Published in final edited form as:

Am J Health Behav. 2019 May 01; 43(3): 543–555. doi:10.5993/AJHB.43.3.9.

Physical Activity Change in an RCT: Comparison of Measurement Methods

Sandahl H. Nelson, PhD, MS, Loki Natarajan, PhD [Professor]

Department of Family Medicine and Public Health, University of California-San Diego, La Jolla, CA.

Ruth E. Patterson, PhD [Professor Emeritus]

University of California-San Diego, La Jolla, CA.

Sheri J. Hartman, PhD [Assistant Professor]

Department of Family Medicine and Public Health, University of California-San Diego, La Jolla, CA.

Caroline A. Thompson, PhD, MPH [Assistant Professor]

School of Public Health, San Diego State University, San Diego, CA.

Suneeta V. Godbole, MPH [Data Manager], Eileen Johnson, BA [Research Assistant]

Department of Family Medicine and Public Health, University of California-San Diego, La Jolla, CA.

Catherine R. Marinac, PhD [Research Fellow]

Division of Population Sciences, Department of Medical Oncology, Dana-Farber Cancer Institute and the Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA.

Jacqueline Kerr, PhD [Professor]

Department of Family Medicine and Public Health, University of California-San Diego, La Jolla, CA.

Abstract

Objectives: We aimed to quantify the agreement between self-report, standard cut-point accelerometer, and machine learning accelerometer estimates of physical activity (PA), and examine how agreement changes over time among older adults in an intervention setting.

Methods: Data were from a randomized weight loss trial that encouraged increased PA among 333 postmenopausal breast cancer survivors. PA was estimated using accelerometry and a validated questionnaire at baseline and 6-months. Accelerometer data were processed using standard cut-points and a validated machine learning algorithm. Agreement of PA at each time-

Correspondence Dr Nelson; sandahln@gmail.com.

Human Subjects Statement

The Institutional Review Board of the University of California, San Diego granted ethics approval for the trial (IRB# 101977). Participants provided informed consent.

Conflicts of Interest Statement

There are no conflicts of interest to declare.

point and change was assessed using mixed effects regression models and concordance correlation.

Results: At baseline, self-report and machine learning provided similar PA estimates (mean difference = 11.5 min/day) unlike self-report and standard cut-points (mean difference = 36.3 min/day). Cut-point and machine learning methods assessed PA change over time more similarly than other comparisons. Specifically, the mean difference of PA change for the cut-point versus machine learning methods was 5.1 min/day for intervention group and 2.9 in controls, whereas it was 24.7 min/day for other comparisons.

Conclusions: Intervention researchers are facing the issue of self-report measures introducing bias and accelerometer cut-points being insensitive. Machine learning approaches may bridge this gap.

Keywords

breast cancer; accelerometry; weight loss; intervention; machine learning

Evidence regarding the public health importance of physical activity (PA) is incredibly robust, from decreasing incidence of heart disease¹ to warding off and helping to recover from some forms of cancer.^{2,3} Studies are continuing to be designed to help hone PA guidelines, and trials are run to understand ways to increase PA, and to quantify the impact of increasing PA on various health markers and outcomes. It is important to public health that we continue to expand our understanding of PA measures while also developing strategies to increase PA. Thus, it is equally important to understand how the way in which we measure PA may affect our study findings and ultimately shape PA recommendations to the public.

The majority of PA research to date relies on self-report or accelerometry measures. Agreement between self-report and accelerometry measured PA is generally low to moderate, but with a wide range from -0.71 to 0.96.⁴ Thus, choice of measurement method could have an impact on the reported levels of activity, as well as assessments of change in activity over time. This discrepancy is particularly important in randomized trials that need to determine participants' eligibility to enter a trial based on estimates of activity or seek to compare change in PA between conditions.

The lack of agreement between self-report and accelerometry methods does not predicate that one measurement should be used while the other is discarded, as both self-report and accelerometer measures are affected by different sources of systematic error. Self-report PA may be misclassified due to poor recall, social desirability (ie, over-reporting of activity), misunderstanding of terms, or use of survey items that fail to capture the respondents' primary mode of PA.⁵ Recall bias may be of particular concern in older populations or cancer survivors whose treatment may have affected cognitive capacity.^{6,7} Social desirability bias can be problematic in intervention trials, as one group may come into greater contact with study staff, leading them to potentially report greater activity improvements. At the same time, accelerometer measures are subject to measurement error due to differences in non-wear time (eg, if participants only wear devices on active days), number of

accelerometer wear days deemed valid, missing activities such as swimming, or due to different decision rules used when processing the data. Specifically, standard intensity cut-point methods developed in young, healthy participants may not perform well in all populations, such as older adults or cancer patients with treatment related side effects. In these individuals, the threshold that constitutes vigorous activity changes and absolute cut-points may not take into account this physical limitation.^{8,9} For example, in older adults, the correlation between accelerometer counts and caloric expenditure is not linear, especially when assessed on a treadmill.⁹ This difference between relative (activity intensity based on an individual's fitness level) and absolute intensity is one reason self-report and accelerometer measurements in older adults has been shown to have especially low correlations.¹⁰ For example, given that many older adults prefer to walk, recall of this activity may be more accurate while accelerometer cut-points defined for a younger more fit population may underestimate the same walking.

