Comparing Step Counting Algorithms for High-Resolution Wrist Accelerometry Data in NHANES 2011–2014

LILY KOFFMAN, CIPRIAN CRAINICEANU, and JOHN MUSCHELLI

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

ABSTRACT

KOFFMAN, L., C. CRAINICEANU, and J. MUSCHELLI. Comparing Step Counting Algorithms for High-Resolution Wrist Accelerometry Data in NHANES 2011–2014. *Med. Sci. Sports Exerc.*, Vol. 57, No. 4, pp. 746–755, 2025. **Purpose:** To quantify the relative performance of step counting algorithms in studies that collect free-living high-resolution wrist accelerometry data and to highlight the implications of using these algorithms in translational research. **Methods:** Five step counting algorithms (four open source and one proprietary) were applied to the publicly available, free-living, high-resolution wrist accelerometry data collected by the National Health and Nutrition Examination Survey (NHANES) in 2011–2014. The mean daily total step counts were compared in terms of correlation, predictive performance, and estimated hazard ratios of mortality. **Results:** The estimated number of steps were highly correlated (median, 0.91; range, 0.77–0.98), had high and comparable predictive performance of mortality (median concordance, 0.72; range, 0.70–0.73). The distributions of the number of steps in the population varied widely (mean step counts range from 2453 to 12,169). Hazard ratios of mortality associated with a 500-step increase per day varied among step counting algorithms between HR = 0.88 and 0.96, corresponding to a 300% difference in mortality risk reduction ([1–0.88]/[1–0.96] = 3). **Conclusions:** Different step counting algorithms provide correlated step estimates and have similar predictive performance that is better than traditional predictors of mortality. However, they provide widely different distributions of step counts and estimated reductions in mortality risk for a 500-step increase. **Key Words:** ACCELEROMETRY, MORTALITY, PHYSICAL ACTIVITY, PREDICTION

Objective estimation of the number of steps using high-resolution wrist accelerometry data has become increasingly important because: 1) large studies, such the National Health and Nutrition Examination Survey (NHANES) and UK Biobank now routinely collect, store, and share publicly high-resolution accelerometry data from wrist-worn wearable devices on tens or hundreds of thousands of study participants for many days at a time; 2) these data are often linked to high quality demographic, behavioral, and health information; 3) the number of daily steps is easy to understand and communicate and can be used as a target for interventions; 4) self-reported walking time or number of steps is subjective and often affected by uncontrollable bias and measurement error(1); and 5) accelerometry data and its summaries have been shown to be highly predictive of current and future health status (2–4).

Wrist-worn accelerometers provide measurements of the acceleration of the device attached to the wrist of a person. Specifically, data generated by wrist-worn accelerometers represent the acceleration expressed in Earth's gravitational units (g = 9.81 m·s^2) along three axes at high resolution, typically between 20 and 100 observations per second (20–100 Hz). The three axes represent the frame of reference of the device, which is related to, affected by, but not the same as a particular orientation of the hand and/or wrist. Step counting algorithms are applied to these high-resolution accelerometry data to produce an estimated number of steps at the second, minute, and/or day level. The downside is that these algorithms are often developed and validated on small data sets, in well controlled environments (e.g., walking on a treadmill) and with a limited number of tasks that do not represent the activity heterogeneity in the free-living environment (5). The upside is that some of these algorithms are open source, which could improve the harmonization of the "step count" across studies. Ultimately, this could lead to clear, evidence-based age- and sex-specific step counts in the population as well as estimation of the reduction in hazard associated with a specific increase in the number of steps per day (e.g., 500). However, the output of these algorithms are estimators, not true "step counts," and they vary depending on the particular type and version of the algorithm used.

Address for correspondence: Lily Koffman, M.Sc., Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, Baltimore, MD 21231; E-mail: lkoffma2@jh.edu.

Submitted for publication September 2024.

Accepted for publication November 2024.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.acsm-msse.org).

0195-9131/25/5704-0746/0

MEDICINE & SCIENCE IN SPORTS & EXERCISE®

DOI: 10.1249/MSS.0000000000003616

In this article, we compare the results of different step counting algorithms deployed on the same high-resolution accelerometry data collected in the NHANES 2011–2014 study and released in December 2022. The NHANES sample is representative of the US population and the devices were worn continuously for up to seven full days in the free-living environment. The size of the compressed data set is 20 terabytes and contains information for 14,693 study participants. Deploying the algorithms on these data took the equivalent of approximately 2.5 yr of computation time, with some algorithms being much faster than others (see Supplemental Table 2, Supplemental Digital Content, http://links.lww.com/MSS/D155). Because the criterion standard for the number of steps is not available in these data, we focus on studying the distributions (in the population and by age), correlation, predictive performance, and the estimated hazard ratios (HR) of mortality for an increase of 500 steps per day.

Understanding these differences in these quantities is crucial, as substantial confusion persists in the literature about the population and age/sex-specific reference distributions of the number of steps as well as the health effects associated with a specific increase in the number of steps. A big contributor to this confusion could be the bias induced by the step counting algorithms, even though the targets of inference (estimands) are thought to be well defined and intuitive. We summarize relevant estimates of step counts from the published literature in Table 1. Saint-Maurice (6) reported an average of 9124 steps in NHANES 2003–2006 using hip-worn accelerometers among 4840 US adults older than 40 yr. Tudor-Locke et al. (7) used only NHANES 2005–2006 and reported an average number 9676 steps among 3744 US adults older than 20 yr. Interestingly, Tudor Locke et al. (7) also reported an average of 6540 steps (32% fewer) by slightly tweaking the step counting algorithm. Small et al. (8) estimated an average of 9352 steps in the UK Biobank using a wrist-worn accelerometer and open source software on 75,943 UK adults older than 40 yr. This matches the average number of steps reported by Tudor-Locke et al. (7) in a comparable US population. Using the same population and data as Small et al. (8), Chan et al. (9) reported an average of 8016 steps, or 14% fewer. Lee et al. (10) reported an average of 5499 number of steps in the Women's Health Study using hip-worn accelerometers among 16,741 US women older than 62 yr. This is about 40% fewer steps than reported by Saint-Maurice et al. (6) in a younger population. Master et al. (11) reported an average of 7731 steps in the All of Us study using wrist-worn Fitbits among 6042 US adults with a median age of 57 yr. This is 15% fewer steps than reported by Saint-Maurice et al. (6), in a US population comparable in terms of age.

