International Journal of Epidemiology, 2026, **55(2)**, dyag014
https://doi.org/10.1093/ije/dyag014
Original Article

International Epidemiological Association
OXFORD

Reproducibility and associated regression dilution bias of accelerometer-derived physical activity and sleep in UK Biobank

Charilaos Zisou[1,2], Catherine Calvin[1,3], Hannah Taylor[1,3], Ben Lacey[1,3], Imen Hammami[1], Rosemary Walmsley[1,2], Tessa Strain[4,5], Katrien Wijndaele[5], Nicholas Wareham[5], Soren Brage[5], Karl Smith-Byrne[1], Derrick Bennett[1], Sarah Lewington[1,6], Jemma C. Hopewell[1,‡], Aiden Doherty[1,2,‡,*]

[1]Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom
[2]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, United Kingdom
[3]UK Biobank, Stockport, United Kingdom
[4]Physical Activity for Health Research Centre, University of Edinburgh, Edinburgh, United Kingdom
[5]MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom
[6]Health Data Research UK, University of Oxford (HDRUK-Oxford), Oxford, United Kingdom
*Corresponding author. Professor of Biomedical Informatics, Nuffield Department of Population Health, University of Oxford, Big Data Institute, Old Road Campus, Oxford, OX3 7LF, United Kingdom. E-mail: aiden.doherty@ndph.ox.ac.uk
‡Joint senior authors.

Abstract

Background: **Previous studies on the reproducibility of 7-day accelerometer measurements have been limited by small sample sizes and short follow-up periods. We aimed to assess the long-term reproducibility of accelerometer-derived physical activity and sleep, and to illustrate the impact of regression dilution bias on the association between daily step count and coronary heart disease (CHD) in UK Biobank.**

Methods: **We analysed data from 3138 UK Biobank participants in the main accelerometry sub-study with up to four repeat accelerometer measurements after 3–4 years. Nine physical activity and sleep phenotypes were extracted to capture different movement behaviours. Reproducibility was assessed by using intraclass correlation coefficients (ICCs). The impact on disease associations was illustrated by considering daily step count and incident CHD using Cox regression (87 038 participants; 3879 CHD events), before and after correction for regression dilution.**

Results: **Among the 3138 participants, 51% were women and the mean (SD) age was 63.1 (9.4) years. Reproducibility was good for overall activity, with an ICC (95% confidence interval) of 0.75 (0.74–0.76), and moderate for other phenotypes, with ICCs ranging from 0.58 (0.56–0.59) for sleep efficiency to 0.69 (0.68–0.70) for sedentary behaviour. In our example, the inverse association between daily step count and CHD showed a 20% lower risk of CHD per usual 4000 steps after correcting for regression dilution compared with 13% before correction.**

Conclusion: **Accelerometer measurements are moderately reproducible and comparable to measures such as blood pressure. Correction for regression dilution bias is crucial to quantify associations of usual physical activity and sleep with disease risk.**

Keywords wearables, measurement error, within-person variation, intra-individual variability, cardiovascular disease

Received: 16 January 2025. **Accepted:** 14 January 2026

Key Messages

- Accelerometer-derived physical activity and sleep are increasingly used in large cohort studies, but the long-term reproducibility of these measures and the extent to which their associations with health outcomes are impacted by regression dilution bias remain poorly understood.
- In this study of repeat 7-day accelerometer measurements in 3138 UK adults, agreement between measurements 3.7 years apart ranged from 58% for sleep efficiency to 75% for overall activity, and an illustrative example demonstrated how consequent regression dilution bias can attenuate associations with disease risk.
- The incorporation of repeat accelerometer measurements in cohort studies will allow researchers to assess their reproducibility and correct epidemiological associations for regression dilution bias.

Introduction

Physical inactivity is associated with several common diseases, including cardiovascular disease and type 2 diabetes [1]. Variations in sleep behaviour are also linked to health outcomes, such as cardiometabolic and mental health conditions [2]. To better understand these relationships, large cohort studies have increasingly incorporated device-based measurements of movement behaviours using wearable accelerometers [3–5]. These measurements, which are less prone to information biases than self-reported questionnaires and can track movements continuously throughout the day, have provided novel insights into how movement behaviours relate to health. For example, the inverse associations between device-measured physical activity and cardiovascular and all-cause mortality are much stronger than previously observed from self-reported data [6]. Moreover, device-based studies challenge prior findings that long sleep duration is associated with worse health outcomes [7, 8].

Previous study designs have typically instructed participants to wear an accelerometer for 7 consecutive days, balancing the need for extensive data collection with logistical challenges such as device battery life and participant compliance [9]. However, the long-term reproducibility of 7-day accelerometer measurements (i.e. the degree of agreement of repeat measurements within the same participants over time) remains uncertain, largely due to previous studies having small sample sizes and collecting repeat measurements over short intervals [10–14]. Within-person variation from random measurement error, short-term fluctuations, and longer-term changes in activity patterns may lead to measurements that do not accurately reflect an individual's 'usual' movement behaviours. This variation systematically weakens observed associations with health outcomes—an effect known as regression dilution bias [15]. Estimation of the reproducibility of accelerometer-derived phenotypes enables correction for this bias [16]. Although correction methods are commonly applied to exposures such as blood pressure and cholesterol [15, 17, 18], they have rarely been utilized in analyses of movement behaviours [19], especially with device-based data.