A potential mechanism for increasing the agreement between self-report and accelerometry is the use of machine learned algorithms to process accelerometry data. Machine learned algorithms have been developed for multiple population groups to potentially overcome the measurement error associated with employing absolute cut-points. Machine learning methods can employ data from the 3 axes of movement and extract feature vectors from the raw data signal.¹¹ Most algorithms, however, have been developed in the laboratory setting, which limits their application to real world cohorts. In contrast, Ellis et al¹¹ have developed algorithms using real world data, including in breast cancer survivors. This specific machine learned algorithm was developed to classify activity based on behavior rather than intensity. Thus, we seek to assess whether, when developed in a free-living setting and in a population with similar fitness levels, machine learning behavior estimates may help to capture PA estimates that more closely align with relative intensity, while removing biases associated with self-report. Whereas self-report and accelerometer measures have provided disparate estimates, we hypothesize that they could be more closely aligned and that machine learning behavioral algorithms offer a way in which to process accelerometry data, focused on behavior rather than intensity, that could bridge the two and provide estimates that align with both self-report and standard accelerometer measures.

This study leveraged data from *Reach for Health*, a weight loss trial that encouraged increased PA and reduced energy intake among 333 postmenopausal, overweight, or obese breast cancer survivors.¹² The trial collected self-report and accelerometer measured PA at baseline and follow-up (6 months). Accelerometer measures were processed using standard cut-points and using a validated machine learning algorithm. The aim of this study was to quantify agreement between self-report, standard cut-point accelerometer, and machine learning accelerometer activity estimates and to examine how agreement may differ by randomization group over time. All with the goal of providing measurement method recommendations for future PA research to facilitate standardization of methods.

METHODS

For this study, we use data from Reach for Health (RFH), a randomized, controlled, weight loss trial in non-diabetic breast cancer survivors, conducted at the University of California

(UC) San Diego as part of the Transdisciplinary Research in Energetics and Cancer (TREC) Center initiative to examine the role of insulin resistance and inflammation in breast cancer risk. Participants were recruited between August 2011 and May 2015 from San Diego and the surrounding communities.

Details of the RFH Study are published.¹² In short, the RFH study was a 4-arm trial in 333 postmenopausal, overweight/obese (BMI ≥ 25 kg/m²), women diagnosed with stage 1, 2, or 3 breast cancer within the past 10 years, with no eligibility restrictions placed on habitual PA levels. The trial used a 2×2 factorial design with all participants randomly assigned to weight loss counseling versus educational materials and to either Metformin or a placebo, resulting in a 4-arm trial, with all therapies conducted over 6 months. Metformin is a common oral diabetes medication that helps control blood sugar levels and is safe for non-diabetics. Because of the factorial design, Metformin was equally distributed between the counseling and educational material arms, and had no impact on PA measures; thus, analysis only compared the weight loss versus control groups. Participants wore hip accelerometers (ActiGraph GTX3+) and completed the Global Physical Activity Questionnaire (GPAQ) at baseline and at 6-month follow-up.

Lifestyle Intervention versus Control

The lifestyle intervention was delivered by trained lifestyle coaches via 12 motivational interview calls over the 6-month intervention. The 12 calls were scheduled with weekly frequency in the beginning of the intervention followed by monthly calls later in the intervention. Food-related weight loss strategies were discussed, and participants were given pedometers and encouraged to increase their PA levels to 300 min/week of moderate to vigorous physical activity (MVPA), primarily through walking. Women in the control group were given the 2010 US Dietary Guidelines for Americans and were contacted by study staff at 2 weeks, one month, and 3 months to encourage adherence to taking their pills (double-blinded Metformin or placebo).

Data Collection

Self-reported physical activity.—Participants answered the Global Physical Activity Questionnaire (GPAQ) twice during the study period. The first questionnaire was completed at baseline (prior to randomization) and the second at 6-month follow-up. GPAQ is a validated questionnaire developed by the World Health Organization to estimate PA in a typical week.^{13,14} The GPAQ collects information on PA accumulated in 3 domains; at work, in travel to and from places, and in recreational activities. The GPAQ estimates activity in a typical week by asking; time spent per day in both moderate and vigorous intensity activities in each of the 3 domains and asking, in a typical week, how many days one does moderate and vigorous intensity activities as part of each of the 3 domains.¹⁵ The answers to these questions are then processed to calculate the time of total PA on average per day as per the World Health Organization scoring guidelines.¹⁶

Accelerometer measures.—All participants wore the ActiGraph (model GT3X+) at baseline (prior to randomization) and at the 6-month follow-up, in most cases within one week of when the GPAQ data were collected. Participants were instructed to wear the

ActiGraph on the right hip during all waking hours, excluding water-based activities (eg, swimming, showering) for 7 consecutive days, approximately one week prior to their baseline and 6-month follow-up visits. The ActiGraphs were initialized to collect data at 30 Hz. To apply the traditional cut-points, ActiGraph data was processed using the ActiLife software to create 60-s epochs with the low-frequency extension enabled. The data were then processed using the Choi algorithm in R¹⁷ to define wear and non-wear minutes. Minute level PA estimates from the accelerometer were processed using the standard 1952 cut-point.¹⁸ Time in MVPA was used to estimate PA time. As a sensitivity analysis, we also examined accelerometer data processed using 1041 and 760 cut-points as these have been proposed to be more appropriate for older adults.⁸