These differences may be due to the variations in the population, device sensitivity, device location, or algorithmic bias. Regardless of the cause, the differences have consequential implications. For example, the close consensus around an average of 9200 steps per day between NHANES 2003–2006 and UK Biobank would correspond to an average of 4.0 to 4.5 miles (6.5 to 7.2 km), or 1.2 to 1.5 h of walking per day (8.4–10.5 $h \cdot wk^{-1}$). This would be excellent news about physical activity (PA) in the US and the UK, as the current Physical Activity Guidelines for Americans recommend 2.5 to 5 h of moderate intensity activity per week and estimate that only 50% of Americans are meeting these recommendations (12,13). Brisk walking (2.5–4.0 mph) is a moderate intensity activity; we will not discuss walking speed in this article. However, overestimating the true average has negative consequences; an average of 9200 steps could seem out of reach and thereby could discourage efforts to increase activity. Moreover, an increase of 500 steps (about 4–7 min of walking) from an average of 9200 steps would correspond to a smaller relative change in behavior than if the true average number of steps were 5000.

The goals of this article are to: 1) obtain and publish minute-level step counts from the raw accelerometry in NHANES 2011–2014 using open source step counting methods; 2) describe patterns in step counts by age; 3) investigate the agreement between different step counting methods and their association with other summaries of accelerometry data (namely, Monitor Independent Movement Summary [MIMS] and Activity Counts [AC]); 4) compare the predictive performance of step estimates with other accelerometery-derived and non-accelerometry covariates on mortality; and 5) quantify the association between step counts and mortality in NHANES.

METHODS

Study Population

NHANES is a nationally representative study of about 5000 Americans per year. The study includes demographic,

TABLE 1. Summary of literature estimating steps from large epidemiological studies.

Study	Target Population	Age	Sex	Race/Ethnicity	Device	Algorithm	Wear Location	Daily Steps
Saint-Maurice et al. (6)	NHANES 2003–2006 (n = 4840)	Mean: 57 Min: 40	54% female	77% Non-Hispanic White	Actigraph	Actigraph proprietary	Hip	Mean: 9124
Tudor-Locke et al. (7)	NHANES 2005–2006 (n = 3744)	Min: 20	52% female	50% Non-Hispanic White	Actigraph	Actigraph proprietary	Hip	Mean (SE): 9676 (107) 6540 (106)*
Small et al. (8)	UK Biobank (n = 75943)	Min: 40	58% female	97% White	Axivity AX3	Open-source ML	Wrist	Median (IQR): 9352 (7099–11973)
Chan et al. (9)	UK Biobank (n = 78822)	Median: 57	54% female	Not available	Axivity AX3	Proprietary ML	Wrist	Mean (SD): 8016 (3321)
Lee et al. (10)	Women's Health Study (n = 16741)	Mean: 72 Min: 62	100% female	Not available	Actigraph	Actigraph proprietary	Hip	Mean: 5499
Master et al. (11)	All of Us (n = 6042)	Median: 57	72% female	84% White	Fitbit	Fitbit proprietary	Wrist	Median (IQR): 7731 (5867–9827)

*Censored estimate; steps among minutes with <500 activity counts removed.

socioeconomic, dietary, and health-related questions; data from the study are publicly available. All NHANES study participants 6 yr and older (2011–2012, Wave G) or 3 yr or older (2013–2014, Wave H) were asked to wear an ActiGraph GT3X+ on the nondominant wrist starting on the day of their examination at the NHANES Mobile Examination Center. They were instructed to wear the device at all times for seven consecutive days and remove the device on the morning of the ninth day. The devices collected triaxial acceleration at 80 Hz. A total of 6917 individuals in NHANES 2011–2012 and 7776 individuals in NHANES 2013–2014 had accelerometer data for analysis. 96% of participants with data wore the device until the ninth day and approximately 2% of participants wore the device for fewer than 7 d (14,15). More details on the NHANES accelerometer procedures can be found at https://wwwn.cdc.gov/nchs/data/nhanes/2011-2012/manuals/Physical_Activity_Monitor_Manual.pdf

Inclusion Criteria

A machine learning algorithm was used by NHANES to classify each minute of the day as wake wear, sleep wear, nonwear, or unknown (16). Minutes characterized as unknown were often short in duration (mean of 1.17 min) and the majority of the time were sandwiched between either two periods of wake wear (46% of unknown bouts) or two periods of sleep wear (28% of unknown bouts) (see Supplemental Table 3, Supplemental Digital Content, http://links.lww.com/MSS/D155). Thus, for the purpose of this paper unknown minutes were counted as wear. Minute-level data quality flags were also provided based on a number of rules indicating if any issues were detected (17). A single day was considered valid if it met the following three conditions: 1) at least 1368 min (95% of a full day) were classified as wake wear, sleep wear, or unknown and did not have any data quality flags; 2) at least 7 h (420 min) of the day were classified as wake wear; and 3) at least 7 h (420 min) had nonzero activity level data (MIMS, see below). Data from an individual were considered valid if they had at least 3 d with valid wear. Sensitivity analyses demonstrated that results were similar when including individuals with at least 1 d of valid wear (see Supplemental Tables 4 and 5, Supplemental Digital Content, http://links.lww.com/MSS/D155).

Accelerometery-Derived Physical Activity Summaries

MIMS and activity counts. Monitor Independent Movement Summary, an open source summary of raw accelerometry data (18), was calculated from the raw accelerometry data and provided by NHANES at the minute level in the PAXMIN files. Log base 10 MIMS, referred to throughout the rest of the manuscript as "log10 MIMS," were calculated by applying the $\log_{10}(1 + \text{MIMS})$ transformation at the minute level. One is added so that zero values are not excluded.

The ActiGraph Activity Count (AC) has been used in thousands of manuscripts. While previously only available as a proprietary algorithm, it was recently made open-source (19). This algorithm was implemented in R (20) using the package agcounter (21), which wraps ActiGraph's Python code to create activity counts (22), and applied to the raw accelerometry data to calculate activity counts at the second level. Log base 10 AC, referred to throughout the rest of the manuscript as "log10 AC," were calculated by summing the AC at the second level to obtain minute level ACs and applying the $\log_{10}(1 + \text{AC})$ transformation at the minute level.

Steps. Four open source step counting methods: ADaptive Empirical Pattern Transformation (ADEPT) (23), Oak (24), Verisense (25,26), and Stepcount(8); and one proprietary algorithm, Actilife (27), were applied to the raw NHANES data to estimate steps for each participant. A summary of the algorithms is included in Supplemental Table 1 (Supplemental Digital Content, http://links.lww.com/MSS/D155). The R package adept (23) was used to implement ADEPT, the R package walking (28) was used to implement Oak and Verisense, and the R package stepcount 29 was used to implement both the random forest (RF) and self-supervised learner (SSL) versions of Stepcount (8). Both the original (26) and revised (25) version of Verisense were used; for more details on each algorithm, see (5). The estimated steps for the entire NHANES data set using these algorithms, the associated software, and accompanying vignette are publicly available at https://github.com/lilykoff/nhanes_steps_mortality.