This study aimed to (i) investigate the long-term reproducibility of accelerometer-derived phenotypes of physical activity and sleep by using repeat 7-day measurements and (ii) illustrate the impact of regression dilution bias through an example examining the association between daily step count and incident coronary heart disease (CHD).

Methods

Study design and population

Details of the UK Biobank study design have been extensively reported elsewhere [20]. Briefly, between 2006 and 2010, 0.5 million participants aged 40–69 years from England, Scotland, and Wales were recruited in a prospective cohort study (response rate: 5.5%). During a baseline assessment-centre visit, participants completed touchscreen questionnaires, were interviewed by trained nurses, and had anthropometric measurements and biological samples taken [21]. A subset of 103 712 individuals also participated in the main accelerometry sub-study between 2013 and 2015 [4]. Participants were mailed an Axivity AX3 triaxial accelerometer (100 Hz) and asked to wear it continuously on their dominant wrist for 7 days and return it by post.

Between 2017 and 2019, a sample of 5067 participants from the main accelerometry sub-study with good compliance (near-complete 7-day wear and prompt device return) were invited by e-mail to undergo repeat measurements (Fig. 1). This sample was selected by using a stratified random sampling procedure to ensure equal representation across sex and age groups (10-year intervals). Participants were asked to provide four measurements ∼3 months apart, allowing some to miss initial measurements but complete subsequent ones. Of those invited, 3208 (63%) accepted the invitation. Devices were dispatched seasonally over four cycles: November 2017–May 2018 (3202 devices dispatched, 2634 measurements released), February–August 2018 (3125 devices dispatched, 3095 measurements released), May–November 2018 (3070 devices dispatched, 3039 measurements released), and August 2018–February 2019 (3051 devices dispatched, 3012 measurements released). Minor differences between devices returned and data released were largely due to reading issues; however, the first cycle was additionally subject to device malfunctions in 17% of devices distributed.

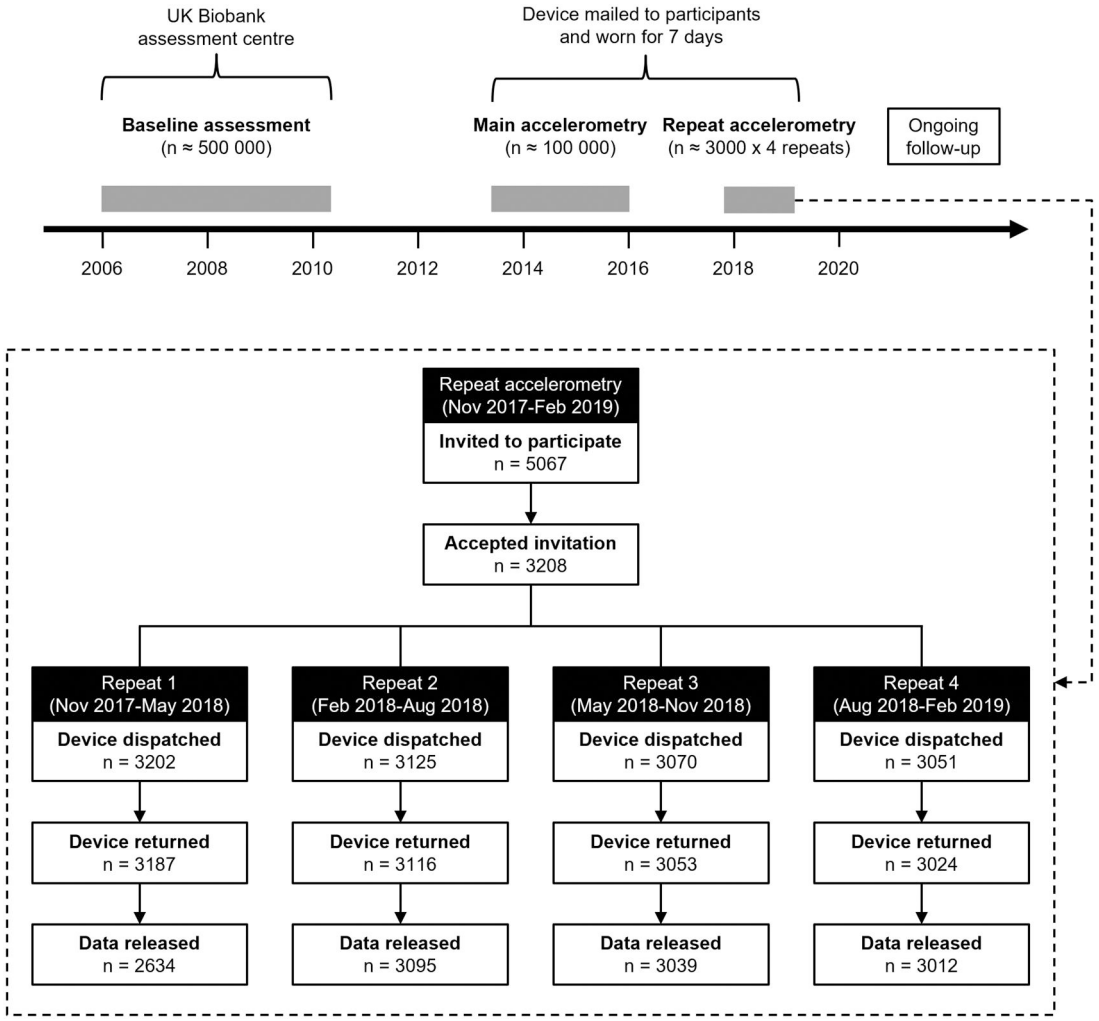

Figure 1 Timeline of UK Biobank study assessments and flow diagram of repeat accelerometry sub-study participants. Differences in devices dispatched between repeat cycles are largely due to participants' opting out. Differences between devices returned and data released within each repeat cycle are largely due to issues with reading the accelerometer data; however, the first cycle was additionally subject to device malfunction in 17% of the devices distributed.