Machine learning processed physical activity.—The machine learning algorithm used in this analysis was specifically developed for this population of overweight/obese postmenopausal breast cancer survivors and its creation has been described in detail elsewhere.¹¹ In brief, for development of the machine learning algorithm 38 women wore an ActiGraph GT3X+ accelerometer on the right hip and a camera (a SenseCam) around their neck which captured images at 15 second intervals during waking hours over 7 free-living days. The activities captured by the pictures were manually annotated and served as the ground truth by which to develop the algorithm to process accelerometer measured activity. The machine learning algorithm used 39 statistical features with a random forest classifier and a hidden Markov model to predict 4 behaviors (sitting, standing, ambulation, and riding in a vehicle). The development and testing process was conducted in 38 women not eligible for (or included in) the current trial but who represented the population. The algorithm was further validated in another sample of 222 older women, independent from the testing phase.¹⁹ In this analysis the machine learning estimate of ambulation time serve as the machine learning estimate of PA time.

Participant characteristics.—Demographics data (age, ethnicity, and educational attainment) as well as Short Form Health Survey (SF-36) measures of mental and physical health were obtained via self-report at baseline. Weight and height were measured by trained study staff at baseline and were used to calculate BMI. Cancer related variables were obtained via medical chart abstraction. Details of measurement procedures are published.¹²

Data Analysis

Data preparation.—Of the initial 333 women enrolled in the RFH trial, 289 (87%) had self-report and accelerometer measured PA at baseline and follow-up. For the purposes of this analysis only the 289 with complete PA data were included in the analytic sample.

For all PA analyses, the mean of the day-level accelerometer and mean of the day-level machine learning measurements were calculated (separately at baseline and follow-up) to provide single estimates of minutes per day. Collapsing daily measurements to a single estimate was done to match the units in self-report, which asks about average weekly activity and is scored to provide a single estimate of “minutes of total physical activity on average per day” as outlined in the GPAQ analysis guidelines.¹⁶

Descriptive statistics.—Descriptive statistics (mean/SD) were calculated for demographic variables at baseline, stratified by group (lifestyle intervention and control) and overall. We also calculated descriptive statistics (mean/SD) of the minute per day estimates of PA from each measurement technique: self-report (SR), accelerometer using machine learning processing (ML), and accelerometer using each of 3 cut-points: Freedson's 1952 cut-points (1952 CP), Copeland 1041 cut-points (1041 CP), and Matthews 760 cut-points (760 CP).^{8,18,20} Descriptive statistics were calculated at baseline (pre-randomization), follow-up, and for the change between the time-points (follow-up – baseline), with follow-up and change statistics stratified by intervention group assignment. We also assessed estimates of the intervention effect on the change in activity, separately, using each PA measurement technique. This analysis was carried out using a mixed effects regression models of activity with fixed effect terms for intervention group (lifestyle intervention or control), time-point (baseline or follow-up), and the time-by-intervention interaction. These models included a random intercept and used unstructured covariance as determined by AIC comparison. For all models, other than self-report, we also adjusted for the sum of accelerometer wear time at each time-point. The sum of wear time was used rather than daily wear time because daily accelerometer measurements were collapsed to a single estimate (as described previously in “data preparation”). Thus, the sum of wear time, rather than average, allowed us to control for the total amount of information gathered, ie, how many days the accelerometer was worn and for how long each day. Per standard statistical procedures in a randomized controlled trial, no additional covariates were controlled for in the regression models.

Comparison of agreement.—Agreement of the estimates of min/day of PA at baseline, at follow-up, and the change (follow-up – baseline) were examined pairwise between self-report, accelerometer using machine learned processing, and accelerometer using only 1952 cut-points. Statistical analysis of the agreement between measures was carried out by calculating descriptive statistics (mean, SD) and the concordance correlation for each pairs' difference at baseline, follow-up, and for the activity change (follow-up – baseline).²¹ We also calculated the Pearson and Spearman correlations for each of these comparisons. All analyses were stratified by lifestyle intervention with the exception of baseline data which were assessed with the groups combined, as this measurement was prior to randomization (Table 2).

To test for statistical significance of the difference between measurement techniques, in a way that accounts for the correlated nature of the measurements nested within individuals, we used a linear mixed effects model with a random intercept. This modeled PA as the outcome and included a fixed effect terms for measurement method, time (baseline versus follow-up), group (control versus intervention), as well as all 2- and 3-way interactions between these variables. Appropriate model contrasts were used to test if there were measurement method differences in estimates of: (1) activity at baseline and follow-up; (2) change in activity in each arm (2-way interactions); and (3) group differences in change in activity (3-way interactions), the results for which are presented in Table 3. Because the mixed effects regression model accounts for the repetition of measurement in the same

person (thereby holding demographic variables constant), no additional covariates were added to the models.

RESULTS

Participants

At baseline, women in the study were on average 62.8 years old (SD = 7.03), predominantly white (83.4%), non-Hispanic (88.2%), with a college education or greater (50.9%). Participants were an average of 2.6 years post-surgery (SD = 1.93); 48.8% had Stage 1 breast cancer, 51.9% had received chemotherapy, and 70.2% had received hormone therapy. The average BMI was 31.1 kg/cm³ (SD = 4.97), participants had an average self-reported physical health score of 46.2 (SD = 8.81) and mental health score of 51.3 (SD = 9.18). Table 1 presents these demographics, stratified by intervention group. There were no statistically significant baseline differences between study arms (all p s > .05).