For all step estimates, MIMS, log10 MIMS, AC, and log10 AC, values were first summed at the minute level. To obtain day totals, values from valid minutes were summed over each day; again, valid minutes were defined as minutes that did not have any data quality flags and were classified by the wear algorithm as wake wear, sleep wear, or unknown. Day totals were then averaged across valid days to obtain one summary for each individual and physical activity variable. These represent values for the "average day."

Mortality and Other Covariates

The following variables from the NHANES 2011–2014 data were extracted: age, sex (male/female), race/ethnicity (non-Hispanic White, non-Hispanic Black, Mexican-American, other Hispanic, other [including multi-race]), education level (less than high school, high school/high school equivalent, more than high school), body mass index (BMI) category (underweight, normal, overweight, obese), diabetes, coronary heart disease (CHD), congestive heart failure (CHF), heart attack, stroke, cancer, alcohol consumption (never drinker, former drinker, moderate drinker, heavy drinker, missing alcohol), cigarette smoking (never smoker, former smoker, current smoker), mobility problem, and self-reported health status (poor, fair, good, very good, excellent). We refer to these variables throughout the manuscript as "traditional predictors."

Mortality data were obtained from the national death registries public-use linked mortality files and is available publicly for participants ≥18 yr old (see https://www.cdc.gov/nchs/data-linkage/mortality-public.htm). The mortality files include

vital status (assumed alive or assumed deceased) and person-months of follow-up time from the mobile examination center visit to the date of death or the end of the mortality follow-up period (December 31, 2019). Mortality data were merged with the covariates data.

Statistical Analysis

For each step counting algorithm, the survey weighted mean and standard error of step counts by age was estimated using the survey(30) package in R following the NHANES weighting guidelines. The means and associated 95% confidence intervals (CI) were smoothed as functions of age using locally weighted smoothing (31). The agreement between different step counting methods was quantified using the Spearman and Pearson correlations between the mean daily step counts from each method, mean daily AC, and mean daily MIMS.

To evaluate the predictive performance of all covariates we first fit separate univariate Cox proportional hazards regression models on mortality for each predictor among individuals between 50 and 79 yr old at screening (n = 3638; number of deaths = 412; median follow up time = 6.75 yr). Age was restricted to a maximum of 79 because individuals 80 and over are topcoded at 80 yr of age in NHANES. A sensitivity analysis was performed including individuals 80 and older and results were similar (see Supplemental Tables 4 and 5, Supplemental Digital Content, http://links.lww.com/MSS/D155). For each model, a 10-fold survey weighted cross-validated concordance (cvC) (32) was calculated. This calculation was repeated 100 times and the average of these 100-cvC values is reported for stability (33,34,35). Since all PA covariates were right skewed, they were Winsorized at the 99th percentile before model fitting (see Supplemental Fig. 1, Supplemental Digital Content, http://links.lww.com/MSS/D155).

A series of four models was used to investigate the association between covariates and mortality. The first model included traditional predictors. The second model contained the same covariates as the first model plus the mean total daily MIMS. The third model contained the same covariates as the first model and mean total steps, as estimated by the step algorithm with the highest univariate cvC. The final model contained the same covariates as the first model, total MIMS, and total steps. The 100-times repeated 10-fold cvC is calculated for each model, along with the coefficient estimate and P value associated with the steps variable.

To estimate the association of step counts with mortality, separate multivariable Cox proportional hazards regression models were fit for each step count algorithm. All models included the traditional predictors and one step count algorithm. Separate models were fit using both raw and standardized (centered and scaled) step counts. The covariate-adjusted HR of mortality was calculated for: 1) an increase of 500 in mean daily steps; and 2) a one standard deviation increase in mean daily steps. To account for a nonlinear (dose–response) relationship between steps and log hazard of mortality, steps from each algorithm were analyzed in quartiles and using restricted cubic splines, with knots at the 5th, 50th, and 95th percentiles (36).

Finally, three stratified analyses were performed to assess potential differences in both single and multivariable concordance by age (50–62-yr-olds vs 63– to 79-yr-olds), self-reported health (excellent, very good, or good vs. fair or poor), and mobility problem (no mobility problem vs. mobility problem).

All statistical analyses were performed using R version 4.4.1. The R packages tidy models (37) and survival (38) were used for model fitting and cross-validation.

RESULTS

After applying the inclusion criteria described in Section 2.2 and excluding everyone younger than 18 yr, the analytic sample contained 4303 individuals from 2011 to 2012 and 4361 from 2013 to 2014. For the mortality prediction analysis, individuals who were younger than 50 yr, older than 79 yr, or had missing covariates were excluded, resulting in an analytic sample of 3368 (1795 from 2011 to 2012 and 1843 from 2013 to 2014). Table 2 presents survey weighted characteristics of individuals 18 yr and older with valid accelerometry data by wave and overall. Supplemental Figure 2 summarizes the inclusion/exclusion process, and Supplemental Table 6 (Supplemental Digital Content, http://links.lww.com/MSS/D155) presents the nonsurvey weighted characteristics of the individuals included in the mortality analysis.

Step Counts by Age

Table 3 presents the estimated survey weighted means and standard deviations for step counts from each algorithm by wave and age category. Across all groups, Actilife consistently estimates the highest mean step count, although the estimates from Actilife, Oak, and Stepcount RF are similar. ADEPT estimates the lowest step counts. For all methods, the average number of steps decreases for individuals 50 yr and older. Verisense (original and revised) and stepcount SSL produce similar step count estimates. The final column of Table 3 presents the absolute difference divided by the mean in the 2011–2012 and 2013–2014; a higher number indicates larger differences between waves. Interestingly, the percent difference in step estimates between waves among all adults ranges from 3.1% (Actilife) to 8.5% (ADEPT), and differences observed for any step algorithm are larger than those observed for MIMS, AC, log10 MIMS, or log10 AC (range, 0.84–1.8%).