Overall, 38 participants provided one measurement, 91 provided two, 716 provided three, and 2353 provided all four.

Accelerometer-derived phenotypes of physical activity and sleep

Raw accelerometer time series were processed by using established methods and the average overall activity was extracted (Supplementary Methods) [4]. Validated machine-learning models were used to classify movement behaviours at the 30-second level and derive summary phenotypes, averaged over the week to provide daily values. A self-supervised neural network and a peak detection algorithm were used to measure the step count, defined as the median number of steps per day (https://github.com/OxWearables/stepcount) [22]. Peak 30-minute cadence was calculated as the average steps per minute from the 30 most active (not necessarily consecutive) minutes of the day. Daily durations of moderate-to-vigorous physical activity (MVPA), light physical activity, sedentary behaviour, and time spent in bed were estimated by using a random forest and hidden Markov model (https://github.com/OxWearables/biobankAccelerometerAnalysis) [23]. Overnight sleep duration, defined as the median hours of sleep per night, and sleep efficiency, calculated as the ratio of the overnight sleep duration to the total time spent in bed, were derived by using a self-supervised neural network (https://github.com/OxWearables/asleep) [7].

Ascertainment of cardiovascular outcomes

Incident CHD events were identified through linkage to UK hospital inpatient records, operations data, and the national death register from the end of accelerometer wear until the earliest of: (i) the end of the available data linkage (31 October 2022, 31 August 2022, or 31 May 2022 for participants in England, Scotland, and Wales, respectively), (ii) loss to follow-up, or (iii) death (Supplementary Table S1). Additional cardiovascular

outcomes were examined to illustrate generalizability (Supplementary Methods).

Statistical analysis

Participants with low-quality accelerometer data, namely those whose device could not be calibrated, had unrealistically high acceleration values (>100 mg) or insufficient wear time (<3 days of wear or missing the same hour across all days), were excluded [4]. For sleep phenotypes, we additionally excluded days with <22 hours of wear [7]. Data meeting these criteria are referred to as 'valid data'. Accelerometer phenotypes were compared across the main and repeat measurements by using descriptive statistics; continuous variables were approximately normally distributed and expressed as mean [standard deviation (SD)], except for MVPA and sleep efficiency, which were skewed and described as median [interquartile range (IQR)].

To assess reproducibility, participants were divided into fifths based on their main accelerometer measurements and changes in mean values within those fifths were tracked after 4 years. Although individual measurements may reflect random within-person variation, calculating the group means of the repeat measurements yields unbiased estimates of usual group levels at the time of measurement [15]. Additionally, intraclass correlation coefficients (ICCs) were calculated, which assess reproducibility and can also be used to correct for regression dilution bias [16]. The 4-year interval between measurements is appropriate, as it approximates the midpoint of follow-up in the prospective association analyses described next and supports the estimation of disease risk in relation to long-term average physical activity levels during the study period [15]. ICCs and 95% confidence intervals (CIs) were estimated by using a two-way mixed-effects model with absolute agreement and were interpreted as poor (<0.5), moderate (0.5–0.75), good (0.75–0.9), and excellent (>0.9), in accordance with established guidelines [24]. Statistical information was incorporated from all measurements by calculating the ICCs between the main and each of the repeat measurements and obtaining their inverse-variance-weighted average (Supplementary Methods). To account for seasonal variations, measurements were first regressed on the season of accelerometer wear and the residuals were used to calculate the ICCs. For skewed phenotypes (MVPA and sleep efficiency), ICCs were assessed with and without inverse-normal transformations by using a rank-based approach. Stratified analyses were conducted to assess ICCs by subgroups of age in the main accelerometry sub-study (<60, ≥60 years), sex, body mass index (<25, ≥25 kg/m^2), and prior self-reported illness or disability (healthy, prior disease). Sensitivity analyses were conducted among participants with four valid repeat measurements to assess whether variations in measurements across cycles reflected differences in participant characteristics.

Regression dilution bias, resulting from random measurement error and short- and long-term changes in activity levels, leads to a systematic underestimation of the observed epidemiological associations. To illustrate this, we considered the association between daily step count and CHD as an example, which has been consistently observed in previous studies [25, 26]. Among all 103 712 main accelerometry participants, we excluded those who had withdrawn, had low-quality accelerometer data (quality criteria described above), had atherosclerotic cardiovascular disease before the main accelerometer wear (to limit reverse causation), or had missing covariates. Cox regression models were used to estimate the hazard ratios (HRs) and 95% CIs for CHD, with the step count analysed across fifths to assess the shape of the association and as a continuous variable (per 4000 steps, approximately equal to the SD), adjusting for potential confounders identified in prior literature as common determinants of both exposure and outcome (Supplementary Methods and Supplementary Table S2). To assess the potential for reverse causation bias, sensitivity analyses were conducted by excluding CHD events occurring within the first 2 years of follow-up or participants with self-reported poor health, as done in previous studies [23, 26].