Activity Distribution and Intervention Main Effects by Measurement Technique

Table 2 shows the average minutes per day of PA at baseline, follow-up, and the change (follow-up – baseline) for each of the measurement techniques (self-report, machine learning, Freedson 1952 accelerometer cut-point, Copeland 1041 accelerometer cut-point, and Matthews 760 accelerometer cut-point). As the baseline measurement period was pre-randomization and there were no significant differences between control and intervention group ($p > 0.1$) the groups remained combined at baseline. Self-report PA at baseline was 56.2 (SD = 76.48) min/day; the controls showed an increase in 22.7 min/day versus an increase of 36.9 min/day in the intervention group. Machine learning estimates at baseline were similar to self-report, with an average min/day of 67.8 (SD = 41.08). Unlike with self-report, machine learning estimates for controls show a decrease of 4.9 (SD = 29.50) min/day, while the intervention group shows an increase of 7.5 (SD=35.88) min/day. Freedson, 1952, cut-point estimates provide the lowest baseline measures with an average of 20.0 (SD=18.51) min/day. The controls show a decrease of 2.0 (SD = 13.34) min/day, while the intervention group shows an increase of 2.4 (SD = 17.16) min/day. The Copeland, 1041, cut-point baseline estimate was 62.0 (SD = 35.46) min/day, controls show a decrease of 4.3 (SD = 22.75) min/day, while the intervention group shows an increase of 1.6 (SD = 28.95) min/day. Matthews, 760, cut-point estimates provide the highest baseline measures with an average of 95.3 (SD = 45.81) min/day. The controls show a decrease of 5.5 (SD = 28.92) min/day, while the intervention group shows an increase of 0.2 (SD = 36.11) min/day. Assessment of the statistical significance of the intervention effect on PA change (ie, the difference in activity change between groups) was non-significant when measured by self-report ($p = .180$). With both the machine learning and accelerometry using the standard 1952 cut-point we found a significant intervention effect for the change (difference of change) in PA ($p = .003$ and $.028$ respectively). Similar to self-report, the Copeland (1041) and Matthews (760) accelerometer cut-point did not detect a significant intervention effect ($p = .095$ and $.235$ respectively).

Due to non-significance of the intervention effect over time with the lower accelerometer cut-points, we carried out the remaining comparisons using only the 1952 cut-point.

Physical activity over time, stratified by group, for each of the 3 measurement techniques - self-report, accelerometer cut-point (1952), and machine learning - are depicted graphically in Figure 1. This figure shows that baseline estimates are similar when measured using self-report or machine learning but are lower using 1952 cut-points, regardless of intervention group. It also illustrates that the trajectory of change is similar when measured by machine learning or 1952 cut-points, but that the trajectory of change is larger (in both the control and intervention group) when measured by self-report.

Agreement among Self-report, Machine Learning, and Standard Accelerometry

Table 3 presents a formal pairwise comparison of self-report, machine learning, and accelerometry using the 1952 cut-point (chosen given its sensitivity to detect change) stratified by intervention group for post-randomization assessment. Baseline values show statistically significant concordance correlation (r_c) between each pairwise comparison, as well as significant Pearson (r_p) and Spearman (r_s) correlation (Supplemental Table 1). Despite statistical significance, baseline correlation was low ($r < .5$) for all comparisons, with the exception of Pearson and Spearman correlation between machine learning and 1952 cut-points ($r_p = .62$ and $r_s = .58$). At baseline, our mixed effects regression analysis showed a significant difference between the 1952 cut-point estimate and both the self-report and machine-learning activity estimates, with difference estimate of 36 and 48 minutes per day respectively ($p < .001$). Baseline estimates of self-report and machine learning were also significantly different ($p = .002$), but with a smaller magnitude of difference (11.5 min/day).

Comparison of follow-up PA estimates, stratified by group, found a significant concordance correlation between each pairwise comparison for the intervention group, although the correlation values were again low ($r_c = .31$). Within the control group only the comparison of machine learning and 1952 cut-points showed significant concordance correlation. All comparisons still showed significant Spearman correlation, although Spearman correlation estimates were lower in the control group. This is notable as Spearman correlation indicates similar ranking of participants, an attribute that is important to maintain if association between PA and another trait (eg, biomarkers) is of interest. The comparison of follow-up PA estimates from our mixed effects regression analysis showed statistically significant differences in estimates for all comparisons, with the exception of the comparison between machine learning and self-report within the control group only ($p = .061$).

Concordance correlation of the change estimates between self-report and both machine learning and standard accelerometry were low (essentially zero, $r_c = .07$) and non-significant, regardless of group. This was also the case for Pearson correlation ($r_p = .11$). The correlation of ranking, as measured by Spearman correlation, showed statistically significant correlation of the change estimate between all measurement techniques in the intervention group; however, in the control group, only machine learning and the 1952 cut-points showed significant agreement of ranking (Spearman correlation). This indicates that there is a differential level of agreement between measurement techniques in the control and intervention group for the ranking of each participants' PA change.