Figure 1 panel A displays the smoothed, survey weighted mean step counts and associated 95% confidence intervals by age for each step counting method. Actilife (13,000 at age 40 yr and 8500 at age 80 yr), Oak (13,000 at age 40 yr and 6000 at age 80 yr), and Stepcount RF (13,000 at age 40 yr and 5500 at age 80 yr) have the highest step count estimates across all ages. These methods estimate a loss of approximately 110 steps per year from age 40 to 80 yr, or 1% per year. Stepcount SSL and both Verisense versions estimate around 10,000 steps at age 40 yr and 4000 to 5500 steps at age

TABLE 2. Survey weighted population characteristics for individuals aged 18 yr and older with valid accelerometry data

Characteristics	Overall, *N* = 8664	2011–2012, *n* = 4367	2013–2014, n = 4297
Sex			
Female	4552 (53%)	2281 (52%)	2271 (53%)
Male	4112 (47%)	2086 (48%)	2027 (47%)
Age (yr)	48.01 (17.41)	47.99 (17.29)	48.04 (17.53)
Race/ethnicity			
Non-Hispanic White	5785 (67%)	2918 (67%)	2867 (67%)
Non-Hispanic Black	979 (11%)	505 (12%)	475 (11%)
Other race—including multirace	640 (7.4%)	312 (7.2%)	327 (7.6%)
Mexican American	736 (8.5%)	344 (7.9%)	393 (9.1%)
Other Hispanic	524 (6.0%)	288 (6.6%)	236 (5.5%)
Education level			
More than HS	5279 (63%)	2664 (63%)	2615 (63%)
Less than HS	1330 (16%)	703 (17%)	627 (15%)
HS/HS equivalent	1788 (21%)	877 (21%)	911 (22%)
Missing	267	123	144
BMI category			
Normal	2436 (28%)	1258 (29%)	1178 (28%)
Obese	3176 (37%)	1538 (36%)	1639 (38%)
Overweight	2825 (33%)	1445 (33%)	1380 (32%)
Underweight	156 (1.8%)	84 (2.0%)	71 (1.7%)
Missing	70	41	29
Diabetes	887 (10%)	430 (9.9%)	457 (11%)
Missing	2	0	2
Coronary heart failure	247 (2.9%)	135 (3.2%)	113 (2.7%)
Missing	270	127	143
Congenital heart disease	316 (3.8%)	145 (3.4%)	171 (4.1%)
Missing	283	133	150
Stroke	263 (3.1%)	140 (3.3%)	124 (3.0%)
Missing	267	124	143
Alcohol use			
Never drinker	1057 (12%)	477 (11%)	581 (14%)
Former drinker	1233 (14%)	619 (14%)	614 (14%)
Moderate drinker	2657 (31%)	1401 (32%)	1256 (29%)
Heavy drinker	669 (7.7%)	374 (8.6%)	296 (6.9%)
Missing alcohol	3048 (35%)	1496 (34%)	1552 (36%)
Smoking status			
Never smoker	4820 (56%)	2351 (55%)	2470 (57%)
Former smoker	2087 (24%)	1071 (25%)	1016 (24%)
Current smoker	1630 (19%)	820 (19%)	810 (19%)
Missing	127	126	1
Mobility problem	1411 (17%)	638 (15%)	773 (19%)
Missing	270	123	146
General health condition			
Poor	232 (2.7%)	106 (2.4%)	126 (2.9%)
Fair	1323 (15%)	612 (14%)	711 (17%)
Good	3408 (39%)	1670 (38%)	1738 (40%)
Very good	2742 (32%)	1442 (33%)	1300 (30%)
Excellent	959 (11%)	536 (12%)	423 (9.8%)
Died by 5 yr follow-up	648 (7.5%)	354 (8.1%)	294 (6.8%)
Missing	15	12	4

Binary or categorical variables presented as *n* (%) and continuous variables presented as mean (SD).

80 yr; a loss of approximately 110 to 150 steps per year. ADEPT estimates the lowest average step count (3000 at age 40 yr and 1000 at age 80 yr); a loss of 50 steps or approximately 1.6% per year from age 40 to 80 yr.

Panel B plots the estimated per-year percent difference in mean daily steps, which are consistent with 2% to 1% increases between ages 20 and 30 yr, small changes between 30 and 40, 1% to 4% decreases between 50 and 70, and 2% to 4% decreases between 70 and 75. Steeper declines occur after age 75 yr, especially for ADEPT.

Correlation between Step Counting Algorithms

Figure 2 displays the Spearman (top) and Pearson (bottom) correlations between mean daily step counts from each algorithm, MIMS, log10 MIMS, AC, and log10 AC. All Spearman correlations between step counting algorithms are larger than 0.8, with the exception of Actilife and ADEPT (0.77). The highest Spearman correlations are between Verisense and Verisense revised (0.98), Verisense revised and Oak (0.97), Verisense and Oak (0.96), and Actilife and Oak (0.96). Estimates of step counting algorithms are also highly correlated with widely used AC and MIMS measurements. The Spearman correlations are smaller for ADEPT (ranging from 0.5 to 0.6) and Stepcount SSL (ranging from 0.59 to 0.7). The highest Spearman correlations are between MIMS and Actilife (0.93), AC and Actilife (0.92), and Oak and MIMS (0.90). Interestingly, the Spearman correlation between AC and MIMS is extremely high (0.99) as is the correlation between log 10 AC and log 10 MIMS (0.98). The Pearson correlations reflect similar patterns and are most different from the Spearman correlations for ADEPT.

Mortality Prediction

Individual predictors of mortality. Figure 3 and Supplemental Table 7 (Supplemental Digital Content, http://links.lww.com/MSS/D155) present the 100 times repeated 10-fold cross validated survey weighted concordance from univariate Cox regression models. The mean or proportion of each variable by deceased and nondeceased groups are also presented. The nine variables with the highest concordance are all measures of physical activity; the seven models with the highest concordance are all step algorithms, followed by AC (0.688), MIMS (0.682), mobility problem (0.675), and age (0.673). The highest concordance corresponds to steps estimated with the Stepcount RF algorithm (C = 0.732). In stratified analyses, single variable concordance was consistently higher among models with just adults aged 63 to 79 yr compared with models with adults aged 50 to 62 yr; no patterns were observed in models stratified by self-reported health or mobility limitations (see Supplemental Table 10, Supplemental Digital Content, http://links.lww.com/MSS/D155). Concordance was lower for models using quartiles of steps compared to continuous steps, but the respective ordering of the PA variables remained similar, and concordance for all step variables remained higher than any traditional predictor (see Supplemental Table 8, Supplemental Digital Content, http://links.lww.com/MSS/D155).