To illustrate the correction for regression dilution bias visually, we plotted the HR of each step count fifth defined by the main accelerometer measurements against the inverse-variance-weighted average over the repeat measurements, estimating each fifth's usual daily step count [15]. The log HR per 4000 steps and its standard error were divided by the corresponding ICC to obtain the HR per usual 4000 steps [16]. A sensitivity analysis was used to assess differences in estimates when age- and sex-specific corrections were applied (Supplementary Methods). Additional cardiovascular outcomes were analysed as illustrative examples to demonstrate generalizability (Supplementary Methods).

Accelerometer data were processed in Python (version 3.9) and statistical analyses were conducted in R (version 4.3).

Results

Reproducibility analysis—participant characteristics

After quality control, 3138 participants with valid accelerometer data from the main sub-study and at least one valid repeat measurement were included in the reproducibility analysis (Table 1 and Supplementary Fig. S1). The mean (SD) age was 63.1 (9.4) years, 1613 (51.4%) were women, and 3038 (96.8%) identified as White. Baseline characteristics were broadly similar to those invited but not enrolled in the repeat accelerometry sub-study, though non-participants were less likely to have a university degree or be taking blood-pressure medication (Supplementary Table S3). Compared with participants in the main accelerometry sub-study, those with repeat measurements were younger at baseline (53.8 vs. 56.7 years), less likely to be women (51.4% vs. 56.3%), more likely to have a university degree, and less likely to be taking blood-pressure or cholesterol medication (Table 1). Compared with the full UK Biobank cohort, the accelerometry participants generally exhibited higher socio-economic status, lower rates of smoking, a lower body mass index, and a lower prevalence of health conditions and medications, potentially indicative of a healthy volunteer effect [27].

Repeat measurements were obtained 3.2–4.0 years [mean (SD) of 3.6 (0.7)] after the main accelerometry sub-study (Table 2). Participants in all cycles had the same mean age in the main sub-study and the proportion of women remained consistent. Data were primarily collected in winter (70%) during

Table 1 Baseline characteristics of all UK Biobank participants, those who participated in the main accelerometry sub-study, and those who had at least one valid repeat accelerometer measurement.

Characteristic at baseline assessment	Baseline assessment 2006–10 ($n = 502\,364$)	Main accelerometry 2013–15 ($n = 95\,981$)	Repeat accelerometry 2017–19 ($n = 3138$)
Demographic and socio-economic factors			
Age at baseline (years)	57.0 (8.1)	56.7 (7.8)	53.8 (9.3)
Age at each assessment (years)	57.0 (8.1)	62.4 (7.8)	63.1 (9.4)
Women [*n* (%)]	273 297 (54.4)	54 038 (56.3)	1613 (51.4)
White [*n* (%)]	472 569 (94.1)	92 687 (96.6)	3038 (96.8)
College or university degree [*n* (%)]	161 102 (32.1)	41 289 (43.0)	1476 (47.0)
Townsend Deprivation Index	−2.1 (−3.6, 0.5)	−2.4 (−3.8, −0.2)	−2.3 (−3.7, 0)
Lifestyle factors [*n* (%)]			
Current smoker	52 960 (10.5)	6613 (6.9)	195 (6.2)
Daily alcohol drinker	101 745 (20.3)	21 966 (22.9)	681 (21.7)
Fruits and vegetables ≥ 8 servings/day	183 677 (36.6)	35 557 (37.0)	1091 (34.8)
Physical and blood measurements			
Body mass index (kg/m^2)	27.4 (4.8)	26.7 (4.5)	26.4 (4.6)
Systolic blood pressure (mmHg)	137.9 (18.6)	136.5 (18.2)	134.9 (17.8)
Resting heart rate (beats/min)	69.4 (11.3)	68.5 (10.9)	68.3 (11.0)
LDL cholesterol (mmol/L)	3.6 (0.9)	3.6 (0.8)	3.5 (0.8)
Self-reported conditions and medication [*n* (%)]			
Blood-pressure medication	103 982 (20.7)	16 240 (16.9)	408 (13.0)
Cholesterol medication	86 874 (17.3)	13 475 (14.0)	350 (11.2)
Long-standing illness or disability	159 853 (31.8)	26 762 (27.9)	827 (26.4)
Poor overall health	22 768 (4.5)	2397 (2.5)	77 (2.5)
Atherosclerotic cardiovascular disease[a]	37 956 (7.6)	7472 (7.8)	252 (8.0)

LDL, low-density cholesterol.
Mean (SD) or median (IQR) shown for continuous variables. *N* (%) shown for categorical variables. Characteristics were measured during the baseline assessment (2006–10), unless otherwise stated. Participants with low-quality accelerometer data were excluded.
[a] Diagnosed prior to each assessment, as recorded in hospital records or self-reported.

Cycle 1, spring (78%) during Cycle 2, summer (78%) during Cycle 3, and autumn (79%) during Cycle 4. The mean levels of physical activity and sleep were similar across the main and repeat measurement cycles, with small seasonal differences.