DISCUSSION

Baseline Physical Activity Comparisons

In this study we undertook an in-depth comparison of self-report, cut-point accelerometry, and machine-learning accelerometry methods for measuring PA and PA change within the context of a randomized lifestyle intervention trial.

In our comparison of baseline, pre-randomization measurements we found that standard 1952 cut-point accelerometry had much lower estimates of PA than self-report, which aligns with previously published findings comparing the International PAQ (IPAQ) with 1952 cut-point accelerometry.^{22,23} We also replicated results which found that machine learning activity estimates were much higher than the 1952 cut-point estimates.²⁴ A novel finding was that machine learning activity estimates closely matched self-report and 1041 accelerometry cut points, whereas 760 cut-point provided much higher estimates. These findings in lower cut-points were contrary to a 2015 analyses within the Women's Health Study,²⁵ which found both the 1041 and 760 cut-points provided lower estimates for the percent of women meeting 150 min/wk of MVPA than their self-report survey. This discrepancy may be explained by the use of a Women's Health Study specific questionnaire for self-report PA, as opposed to the GPAQ used in our study, and raises the importance of standardizing self-report questionnaires. A second study which compared PA using self-report (Active Australia Survey) and accelerometry (1952 cut-point) in a randomized controlled trial of PA among male and female type 2 diabetics²⁶ also found that, at baseline, self-report measures were lower than accelerometry measures, again a difference possibly explained by the difference in the self-report measure used; it could also be due to the use of PA (based on self-report) as a known eligibility criteria for entry into the study, which was not a criterion for entry into the study we assessed, thus causing participants to underreport PA in a conscious or unconscious effort to gain entry into the study.

The findings from our baseline comparison has implications for cross-sectional studies and for eligibility screening in future randomized trials. In particular, if PA guidelines are developed mostly from self-reported evidence²⁷ then the 1952 cut-point would greatly underestimate the number of people meeting guidelines, especially older adults. Thus, for eligibility screening in intervention trials or cross-sectional analysis interested in how well people are meeting guidelines, the GPAQ self-report, machine learning, or 1041 accelerometry cut-point appear to provide similar estimates and would more closely align with current PA guidelines.

Physical Activity Change Comparisons

In our examination of change in PA, we found notable differences between measurement techniques. Self-report estimates showed large increases in PA among both the intervention and control group, while accelerometry methods (machine learning and cut-points) detected only modest increases in the intervention group and showed decreases among the control group. Our findings are consistent with the examination of change in a previous study assessing agreement in PA change between self-report (IPAQ) and accelerometry (1952 cut point) in a single arm trial of PA in Latina women²³ which found that participants self-

reported far greater increases in moderate PA than were detected by the accelerometer (median 232 min/wk increase vs a 65 min/wk increase respectively). Another study, previously referenced, that compared PA change using self-report (Active Australia Survey) and accelerometry (1952 cut-point) in a randomized controlled trial of PA²⁶ also found that the measurement difference (between self-report and 1952 cut points) increased following the intervention, but unlike in our study, they found that the magnitude of the difference was greater in the intervention group than in the control group. We also observed a weakening of the correlation between self-report and accelerometry (machine learning and 1952 cut point) at follow-up as compared to the baseline. A finding that, in the self-report and 1952 cut point comparison, has been previously seen.^{28,29} A novel finding in our study was the greater decline in correlation between self-report and accelerometry (machine learning and 1952 cut points) seen in the control group as compared to the intervention group, a finding which was not seen between the machine learning and 1952 cut-point measures. Change estimates measured by machine learning and 1952 cut-points were significantly correlated (regardless of intervention group) but were both uncorrelated and meaningfully different (on the order of about 30 min/day) from self-report. The difference in correlation following the intervention, and concomitant difference in change estimates, between self-report and accelerometry that was seen in these studies, and in our own, may stem from an increase in social desirability bias after completion of a study that included contact with study staff. Alternatively, it may stem from a genuine over-estimation of the activity in an 'average day' especially after randomizing participants to an intervention which increased focus on daily PA. Regardless of the underlying reason, these findings support the use of accelerometry, rather than self-report, in studies where change in PA is an outcome of interest, especially when change is compared between groups.

This study also assessed differences, based on measurement technique, in study findings regarding PA in the context of a randomized intervention trial. In randomized trials, it is the difference in PA change between groups that is often the most important statistical parameter. Here we find that only the machine learning and 1952 cut-points provide sensitive enough measurements to detect a statistically significant difference in PA change. This finding indicates that intervention trials which use self-report or lower thresholds may fail to detect significant change and supports the use of machine learning or 1952 cut-points in such trials.

Limitations

It is important to note that this study was conducted in a specific population – a convenience sample of overweight, post-menopausal, breast cancer survivors, and thus, may not be generalizable to another population. Whereas conducting this study in a specific population is a potential limitation, it is also a strength, as it is generally less active populations (eg, older and overweight) for which we target our PA studies and interventions. In addition, understanding measurement differences for PA among older adults is of particular importance in breast cancer patients and survivors, as breast cancer incidence is more than 7 times higher in women over age 50³⁰ and there are well documented benefits of PA in breast cancer patients and survivors^{31–34} make PA research in this population an important and expanding field of study.

