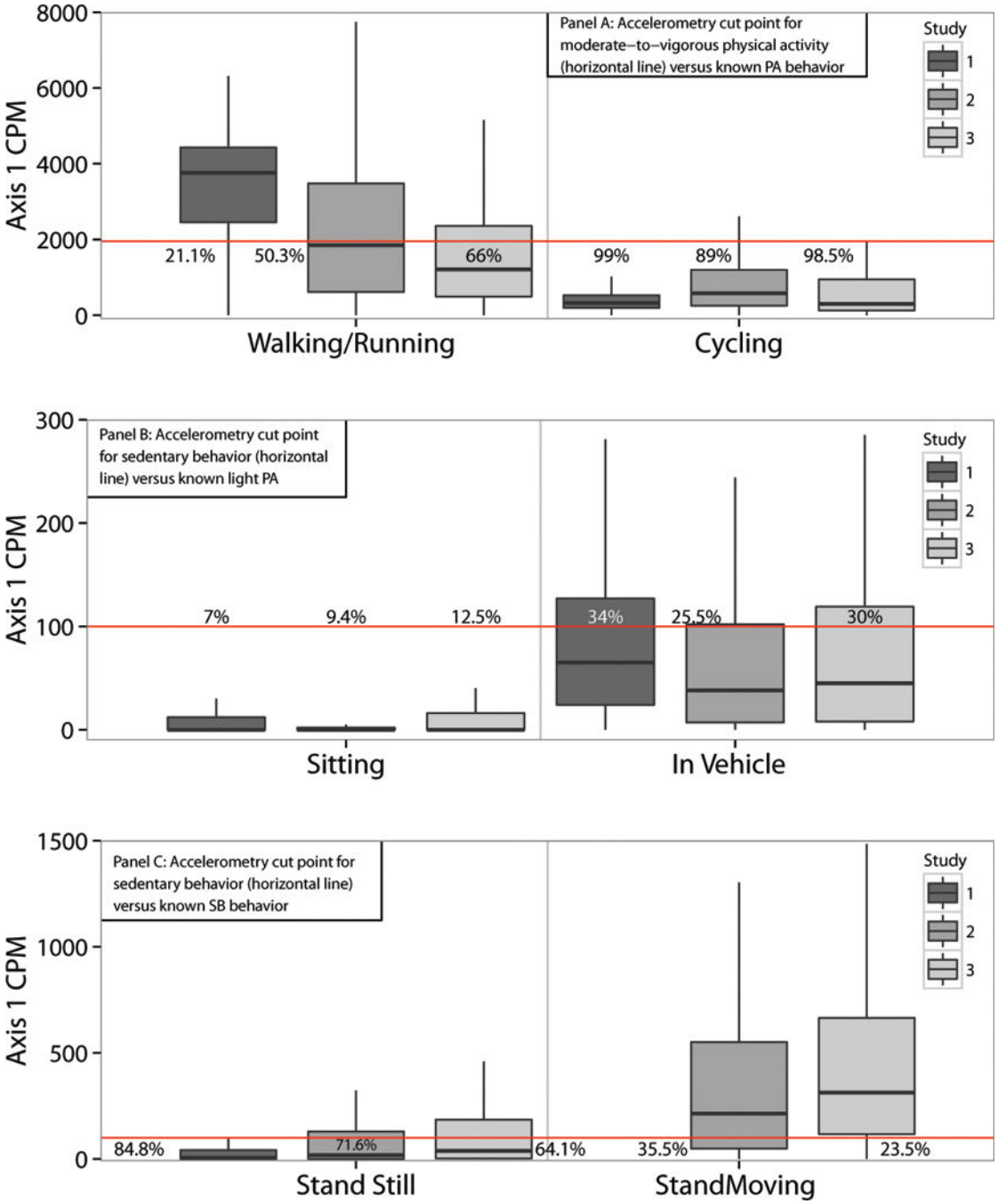

Figure 1.

Comparison of Accelerometer Classification of Physical Activity and Sedentary Behavior to Actual Behaviors using 3 Study Designs and Samples with box and whisker plots (Study 1: Prescribed transportation modes by 2 research assistants; Study 2: Usual daily activities among 40 cyclists; Study 3: Usual daily activities among 38 overweight females).