Reproducibility of accelerometer-derived phenotypes of physical activity and sleep

The mean values of all phenotypes within the fifths defined by the main measurements moderately converged after 4 years, displaying a pattern of regression towards the mean with small seasonal variations (Fig. 2). The top fifth converged more than the bottom fifth for physical activity phenotypes, while the opposite trend was observed for sedentary behaviour and sleep phenotypes. The reproducibility of the examined phenotypes was moderate to good [24], with the ICC (95% CI) for overall activity being 0.75 (0.74–0.76) and the rest ranging from 0.58 (0.56–0.59) for sleep efficiency to 0.69 (0.68–0.70) for sedentary behaviour (Fig. 2 and Supplementary Table S4). There was no material difference between the ICCs with and without inverse-normal transformations; therefore, only the latter are presented. Reproducibility was generally higher for older (≥60 years) compared with younger (<60 years) individuals and for women compared with men (Supplementary Table S5). Minor differences in reproducibility were observed across some physical activity phenotypes between normal/underweight (<25 kg/m^2) and overweight/obese (≥25 kg/m^2) individuals (Supplementary Table S6). Individuals with prior illness or disability exhibited slightly higher reproducibility compared with healthy individuals, with step count showing a more pronounced difference. Restricting the analysis to participants with four valid repeat measurements did not substantially alter the results (Supplementary Tables S7 and S8).

Illustrative example: impact of regression dilution bias on associations between daily step count and cardiovascular diseases

Of the 103 712 participants with accelerometer data, 87 038 remained in the association analysis after the exclusion of those who withdrew from the study ($n = 52$), had low-quality accelerometer data ($n = 7679$), had prevalent atherosclerotic cardiovascular disease ($n = 7472$), or had missing covariates ($n = 1471$) (Supplementary Fig. S2). During a median (IQR) follow-up of 7.9 (7.3–8.4) years, 3879 incident CHD events occurred. Every additional 4000 steps per day were associated with a 13% lower risk of CHD (HR 0.87, 95% CI 0.84–0.90) after adjustment for potential confounders with no correction for regression dilution bias (Fig. 3). Plotting the estimated HRs against the inverse-variance weighted group means of the repeat measurements demonstrated a substantially stronger aassociation.After correction for

Table 2 Accelerometer-derived phenotypes of physical activity and sleep across main and repeat accelerometry sub-studies.

Phenotype	Main accelerometry (*n* = 3138)	Repeat 1 (*n* = 2361)	Repeat 2 (*n* = 2751)	Repeat 3 (*n* = 2665)	Repeat 4 (*n* = 2653)
Time since main accelerometry sub-study (years)	0	3.2 (0.6)	3.5 (0.6)	3.7 (0.6)	4.0 (0.6)
Age at main accelerometry sub-study (years)	59.5 (9.5)	59.6 (9.4)	59.5 (9.5)	59.4 (9.4)	59.5 (9.5)
Age at each assessment (years)	59.5 (9.5)	62.9 (9.4)	63.0 (9.5)	63.1 (9.4)	63.4 (9.4)
Women [*n* (%)]	1613 (51.4)	1220 (51.7)	1411 (51.3)	1365 (51.2)	1348 (50.8)
Season [*n* (%)]					
Spring	690 (22.0)	694 (29.4)	2153 (78.3)	9 (0.3)	0 (0)
Summer	910 (29.0)	0 (0)	588 (21.4)	2085 (78.2)	12 (0.5)
Autumn	948 (30.2)	11 (0.5)	0 (0)	571 (21.4)	2091 (78.8)
Winter	590 (18.8)	1656 (70.1)	10 (0.4)	0 (0)	550 (20.7)
Physical activity phenotypes					
Overall activity (mg)	29.0 (8.7)	27.1 (8.9)	28.5 (8.9)	28.1 (8.7)	27.1 (8.5)
Daily step count (steps)	9618 (3863)	9035 (4001)	9845 (3982)	9828 (4181)	9148 (3886)
Peak 30-minute cadence (steps/minute)	93.6 (16.5)	92.3 (18.5)	93.6 (17.0)	92.6 (16.8)	92.5 (17.7)
MVPA (hours)	0.60 (0.31, 1.03)	0.50 (0.24, 0.91)	0.63 (0.31, 1.06)	0.62 (0.30, 1.05)	0.56 (0.26, 0.97)
Light physical activity (hours)	5.0 (1.6)	4.8 (1.6)	5.0 (1.6)	4.9 (1.6)	4.8 (1.6)
Sedentary behaviour (hours)	10.4 (1.8)	10.6 (1.9)	10.5 (1.9)	10.6 (1.9)	10.6 (1.9)
Sleep phenotypes[a]					
Total daily time in bed (hours)	7.8 (1.0)	7.9 (1.1)	7.8 (1.0)	7.7 (1.0)	7.9 (1.1)
Overnight sleep duration (hours)	6.8 (1.0)	6.9 (1.0)	6.7 (1.0)	6.8 (1.0)	6.9 (1.0)
Sleep efficiency	0.86 (0.81, 0.90)	0.86 (0.80, 0.90)	0.86 (0.81, 0.91)	0.87 (0.81, 0.91)	0.86 (0.81, 0.91)

Mean (SD) or median (IQR) shown for continuous variables. *N* (%) shown for categorical variables. Seasons were defined as: Winter (December, January, February), Spring (March, April, May), Summer (June, July, August), and Autumn (September, October, November).

[a] After additional measurement-quality-related exclusions, descriptive statistics for sleep phenotypes were calculated from the following number of participants: Main (*n* = 3119), Repeat 1 (*n* = 2291), Repeat 2 (*n* = 2707), Repeat 3 (*n* = 2619), and Repeat 4 (*n* = 2610).

regression dilution bias (by dividing the logHR and standard error by the corresponding ICC of 0.62), every additional usual 4000 steps per day were associated with a 20% lower risk of CHD (HR 0.80, 95% CI 0.76–0.85; excluding the first 2 years of follow-up or participants with poor health did not materially alter the estimate; Supplementary Table S9). Sensitivity analysis using age- and sex-specific corrections for regression dilution bias resulted in highly consistent estimates (Supplementary Table S10). Across different cardiovascular outcomes, a consistent pattern of underestimation was observed before correction, as expected, as regression dilution bias is a consequence of within-person variation in the exposure (Supplementary Fig. S3).