The added predictive performance of PA summaries to the traditional risk factors of mortality. For the multivariable mortality analysis, four models were fit. Each model included traditional predictors; the model with just these variables is referred to as "traditional predictors." Then, three other models are considered: traditional predictors and Stepcount RF, traditional predictors and MIMS, and traditional predictors, Stepcount RF, and MIMS. Table 4 presents the 100 times repeated 10-fold cross-validated survey weighted concordance for each model, along with the estimated coefficients and *P* values from a model fit on all of the data. The concordance is highest for the model with traditional predictors and

TABLE 3. Survey weighted mean (SD) physical activity totals by wave and age among individuals with at least three valid days of accelerometry data

	2011–2012		2013–2014		% Difference in Waves	
	Age 50+ yr $n = 2107$	All Adults $N = 4303$	Age 50+ yr $n = 2146$	All Adults $N = 4361$	Age 50+ yr	All Adults
Actilife	11,195 (4001)	12,169 (3995)	10,850 (4008)	11,902 (4048)	3.1	2.2
ADEPT	2342 (1593)	2659 (1569)	2151 (1680)	2453 (1575)	8.5	8.1
Oak	10,254 (5027)	11,794 (5061)	9733 (4886)	11,381 (5065)	5.2	3.6
Stepcount						
RF	9888 (5542)	11,509 (5722)	9502 (5442)	11,263 (5764)	4.0	2.2
SSL	8358 (4402)	9144 (4400)	8027 (4388)	8846 (4399)	4.0	3.3
Verisense				2453 (1575)		
Original	8337 (4027)	9497 (4062)	7974 (4019)	9163 (4065)	4.5	3.6
Revised	7532 (4521)	9102 (4756)	7122 (4402)	8725 (4730)	5.6	4.2
MIMS						
Raw	12,435 (3642)	13,572 (3758)	12,243 (3574)	13,467 (3784)	1.6	0.78
log10	960 (158)	998 (154)	952 (160)	994 (155)	0.84	0.40
AC						
Raw	2,333,927 (791,711)	2,573,262 (815,192)	2,291,942 (766,303)	2,549,846 (816,305)	1.8	0.91
log10	2878 (383)	2955 (366)	2847 (394)	2933 (370)	1.1	0.75

The final column shows the percent difference in estimates between 2011–2012 and 2013–2014: $\frac{|est_{2011-12} - est_{2013-14}|}{mean(est_{2011-12}, est_{2013-14})}$.

steps (0.776), though traditional predictors and MIMS (0.773) and traditional predictors, MIMS, and steps (0.774) were close. Furthermore, the coefficient for steps remains statistically significant even after the addition of MIMS, suggesting that total steps per day may confer additional information about mortality beyond that provided by MIMS and the traditional predictors

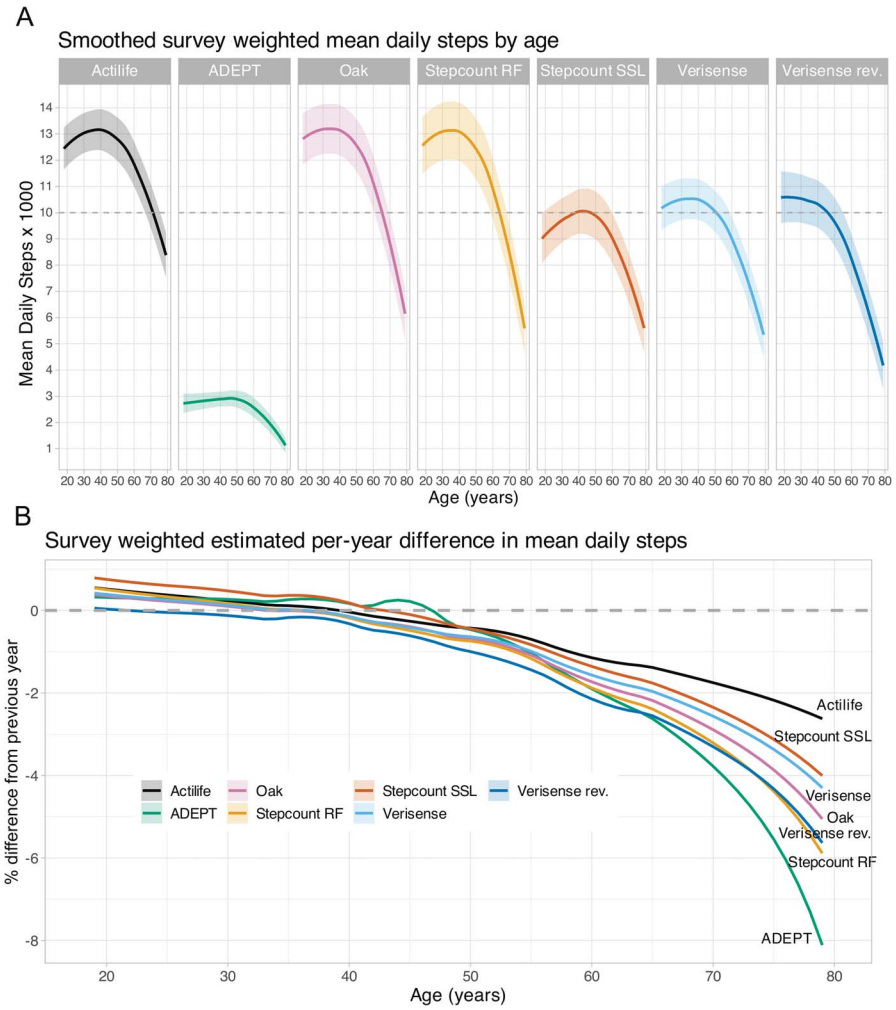

FIGURE 1—Panel A: survey weighted mean and associated 95% confidence intervals for total daily step counts, by algorithm and age. The horizontal dashed line indicates the 10,000 steps for reference. Panel B: percent change in smoothed survey weighted mean daily steps for each year compared with the previous year, by algorithm. The horizontal dashed line indicates no change (0%) for reference.

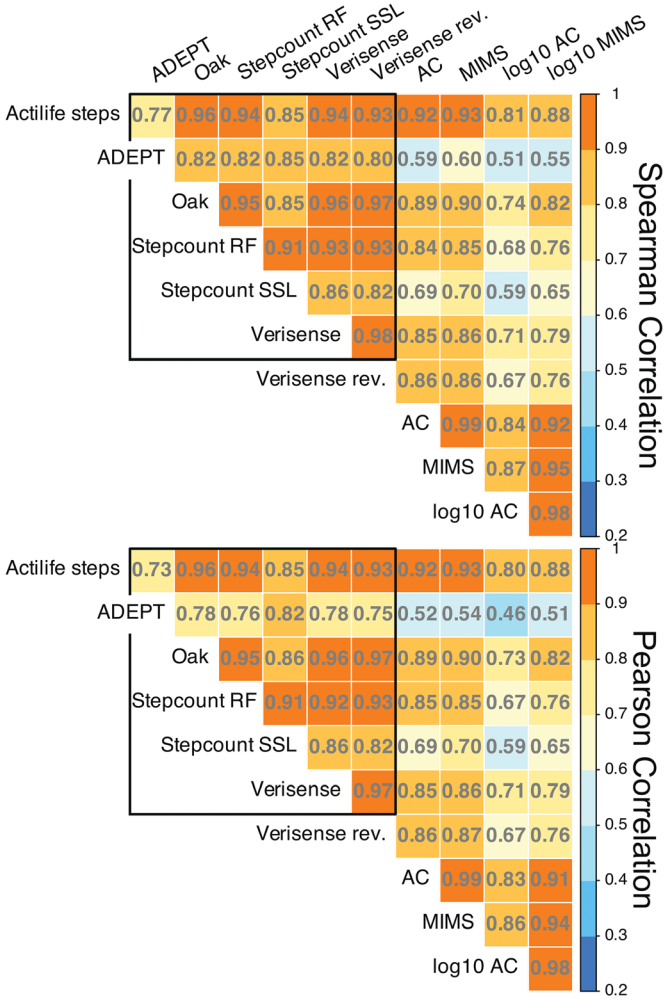

FIGURE 2—Spearman correlation (top) and Pearson correlation (bottom) between mean daily step counts from different algorithms and mean daily PA summaries. Correlations are not survey weighted since the interest is in correlation between raw estimates.

used in the model. Furthermore, the confidence intervals for the HR associated with an 500-step increase overlap, indicating the effect of steps on mortality does not change significantly even after adjusting for MIMS.