Another potential limitation is the 13% of the original 333 RFH participants that were missing a PA measurement at follow-up. We choose not to impute their activity measurement as we did not feel comfortable assuming data missing as random (ie, we felt failure to provide follow-up PA measurements may be associated with their PA level). We attempted to mitigate this limitation by removing them from baseline comparisons as well, thereby ensuring that differences in measurements at baseline and follow-up were not due to differences in the makeup of the sample. However, it is possible that those who failed to provide a follow-up activity measurement would have had a different level of agreement between the measures.

The examination of different cut-points is also limited by only including women, as PA in men may be detected differently by cut-point accelerometry. Another important consideration is the fact that the machine learning algorithm used in this study was designed specifically for this population; thus, use of a less specific algorithm may not perform as well and designing an algorithm in the case of each study may not be feasible. Lastly is the fact that machine learning categorizes PA by presence of ambulation but cut-points categorize PA by magnitude of the activity; thus, MVPA no longer would be an applicable term. At the same time, it is potentially more feasible for the general public to interpret prescribed activity (ie, “walk 30 min/day”) than to interpret prescribed intensity (ie, “spend 30 min/day being moderate to vigorously active”).

Given these limitations, future studies should include participants of various age and activity ranges, as well as men, to validate our findings further, especially those regarding the use of machine learning accelerometry. It also would be helpful to undertake a comparison using more than one self-report measure to quantify how much of the variation in agreement between self-report and accelerometry is due to the measure chosen. In addition, studies should be carried out which use less precisely tailored machine learning algorithms to assess how closely the training population must match the study population before the gains in sensitivity seen in machine learning over cut-point accelerometry are lost. Such work would be of value to the broader scientific community, as developing tailored machine learning algorithms can be computationally intensive (often requiring a computer scientist) and may not be valid if applied to populations that are different than the one on which it was trained.

Interpretation of Findings

The goal of this study was to present findings to inform measurement methods for future PA studies and in interpreting the findings of existing studies. In populations such as the one examined here, if the parameter of interest is accurate cross-sectional PA estimates, then self-report, accelerometry with 1041 cut-points, and machine learning appear to provide similar estimates. If the parameter of interest is change in PA, then self-report should be avoided, and an accelerometry based method should be used. Lastly, if the goal is detecting a difference of change between groups then machine learning and 1952 cut-points both appear to have appropriate sensitivity. Thus, if a study seeks to use one measurement method to examine all 3 of these things, baseline PA levels, estimates of change, and detection of a difference in PA change between groups, then our findings recommend machine learning as

an objective and minimally biased measurement option that still provides adequate sensitivity to detect change between groups.

Conclusion

We found that standard 1952 cutpoints provided estimates that were much lower than self-report. We also found that following the intervention both groups self-reported large increases in activity whereas accelerometry showed decreases in the control group and only slight increases in the intervention group. Thus, intervention researchers are facing the issue of self-report measures introducing bias and accelerometer cut-points being insensitive to perceived intensity. Machine learning algorithms can provide a new approach to bridge this gap. Although we may still wish to analyze multiple approaches to be able to compare to previous literature, researchers who wish to conduct studies that assess change in activity and provide estimates of activity that align with perceived intensity should consider the use of machine learning algorithms to process accelerometry data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This study was funded in part by the National Institutes of Health U54 CA155435, F32 CA220859.

References

1. Humphreys BR, McLeod L, Ruseski JE. Physical activity and health outcomes: evidence from Canada. *Health Econ.* 2014;23(1):33–54. [PubMed: 23364850]
2. Lee IM. Physical activity and cancer prevention – data from epidemiologic studies. *Med Sci Sports Exerc.* 2003;35(11):1823–1827. [PubMed: 14600545]
3. Speck RM, Courneya KS, Masse LC, et al. An update of controlled physical activity trials in cancer survivors: a systematic review and meta-analysis. *J Cancer Surviv.* 2010;4(2):87–100. [PubMed: 20052559]
4. Prince SA, Adamo KB, Hamel ME, et al. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *Int J Behav Nutr Phys Act.* 2008;5:56. [PubMed: 18990237]
5. Sallis JF, Saelens BE. Assessment of physical activity by self-report: status, limitations, and future directions. *Res Q Exerc Sport.* 2000;71(2 Suppl):S1–S14. [PubMed: 10925819]
6. Marinac CR, Godbole S, Kerr J, et al. Objectively measured physical activity and cognitive functioning in breast cancer survivors. *J Cancer Surviv.* 2015;9(2):230–238. [PubMed: 25304986]
7. Shilling V, Jenkins V, Morris R, et al. The effects of adjuvant chemotherapy on cognition in women with breast cancer--preliminary results of an observational longitudinal study. *Breast.* 2005;14(2):142–150. [PubMed: 15767184]
8. Copeland JL, Eslinger DW. Accelerometer assessment of physical activity in active, healthy older adults. *J Aging Phys Act.* 2009;17(1):17–30. [PubMed: 19299836]
9. Evenson KR, Wen F, Herring AH, et al. Calibrating physical activity intensity for hip-worn accelerometry in women age 60 to 91 years: the Women's Health Initiative OPACH Calibration Study. *Prev Med Rep.* 2015;2:750–756. [PubMed: 26527313]
10. Cerin E, Cain KL, Oyeyemi AL, et al. Correlates of agreement between accelerometry and self-reported physical activity. *Med Sci Sports Exerc.* 2016;48(6):1075–1084. [PubMed: 26784274]