Discussion

In the largest study of repeat 7-day accelerometer measurements to date, the mean levels of physical activity and sleep within the groups defined by their initial accelerometer measurements moderately converged after 4 years, indicating regression towards the mean. Reproducibility was moderate to good, with overall activity reaching 75% and individual phenotypes ranging from 58% for sleep efficiency to 69% for sedentary behaviour. Reproducibility was generally higher in women compared with men and in older compared with younger individuals. Furthermore, correction for regression dilution bias was shown to have a material impact on the epidemiological associations, underestimating the effects by nearly 40% in our example of daily step count and CHD.

Our reproducibility results are broadly consistent with those of previous smaller studies using hip-worn accelerometers, which reported overall activity estimates of between 76% and 83% [10, 13, 14]. Furthermore, a study of 1679 adults reported a correlation of 55% between repeat measurements of step counts taken ∼3.7 years apart, which closely aligns with our estimate [28]. For MVPA, light physical activity, and sedentary behaviour, the reproducibility estimates are more variable, at between 59% and 77% [10, 12–14]. This heterogeneity is possibly driven by differences in the study time frames and population characteristics, particularly age and sex. For sleep phenotypes, a study with repeat wrist-worn accelerometer measurements over 1 year reported correlations of 67% for total time in bed, 76% for overnight sleep duration, and an unexpectedly high 90% for sleep efficiency, though these measurements relied on participant input [29]. In contrast to these smaller studies, our analysis identified clear differences in reproducibility by age, sex, and prior disability status.

Compared with self-reports, accelerometer-derived physical activity and sleep data exhibit markedly higher reproducibility, further highlighting the value of incorporating device-based measurements in large-scale epidemiological studies. A study of 19 000 UK Biobank participants reported the reproducibility of self-reported total physical activity at ∼50% after 4.3 years, with