*Percentages reflect the amount of behavior falling within established accelerometer intensity thresholds. Within the box is the 25–75th percentile and the solid line is the median. The lines outside the box (whiskers) represent variability outside the quartiles

Table 1

Studies in Which Prescribed Activity or Wearable Cameras Were Used to Capture Physical Activity Behaviors in Free Living Adults.

	Study 1	Study 2	Study 3
Behavior Types	**Prescribed transportation**	**Usual daily activities**	**Usual daily activities**
Sample Description (N)	**Research assistants (2)**	**Cyclists (40)**	**Overweight females (36)**
% overweight or obese	NA	25	100
Age, years (mean, SD)	NA[a]	36 (12)	55 (16)
Female (%)	100%	30%	100%
Overweight or obese (%)[b]	0%	25%	100%
Days of monitoring physical activity	35	171	314
Proportion of time spent in activities[c]			
Bicycling	9.7%	6.8%	0.1%
Riding in vehicle	32.2%	5.8%	11.5%
Sitting (not in a vehicle)	8.9%	51.3%	54.8%
Standing still	17.5%	8.7%	6.6%
Standing & moving	NA	10.5%	7.2%
Walking/running	28.2%	6.7%	6.1%
Uncodeable	3.5%	10.2%	13.8%
Total	100%	100%	100%

[a]Aged under 30 years old

[b]Body mass index ≥ 25 kg/m^2

[c]As determined by use of prescribed activities (Study 1) or a wearable camera in which images were manually coded for the type of activity (Study 2 and 3)

Table 2

Accuracy of Machine Learning Classifiers of Physical Activity Behaviors Compared to Traditional Accelerometer Cut points When Applied to 3 Studies

Study 1: Prescribed transportation modes by 2 research assistants; Study 2: Usual daily activities among 40 cyclists; Study 3: Usual daily activities among 36 overweight females.

	Walking/Running	Cycling	Riding in vehicle	Sitting	Standing	Standing Moving	Mean[c]
Comparison to known physical activity behaviors							
Accelerometer cut points[a]	0.43	0.09	0.68	0.84	0.34	0.27	–
Accuracy of machine learning classifiers							
Study 1 when trained on:							
Study 1 (itself)	0.98	0.99	0.96	0.87	0.91	–	0.93
Study 2	0.97	0.99	0.88	0.87	0.89	–	0.90
Study 3	0.92	0.50[b]	0.94	0.75	0.82	–	0.86
Study 2 when trained on:							
Study 1	0.87	0.84	0.91	0.86	0.81	–	0.86
Study 2 (itself)	0.82	0.97	0.93	0.92	0.70	0.83	0.84
Study 3	0.85	0.54[b]	0.93	0.92	0.70	0.74	0.83
Study 3 when trained on:							
Study 1	0.84	0.78[b]	0.87	0.80	0.68	–	0.80
Study 2	0.73	0.82[b]	0.85	0.89	0.64	0.81	0.78
Study 3 (itself)	0.86	–	0.94	0.92	0.70	0.82	0.85

[a]Moderate-to-vigorous physical activity cut point (≥1952 counts per minute) applied to Walking/Running and Cycling and sedentary behavior cut point (≤100 counts per minute) applied to Riding in a Vehicle, Sitting, and Standing, for all three studies combined.

[b]Only one woman provided 135 minutes of cycling in Study 3

[c]Mean does not include cycling minutes that were overly influenced by the single cyclist in Study 3.

Table 3

Mean difference and agreement for minutes per day of behaviors in Study 3

Usual daily activities among 36 overweight females. N = 295 days

	Estimated Mean (SE) Minutes/day[a]			
	Trained on Study 2	**Trained on Study 3**	**Difference**	**p**
Walking/Running	31.7 (5.3)	57.7 (5.3)	−26.0 (2.6)	<0.001
Cycling	17.1 (1.7)	0.1 (1.7)	17.0 (1.3)	<0.001
Riding in Vehicle	54.1 (5.4)	73.3 (5.4)	−19.2 (3.6)	<0.001
Sitting	353.5 (15.6)	363.2 (15.6)	−9.6 (6.6)	0.142
Standing	73.0 (6.7)	62.3 (6.7)	10.7 (4.1)	0.009
Standing Moving	144.3 (10.8)	117.1 (10.8)	27.1 (5.1)	<0.001

[a]From mixed-effects linear regression models adjusted for wear time and nesting of days within participants

Mean wear time = 674 minutes/day