11. Ellis K, Godbole S, Kerr J, et al. Multi-sensor physical activity recognition in free-living. *Proc ACM Int Conf Ubiquitous Comput.* 2014;2014:431–440. [PubMed: 26247061]
12. Patterson RE, Marinac CR, Natarajan L, et al. Recruitment strategies, design, and participant characteristics in a trial of weight-loss and metformin in breast cancer survivors. *Contemp Clin Trials.* 2016;47:64–71. [PubMed: 26706665]
13. Bull FC, Maslin TS, Armstrong T. Global Physical Activity Questionnaire (GPAQ): nine country reliability and validity study. *J Phys Act Health.* 2009;6(6):790–804. [PubMed: 20101923]
14. Chu AH, Ng SH, Koh D, et al. Reliability and validity of the self- and interviewer-administered versions of the Global Physical Activity Questionnaire (GPAQ). *PLoS One.* 2015;10(9):e0136944. [PubMed: 26327457]
15. Armstrong T, Bull F. Development of the World Health Organization Global Physical Activity Questionnaire (GPAQ). *J Public Health.* 2006;14(2):66–70.
16. Global Physical Activity Questionnaire (GPAQ) analysis guide. Available at: https://www.who.int/ncds/surveil-lance/steps/resources/GPAQ_Analysis_Guide.pdf. Accessed November 12, 2018.
17. Choi L, Liu Z, Matthews CE, et al. Validation of accelerometer wear and nonwear time classification algorithm. *Med Sci Sports Exerc.* 2011;43(2):357–364. [PubMed: 20581716]
18. Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. *Med Sci Sports Exerc.* 1998;30(5):777–781. [PubMed: 9588623]
19. Rosenberg D, Godbole S, Ellis K, et al. Classifiers for accelerometer-measured behaviors in older women. *Med Sci Sports Exerc.* 2017;49(3):610–616. [PubMed: 28222058]
20. Matthews CE. Calibration of accelerometer output for adults. *Med Sci Sports Exerc.* 2005;37(11 Suppl):S512–S522. [PubMed: 16294114]
21. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989;45(1):255–268. [PubMed: 2720055]
22. Dyrstad SM, Hansen BH, Holme IM, et al. Comparison of self-reported versus accelerometer-measured physical activity. *Med Sci Sports Exerc.* 2014;46(1):99–106. [PubMed: 23793232]
23. Nicaise V, Crespo NC, Marshall S. Agreement between the IPAQ and accelerometer for detecting intervention-related changes in physical activity in a sample of Latina women. *J Phys Act Health.* 2014;11(4):846–852. [PubMed: 23575348]
24. Kerr J, Marinac CR, Ellis K, et al. Comparison of accelerometry methods for estimating physical activity. *Med Sci Sports Exerc.* 2017;49(3):617–624. [PubMed: 27755355]
25. Shiroma EJ, Cook NR, Manson JE, et al. Comparison of self-reported and accelerometer-assessed physical activity in older women. *PLoS One.* 2015;10(12):e0145950. [PubMed: 26713857]
26. Winkler E, Waters L, Eakin E, et al. Is measurement error altered by participation in a physical activity intervention? *Med Sci Sports Exerc.* 2013;45(5):1004–1011. [PubMed: 23190599]
27. Haskell WL, Lee IM, Pate RR, et al. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. *Med Sci Sports Exerc.* 2007;39(8):1423–1434. [PubMed: 17762377]
28. Napolitano MA, Borradaile KE, Lewis BA, et al. Accelerometer use in a physical activity intervention trial. *Contemp Clin Trials.* 2010;31(6):514–523. [PubMed: 20723619]
29. Sloane R, Snyder DC, Demark-Wahnefried W, et al. Comparing the 7-day physical activity recall with a tri-axial accelerometer for measuring time in exercise. *Med Sci Sports Exerc.* 2009;41(6):1334–1340. [PubMed: 19461530]
30. Howlader N, Noone AM, Krapcho M, et al., eds. *SEER Cancer Statistics Review, 1975–2014.* Bethesda, MD: National Cancer Institute; 2016 Available at: https://seer.cancer.gov/csr/1975_2014/. Accessed March 16, 2019.
31. Courneya KS, Segal RJ, Mackey JR, et al. Effects of exercise dose and type on sleep quality in breast cancer patients receiving chemotherapy: a multicenter randomized trial. *Breast Cancer Res Treat.* 2014;144(2):361–369. [PubMed: 24554388]
32. Courneya KS, Segal RJ, McKenzie DC, et al. Effects of exercise during adjuvant chemotherapy on breast cancer outcomes. *Med Sci Sports Exerc.* 2014;46(9):1744–1751. [PubMed: 24633595]

33. Carayol M, Delpierre C, Bernard P, et al. Population-, intervention- and methodology-related characteristics of clinical trials impact exercise efficacy during adjuvant therapy for breast cancer: a meta-regression analysis. *Psychooncology*. 2015;24(7):737–747. [PubMed: 25483860]
34. Schmidt ME, Wiskemann J, Armbrust P, et al. Effects of resistance exercise on fatigue and quality of life in breast cancer patients undergoing adjuvant chemotherapy: a randomized controlled trial. *Int J Cancer*. 2015;137(2):471–480. [PubMed: 25484317]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.
Physical Activity Change over Time in a Randomized Controlled Trial, Stratified by Group,
for Each of the 3 Physical Activity Measurement Techniques (N = 289)