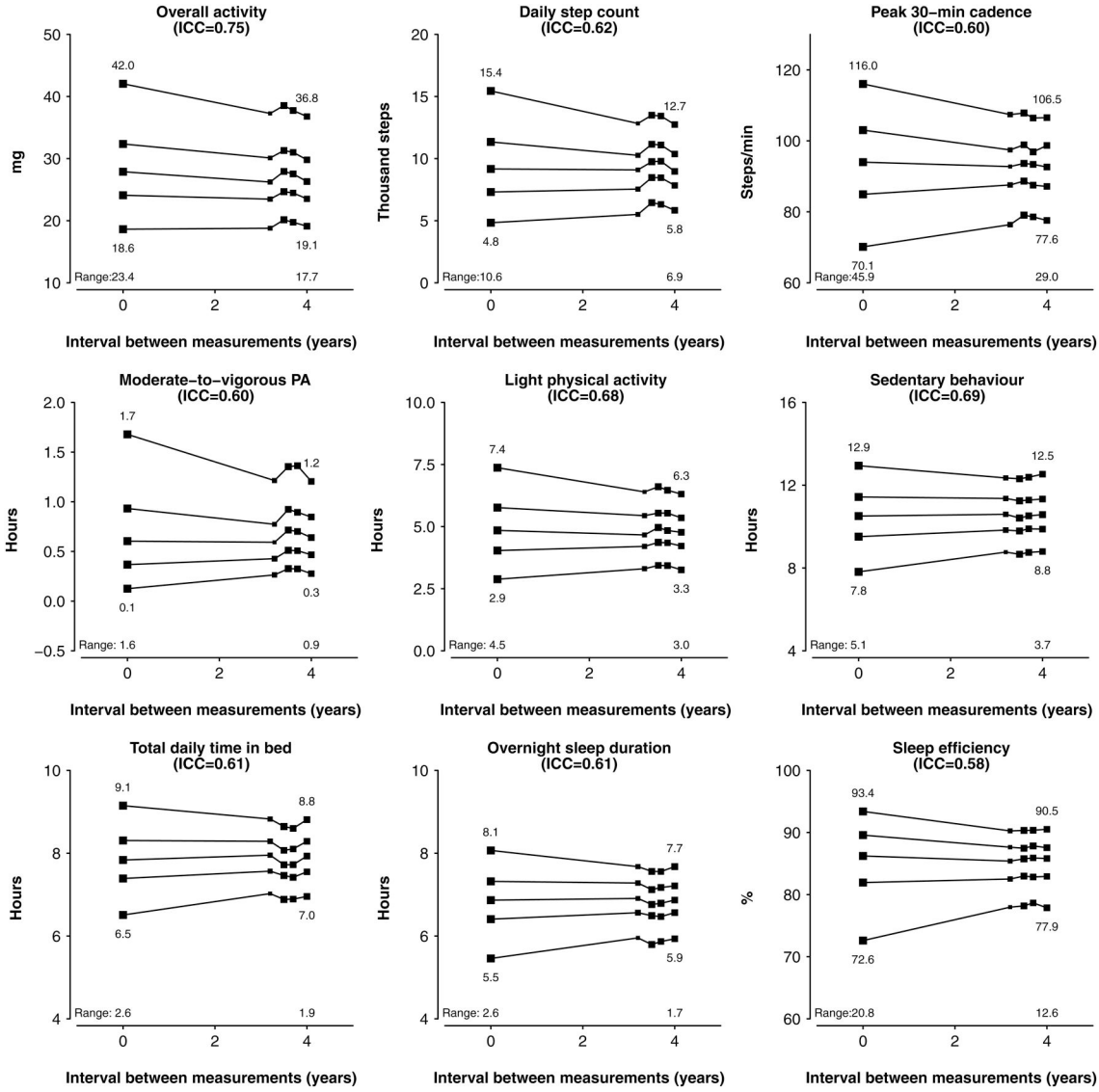

Figure 2 Changes in accelerometer-derived phenotypes of physical activity and sleep within fifths defined by the main accelerometer measurements. Analysis was performed in 3138 participants with valid accelerometer data from the main sub-study and at least one valid repeat measurement. Squares represent mean values initially and at subsequent repeat measurements for participants divided into fifths according to their main accelerometer measurement. The size of each square is proportional to the number of participants available. The mean values in the top and bottom groups and the absolute differences (ranges) between them are given at Years 0 and 4. PA, physical activity.

estimates for different movement behaviours ranging from 45% to 60% [30]. Similar results were obtained in a Chinese population of 20 000 participants over a 3-year period [19]. For self-reported sleep, two UK studies ($n = 13$ 000 over 1.7 years and $n = 19$ 000 over 4.3 years) found kappa scores of ∼50% for sleep duration categories [30, 31]. Notably, the reproducibility of accelerometer-derived phenotypes is comparable to that of well-established CHD risk factors such as systolic blood pressure and total cholesterol, which is 65%–70% within a 5-year period [17, 18]. In broader epidemiological research, accelerometer-derived phenotypes exhibit moderately lower reproducibility than the most-reproducible imaging measures (median = 85%) and substantially higher reproducibility than the least-reproducible 24-hour dietary-recall measures collected through online questionnaires (median = 35%), based on a UK Biobank study of 2858 variables [32].

The reproducibility of accelerometer measurements in our study supports their use in epidemiological research, but measurement error and fluctuations in physical activity and sleep inevitably occur. These deviations from an individual's usual behaviour can attenuate associations with health outcomes when only a single measurement is used, as illustrated by our examples. Very few large-scale studies of physical activity and sleep have employed repeat assessments (those that have are limited to self-reporting). For example, studies using the average physical activity from repeated surveys have reported stronger associations with health outcomes, suggesting prior underestimation of ∼60% [33, 34]. The approach used in our study has

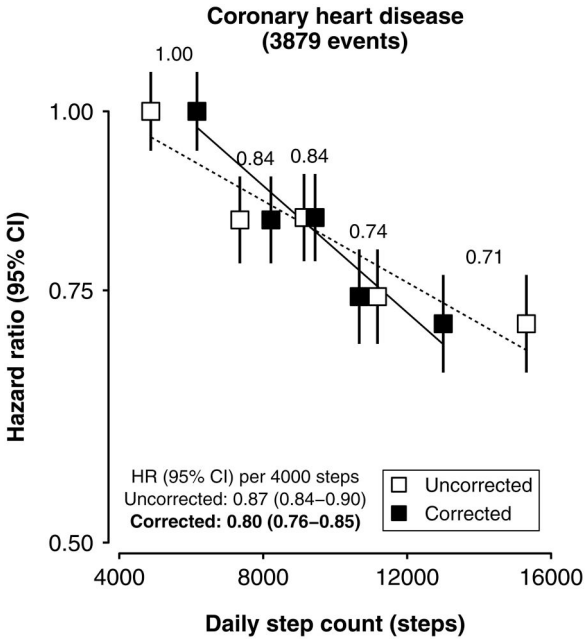

Figure 3 Illustrative example of the association between daily step count and incident CHD, before and after correction for regression dilution bias. Analysis was performed in 87 038 participants after excluding those with low-quality accelerometer data, prevalent atherosclerotic cardiovascular disease, or missing covariates. Cox models were undertaken by using age as the timescale, stratified by sex, and adjusted for season of accelerometer wear, ethnicity, geographical area, education, Townsend deprivation index, alcohol intake, smoking status, and fresh fruit and vegetable intake. For each step count group, the square represents the HR, with its value shown above and its size being inversely proportional to the variance of the group-specific log risk. Vertical lines represent group-specific 95% CIs. Diagonal lines illustrate the trend across groups based on inverse-variance-weighted regression. The HRs per 4000 steps reflect models in which the step count was treated as a continuous variable.

the advantage of requiring repeat measurements from only a subset of the original cohort—an efficient method previously applied in UK and Chinese populations [19, 31]. It is important to note that, while our analysis focuses on the relative change in risk after correction for regression dilution, interpretation of disease associations also requires careful consideration of other sources of bias (particularly residual confounding and reverse causation) when informing public health recommendations.