Table 5 presents the adjusted estimated HR and 95% confidence intervals associated with a 500-step increase (raw) and a one standard deviation increase (scaled), which is algorithm-specific, in mean steps per day. The HR for all methods except ADEPT associated with a 500-step increase in mean steps are between 0.95 and 0.96; ADEPT is 0.88. Achieving a HR of 0.88 would be require a 11,000 to 14,000 increase in the number of steps for the other algorithms. The scaled HR associated with a one standard deviation increase are in close agreement for Actilife, ADEPT and Stepcount SSL (0.67), though a one standard deviation increase from ADEPT would correspond to an increase of 1500 steps, whereas all other algorithms would require an increase of at least 4000 steps to achieve the same reduction in risk. Oak and Verisense revised also agree very closely with each other (HR, 0.63), though Oak would require an increase of almost 5000 steps whereas Verisense revised would require an increase of 4000 steps to achieve the same reduction in risk. Verisense estimates a HR of 0.65 for an increase of one standard deviation (equivalent to 4000) steps, somewhere between Oak and Actilife. Stepcount RF estimates the smallest HR (most improvement) for an increase of one standard deviation (equivalent to 5400) steps. Note that all these algorithms are based on the same raw accelerometry data and differences are due only to the step counting algorithms. While ADEPT estimates a smaller number of steps than other algorithms for the same reduction in mortality risk, achieving an increase of 500 ADEPT estimated steps may be as difficult as an increase of 2000 Oak estimated steps. Indeed, recall that we only have the outcomes of step counting algorithms and not the actual gold standard of the number of steps.

Dose–response relationship between steps and mortality. Supplemental Table 9 and Supplemental Figure 3 (Supplemental Digital Content, http://links.lww.com/MSS/D155) present the results of the multivariable regressions modeling physical activity variables as quartiles, and Figure 4 (Supplemental Digital Content, http://links.lww.com/MSS/D155) presents the results of the cubic spline analysis. Interestingly, a dose–response relationship between steps and

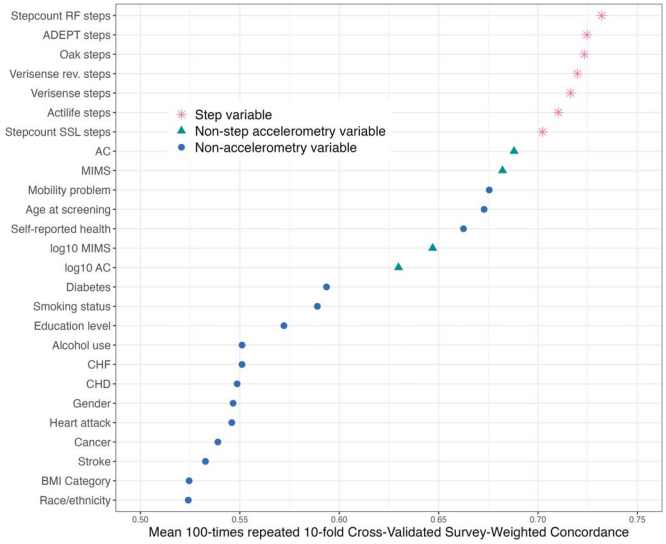

FIGURE 3—100 times repeated 10-fold cross-validated survey weighted concordance from univariate Cox proportional hazards models.

mortality is observed for both Stepcount and Verisense methods, but less so for Oak, and not at all for Actilife.

DISCUSSION

We estimated steps in NHANES 2011–2014 using four open-source and one proprietary step counting algorithm based on the high-resolution three axial wrist accelerometry data. The open-source data (minute-level step counts for all algorithms and individuals in NHANES) published with this manuscript and accompanied by a vignette detailing analyses in this article will provide a valuable resource for future refinements of step counting algorithms and analysis of walking patterns among a nationally representative sample of Americans.

There was substantial heterogeneity between step counting algorithms. The estimates for mean total steps among all adults by wave varied from 2453 (2013–14; ADEPT) to 12,169 (2011–12; Actilife). Differences between waves and within methods ranged from 2.2% (Actilife, Stepcount RF) to 8.1% (ADEPT), which are all larger than the differences observed between waves for MIMS or AC (<1%); this discrepancy warrants further investigation. The differences in estimates between algorithms can be partially explained by the details of the algorithms. ADEPT was developed for accurate stride segmentation and was thus optimized for specificity, not sensitivity (39) Actilife was developed explicitly for hip wear (40), and thus may not be appropriate in the wrist-worn setting. Oak, based on the continuous wavelet transform, is a "one-size fits most" approach and was designed to work relatively well in a variety of settings and populations. Verisense is based on peak finding in the acceleration signal and requires several tuning parameters for both peak identification and filtering to remove peaks that are not walking; the revised algorithm was intended to overcome bias of the original algorithm in free-living settings. Stepcount is also based on peak finding, but first uses machine learning to identify periods of walking. Excluding ADEPT and Actilife, the estimates from the other algorithms are more similar: mean total steps range from 8725 (Verisense revised, 2013–14) to 11,792 (Oak, 2013–14); a difference of approximately 30%.

The number of steps for all methods were very highly correlated with a minimum Spearman correlation of 0.77, a median of 0.91, and a maximum of 0.98. In univariate Cox models predicting mortality, the mean total step count obtained by any algorithm outperformed all other accelerometry-derived summaries of PA (AC, MIMS, log10 AC, log10 MIMS) as well as traditional risk factors including age, self-reported health, and mobility problem.

In multivariable models, steps remained highly significant even after the inclusion of MIMS, the NHANES-provided summary of PA. For all step counting methods, an increase in steps was associated with a decrease in the hazard of mortality, but HR associated with a 500-step increase varied from 0.88 to 0.96, again indicating t.he scale of the differences in these step estimates. The HR for a one standard deviation increase in the number of steps also varied between methods from 0.59 to 0.67.