Characteristics of Overweight/Obese Breast Cancer Survivors in a Randomized Controlled Trial, Stratified by Intervention Group (N = 289)

Table 1

Variable	Control (N = 146)		Lifestyle Intervention (N = 143)	
	mean (SD)	mean (SD)	mean (SD)	mean (SD)
Age	62.6 (6.94)	63.0 (7.13)		
Age at diagnosis	60.0 (6.76)	60.3 (7.30)		
Years from Diagnosis to study	2.5 (1.78)	2.7 (2.09)		
Physical Health (SF-36 ^a)	46.0 (9.06)	46.3 (8.57)		
Mental Health (SF-36 ^a)	50.5 (9.54)	52.1 (8.76)		
BMI	31.1 (5.01)	31.1 (4.94)		
	N (%)	N (%)		
White	123 (84.2%)	118 (82.5%)		
Non-Hispanic	132 (90.4%)	123 (86.0%)		
College education or greater	74 (50.7%)	73 (51.0%)		
Received Chemotherapy	77 (52.7%)	73 (51.0%)		
Received Hormone Therapy	107 (73.3%)	96 (67.1%)		
Breast cancer stage				
I	67 (45.9%)	74 (51.7%)		
II	50 (34.2%)	49 (34.3%)		
III	29 (19.9%)	20 (14.0%)		

Note.

^a: Score based on the SF-36 self-report questionnaire, ranges 0–100 with higher scores corresponding to better health

Table 2

Distribution of Physical Activity in a Randomized Controlled Trial, as Measured by Self-report, Machine Learning, and Standard Cut-point Accelerometry – Stratified by Time and Intervention Group (N = 289; N_{control} = 146, N_{lifestyle intervention} = 143)

MVPA min/day by measurement (SD)	Baseline		Follow-up (6 month)		Change (Follow-up - Baseline)		Group difference of change	
	Pre-randomization		Control	Lifestyle Intervention	Control	Lifestyle Intervention	β	p-value
Self-report								
	56.2 (76.48)		72.6 (83.34)	99.7 (121.12)	22.7 (77.12)	36.9 (101.15)	14.21	.18
Machine Learning^a	67.8 (41.08)		61.5 (35.95)	76.7 (46.33)	-4.9 (29.50)	7.5 (35.88)	11.5	.003 [*]
Accelerometer								
1952 CP	20.0 (18.51)		17.1 (14.81)	23.3 (22.74)	-2.0 (13.34)	2.4 (17.16)	3.96	.028 [*]
1041 CP	62.0 (35.46)		55.0 (31.54)	66.3 (36.47)	-4.3 (22.75)	1.6 (28.95)	5.09	.095
760 CP	95.3 (45.81)		85.8 (41.59)	99.6 (45.47)	-5.5 (28.92)	0.2 (36.11)	4.54	.235

Note.

^{*}: Significant difference of change between groups (p < .05)

^a: Ambulation min/day used as estimate of MVPA min/day

Table 3

Comparison of Self-report, Machine Learning, and 1952 Cut-point Physical Activity Measures in a Randomized Controlled Trial (N = 289)

Time	Comparison	Mean difference min/day(SD)	Concordance Correlation	CC Lower	CC upper	Mixed effects p-value
Baseline	SR -ML	-11.5(75.10)	0.26*	0.14	0.38	.002
	SR - Accel	36.3 (72.09)	0.13*	0.07	0.18	<.001
	ML - Accel	47.8 (33.43)	0.23*	0.17	0.29	<.001
Lifestyle Intervention (N = 143)						
Follow-up (6 months)	SR -ML	23.0(114.19)	0.22*	0.12	0.31	<.001
	SR - Accel	76.5(115.33)	0.09*	0.05	0.13	<.001
	ML - Accel	53.4 (34.65)	0.27*	0.2	0.33	<.001
	Control (N = 146)					
	SR -ML	11.1(87.84)	0.06	-0.06	0.18	.061
Change (Follow-up - Baseline)	SR - Accel	55.5 (82.20)	0.04	0	0.08	<.001
	ML - Accel	44.4 (27.65)	0.21*	0.16	0.27	<.001
	Lifestyle Intervention (N = 143)					
	SR -ML	29.4(103.50)	0.07	-0.03	0.16	.001
	SR - Accel	34.6 (100.72)	0.03	-0.02	0.08	<.001
Change (Follow-up - Baseline)	ML - Accel	5.1 (30.55)	0.4*	0.3	0.5	.545
	Control (N = 146)					
	SR -ML	27.6(81.32)	0.03	-0.07	0.13	.001
	SR - Accel	24.7 (77.75)	0.01	-0.04	0.06	.003
	ML - Accel	-2.9 (26.76)	0.31*	0.21	0.42	.727

Note.

*. Significant correlation coefficient (p < .05)

(SR) = Self-Report, (ML)= Machine learning, (Accel) = 1952 Cut-point accelerometry, (CC) = Concordance correlation