TABLE 4. Multivariable model summaries

Model	Steps HR*	Steps *P*	Model Concordance
Traditional predictors only	—	—	0.769 (0.769, 0.770)
Traditional predictors + MIMS	—	—	0.773 (0.772, 0.774)
Traditional predictors + Stepcount RF steps	0.955 (0.940, 0.970)	<0.001	0.776 (0.775, 0.777)
Traditional predictors + Stepcount RF steps + MIMS	0.961 (0.939, 0.983)	0.036	0.774 (0.773, 0.775)

For each model, the HR and associated *P* value for steps per day is reported, along with the average of 100 times repeated 10-fold cross-validated model concordance.
*HR associated with an increase of 500 steps per day.

TABLE 5. Adjusted HR and associated 95% confidence intervals associated with a 500-unit increase in steps (raw) and one standard deviation increase in steps (scaled).

Step Algorithm	Adjusted HR; Raw	Adjusted HR; Scaled	SD of Steps (×100)
Actilife	0.95 (0.93, 0.97)	0.67 (0.58, 0.78)	4.0
ADEPT	0.88 (0.83, 0.93)	0.67 (0.56, 0.80)	1.5
Oak	0.95 (0.94, 0.97)	0.63 (0.54, 0.74)	4.9
Verisense	0.95 (0.93, 0.97)	0.65 (0.55, 0.76)	4.0
Verisense revised	0.95 (0.93, 0.97)	0.63 (0.54, 0.75)	4.4
Stepcount RF	0.95 (0.94, 0.97)	0.59 (0.50, 0.70)	5.4
Stepcount SSL	0.96 (0.94, 0.97)	0.67 (0.57, 0.79)	4.4

The standard deviation is in thousands of steps and can be considered what a one-unit increase in the scaled predictor represents in terms of steps. HR are obtained from weighted Cox proportional hazards regression models adjusting for the traditional predictors described: age, sex, BMI, race, diabetes, CHF, self-reported health, CHD, heart attack, stroke, cancer, alcohol use, smoking, mobility problem, and education.

Analyses in this paper indicate that studies that use one of these step counting algorithms are likely to provide similar findings with respect to prediction of mortality. However, they also indicate that the estimators of the number of steps from different studies may not be comparable when using different algorithms or even slight modifications of the same algorithm. The most straightforward way to harmonize step counts across studies is to apply the same algorithm to the raw accelerometry data. From a translation perspective, choosing a particular step counting algorithm may have substantial effects on health recommendations and interventions. For example, if Stepcount RF were used to estimate the number of steps for health recommendations, the average number of steps for a person older than 50 yr in the United States would be around 12,000 and an increase of 500 steps per day would correspond to a 5% reduction in risk (as measured by the mortality hazard). These estimators for the number of steps are larger than those published by Saint-Maurice (6) (average of 9124 steps in US adults older than 40 yr), Tudor-Locke et al. (7) (average of 9676 steps in US adults older than 20 yr), and Small et al. (8) (average of 9352 in UK adults older than 40 yr). This number of steps could appear daunting for individuals older than 50 yr and may result in discouraging people from even attempting to increase their PA. We argue that there is need to 1) closely investigating the accuracy of these estimates and identify potential sources of over-estimation of the number of steps; and 2) create a trusted source of open-source step counting algorithms that are version-controlled and calibrated to NHANES and, possibly, other studies.

Steps provide a measure of physical activity that is highly predictive of mortality. Indeed, our results support other findings on the relationship between step counts and mortality (8,10,11,41). Monitor Independent Movement Summary and AC are widely used activity measures, but are not directly translatable to the public. Minutes in conditions, such as light, moderate, and moderate-to-vigorous activity are translatable, but require thresholds and these thresholds have not been agreed upon. Steps do not require thresholding to estimate and are translatable to the general public. However, the large variation in estimates derived from different step counting algorithms limits exactly what is translated to the public, affects public health recommendations, and ultimately undermines the credibility of using step counts in research. Currently, we can recommend that "more steps are better," but we cannot say how many more are optimal. More development is needed to validate existing step counting algorithms in free living settings; in order to do so, more free-living accelerometry data sets with ground truth step counts and activity labels are needed.

Strengths of our analysis include the use of a large, nationally representative study. Our code is open-source and our analysis is reproducible. Limitations include the exclusion of individuals with insufficient wear; it is possible that these individuals have different patterns of PA, though assessing this exceeds the scope of this article. Inaccuracy in the NHANES wear detection algorithm could also have affected the inclusion of study participants. Furthermore, the association between steps and mortality found in our analysis should not be interpreted as causal and the short follow up time (median, 6.75 yr) in the study creates concern about reverse causality. Future directions include exploring the relationship between steps estimated by various step counting algorithms and mortality in other large, nationally representative studies.

This work was supported by the National Institutes of Health under Grant R01NS060910 and Grant R01AG075883. Ciprian Crainiceanu is consulting for Bayer and Johnson and Johnson on methods development for wearable and implantable technologies. The details of these contracts are disclosed through the Johns Hopkins University eDisclose system. The research presented here is not related to and was not supported by this consulting work. The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation. The results of the present study do not constitute endorsement by the American College of Sports Medicine.

This work was supported by the National Institutes of Health under Grant R01NS060910 and Grant R01AG075883. Ciprian Crainiceanu is consulting for Bayer and Johnson and Johnson on methods development for wearable and implantable technologies.

REFERENCES

1. Prince SA, Adamo KB, Hamel M, Hardt J, Connor Gorber S, Tremblay M. 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.
2. Leroux A, Xu S, Kundu P, et al. Quantifying the predictive performance of objectively measured physical activity on mortality in the UK Biobank. *J Gerontol A Biol Sci Med Sci*. 2021;76(8):1486–94.
3. Wanigatunga AA, Di J, Zipunnikov V, et al. Association of total daily physical activity and fragmented physical activity with mortality in older adults. *JAMA Netw Open*. 2019;2(10):e1912352.
4. Smirnova E, Leroux A, Cao Q, et al. The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: National Health and Nutritional Examination Survey 2003–2006. *J Gerontol A Biol Sci Med Sci*. 2020;75(9):1779–85.
5. Koffman L, Muschelli J. Evaluating step counting algorithms on subsecond wristworn accelerometry: a comparison using publicly available data sets. *J Meas Phys Behav*. 2024;7(1):jmpb.2024-0008.
6. Saint-Maurice PF, Troiano RP, Bassett DR Jr., et al. Association of daily step count and step intensity with mortality among US adults. *JAMA*. 2020;323(12):1151–60.

7. Tudor-Locke C, Johnson WD, Katzmarzyk PT. Accelerometer-determined steps per day in US adults. *Med Sci Sports Exerc*. 2009; 41(7):1384–91.
8. Small SR, Chan S, Walmsley R, et al. Development and validation of a machine learning wrist-worn step detection algorithm with deployment in the UK Biobank. *medRxiv [Preprint]*. 2023;20.23285750.
9. Chan LLY, Choi TCM, Lord SR, Brodie MA. Development and large-scale validation of the watch walk wrist-worn digital gait biomarkers. *Sci Rep*. 2022;12(1):16211.
10. Lee IM, Shiroma EJ, Kamada M, Bassett DR, Matthews CE, Buring JE. Association of step volume and intensity with all-cause mortality in older women. *JAMA Intern Med*. 2019;179(8):1105–12.
11. Master H, Annis J, Huang S, et al. Association of step counts over time with the risk of chronic disease in the all of us research program. *Nat Med*. 2022;28(11):2301–8.
12. Piercy KL, Troiano RP, Ballard RM, et al. The physical activity guidelines for Americans. *JAMA*. 2018;320(19):2020–8.
13. Abildso CG, Daily SM, Umstattd Meyer MR, Perry CK, Eyler A. Prevalence of meeting aerobic, muscle-strengthening, and combined physical activity guidelines during leisure time among adults, by rural-urban classification and region—United States, 2020. *MMWR Morb Mortal Wkly Rep*. 2023;72(4):85–9.
14. Centers for Disease Control and Prevention. NHANES 2011–2012: Physical Activity Monitor Data (PAXMIN_G). Accessed: 2024-05-01. https://wwwn.cdc.gov/Nchs/ Nhanes/2011–2012/PAXMIN_G.htm.
15. Centers for Disease Control and Prevention. NHANES 2013–2014: Physical Activity Monitor Data (PAXMIN_H). Accessed: 2024-05-01. https://wwwn.cdc.gov/Nchs/Nhanes/2013–2014/PAXMIN_H.htm.
16. Thapa-Chhetry B, Arguello DJ, John D, Intille S. Detecting sleep and nonwear in 24-h wrist accelerometer data from the National Health and nutrition examination survey. *Med Sci Sports Exerc*. 2022; 54(11):1936–46.
17. National Center for Health Statistics (NCHS). Data quality flag summary table for the Physical Activity Monitor (PAM) data collected in NHANES 2011–2014 and NNYFS; 2024. Accessed: 2024-09-24. https://wwwn.cdc.gov/nchs/nhanes/Pam/ Default.aspx.
18. John D, Tang Q, Albinali F, Intille S. An open-source monitor-independent movement summary for accelerometer data processing. *J Meas Phys Behav*. 2019;2(4):268–81.
19. Neishabouri A, Nguyen J, Samuelsson J, et al. Quantification of acceleration as activity counts in ActiGraph wearable. *Sci Rep*. 2022; 12(1):11958.
20. R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria; 2024. Available from: https://www.R-project.org/.
21. Muschelli J. agcounter: processes accelerometry data to Actigraph Activity Counts (AC); 2024. R package version 0.2.0.9000. Available from: https://github.com/jhuwit/agcounter.
22. LLC A. Code for the technical report on the ActiLife counts algorithm.; 2024. Python module version 0.2.6. Available from: https://github.com/actigraph/agcounts.
23. Karas M, Straczkiewicz M, Fadel W, Harezlak J, Crainiceanu CM, Urbanek JK. Adaptive empirical pattern transformation (ADEPT) with application to walking stride segmentation. *Biostatistics*. 2021; 22(2):331–47.
24. Straczkiewicz M, Huang EJ, Onnela JP. A "one-size-fits-most" walking recognition method for smartphones, smartwatches, and wearable accelerometers. *NPJ Digit Med*. 2023;6(1):29.
25. Rowlands AV, Maylor B, Dawkins NP, et al. Stepping up with GGIR: validity of step cadence derived from wrist-worn research-grade accelerometers using the verisense step count algorithm. *J Sports Sci*. 2022;40(19):2182–90.
26. Patterson MR. Verisense toolbox. GitHub;2020. https://github.com/ShimmerEngineering/Verisense-Toolbox/tree/master/Verisense_step_algorithm.
27. ActiGraph. *ActiLife Software*. Pensacola, FL; 2015.
28. Muschelli J. walking: Segments Accelerometry Data into Walking using the Python 'forest' Module; 2024. R package version 0.3.0.
29. Muschelli J. stepcount: Esimate Step Counts from Accelerometry Data; 2024. R package version 0.1.1.
30. Lumley T. survey: analysis of complex survey samples; 2023. R package version 4.2. [31] Cleveland WS, Grosse E, Shyu WM. Local Regression Model. In: Chambers JM, Hastie TJ, editors. *Statistical Models in S. Wadsworth & Brooks/Cole*. 1992.
31. Cleveland WS, Grosse E, Shyu WM. Local regression models. In: Chambers JM, Hastie TJ, editors. *Statistical models in S*. New York, New York, Wadsworth & Brooks/Cole; 1992.
32. Harrell FE Jr. Evaluating the yield of medical tests. *JAMA*. 1982; 247(18):2543–6.
33. Meng Q, Cui E, Leroux A, Mowry EM, Lindquist MA, Crainiceanu CM. Quantifying the association between objectively measured physical activity and multiple sclerosis in the UK Biobank. *Med Sci Sports Exerc*. 2023;55(12):2194–202.
34. Zhao A, Cui E, Leroux A, Lindquist MA, Crainiceanu CM. Evaluating the prediction performance of objective physical activity measures for incident Parkinson's disease in the UK Biobank. *J Neurol*. 2023;270(12):5913–23.
35. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J Chem*. 2014;6(1):10.
36. Desquilbet L, Mariotti F. Dose–response analyses using restricted cubic spline functions in public health research. *Stat Med*. 2010;29(9): 1037–57.
37. Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.; 2020. Available from: https://www.tidymodels.org.
38. Therneau TM. A Package for Survival Analysis in R; 2023. R package version 3.5–5. Available from: https://CRAN.R-project.org/package=survival.
39. Karas M, Urbanek JK, Illiano VP, Bogaarts G, Crainiceanu CM, Dorn JF. Estimation of free-living walking cadence from wrist-worn sensor accelerometry data and its association with SF-36 quality of life scores. *Physiol Meas*. 2021;42(6).
40. Toth L, Paluch AE, Bassett DR Jr., et al. Comparative analysis of ActiGraph step counting methods in adults: a systematic literature review and meta-analysis. *Med Sci Sports Exerc*. 2023;56(1):53–62.
41. Hamer M, Blodgett JM, Stamatakis E. Dose–response association between step count and cardiovascular disease risk markers in middle-aged adults. *Scand J Med Sci Sports*. 2022;32(7):1161–5.