This study has several strengths, including being the largest reproducibility study of device-measured physical activity and sleep to date, with four repeat measurements, which enabled us to assess reproducibility across key subgroups. Additionally, it focused on wrist-worn accelerometers, which are increasingly used in large-scale cohort studies, and examined a broad range of derived phenotypes, capturing different aspects of movement behaviour. However, the study also has limitations. First, participant characteristics were measured at baseline, a few years before the main accelerometer measurements, and may not accurately reflect participants' characteristics at the time of wear (e.g. body mass index and self-reported disability used in the stratified analyses). Second, UK Biobank is not representative of the UK population and movement behaviours may differ further in other cultural and geographic settings. The repeat accelerometry sub-study also includes a highly compliant sample, which may exhibit more stable routines and contribute to higher ICCs. Therefore, our ICCs may not generalize widely, though they align with previous estimates. Finally, measurement error in confounders can lead to residual confounding, which may bias exposure–outcome associations in either direction. Methods such as regression calibration can be used to correct for this [35], which may further shift the corrected association, depending on the direction and strength of confounding as well as the extent of measurement error in the confounders. However, this approach requires repeated measurements of both exposure and confounders, which were not collected concurrently in our study. While we considered using covariate data from the first repeat assessment visit (2012–13), the overlap with repeat accelerometry participants was too limited.

Future work should assess reproducibility in more representative cohorts, including populations with different movement behaviour patterns such as those in non-Western settings. In addition, the impact of regression dilution bias should be investigated across a wide range of health outcomes and extended to other physical activity and sleep phenotypes, including time-based and intensity-specific metrics, by using appropriate analytical frameworks such as compositional data analysis.

Conclusion

This study provides new insights into the moderate-to-good long-term reproducibility of accelerometer-derived physical activity and sleep among UK adults. Similarly to other measures, such as blood pressure, this level of reproducibility is sufficiently high for epidemiological research, but associations with health outcomes remain subject to regression dilution bias. This highlights the need for cohorts to incorporate repeat accelerometer measurements and for researchers to account for regression

dilution bias in studies of accelerometer-derived physical activity and sleep.

Ethics Approval

Ethics approval for this study is covered by the general ethics approval for UK Biobank studies from the North West Multicentre Research Ethics Committee (Ref 11/NW/0382), which is renewed every 5 years. All study participants provided written informed consent.

Acknowledgements

We thank Simon Sheard from the UK Biobank team for coordinating the repeat accelerometry sub-study, and Howard Callen and Janet Maccora for providing key information and guidance on its study design. This research was conducted by using the UK Biobank resource under application number 59070. We are grateful to all participants for generously contributing their data to advance research in population health. This work used data provided by patients and collected by the National Health Service (NHS) as part of their care and support. Computation was performed via the Oxford Biomedical Research Computing (BMRC) facility—a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the National Institute for Health and Care Research (NIHR) Oxford Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

Author contributions

C.Z., D.B., S.L., J.C.H., and A.D. conceived the project and developed an analysis plan. N.W. and S.B. designed the repeat accelerometry sub-study. C.Z. performed all analyses and wrote the first draft of the manuscript. J.C.H. and A.D. supervised the work. All authors discussed the results and contributed to the final manuscript.

Supplementary material

Supplementary material is available at *IJE* online.

Conflicts of interest

The Nuffield Department of Population Health receives research grants from industry that are governed by University of Oxford contracts that protect its independence and has a staff policy of not taking personal payments from the pharmaceutical and food industries; further details can be found at https://www.ndph.ox.ac.uk/about/independence-of-research.

Funding

C.Z. is supported by the Oxford British Heart Foundation (BHF) Centre of Research Excellence (RE/18/3/34214). C.C., H.T., and B. L. acknowledge support from UK Biobank, funded largely by the UK Medical Research Council (MRC) and Wellcome. I.H. is supported by grants to the University of Oxford from the UK MRC, the BHF, and Health Data Research (HDR) UK. R.W. is supported by a MRC Industrial Strategy Studentship (MR/S502509/1) and by HDR UK—an initiative funded by UK Research and Innovation (UKRI), Department of Health and Social Care (England) and the devolved administrations. T.S., K.W., N.W., and S.B. are supported by the UK MRC (MC_UU_00006/1, MC_UU_00006/4) and NIHR Cambridge Biomedical Research Centre (NIHR203312). K. S.-B. is supported by Cancer Research UK (C8221/A29017, C16077/A29186) and UKRI (10063259). D.B. is supported by Novo Nordisk and Swiss Re. S.L. reports grants from HDR UK Ltd (HDRUK2023.0028) funded by the MRC, Engineering and Physical Sciences Research Council (EPSRC), Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), BHF, and Cancer Research UK, outside the submitted work. J.C.H. acknowledges funding from the BHF, NIHR Oxford Biomedical Research Centre (BRC), BHF Centre of Research Excellence, and the Nuffield Department of Population Health. A.D.'s research team is supported by a range of grants from the Wellcome Trust (223100/Z/21/Z, 227093/Z/23/Z), Novo Nordisk, Swiss Re, Boehringer Ingelheim, National Institutes of Health's Oxford Cambridge Scholars Program, EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1), and the BHF Centre of Research Excellence (RE/18/3/34214). For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Data availability

Researchers can request data access by following UK Biobank's data-access guidelines, available at https://www.ukbiobank.ac.uk/enable-your-research. Code to reproduce the results and apply the methods used in this manuscript is available at https://github.com/OxWearables/ukb_repeat_accelerometry.

Use of Artificial Intelligence (AI) Tools

ChatGPT was used to proofread the manuscript for grammar and clarity. All scientific aspects of the study, including the design, data analysis, interpretation of results, and image generation, were conducted by the authors without the use of AI.

References

1. Bull FC, Al-Ansari SS, Biddle S *et al.* World Health Organization 2020 guidelines on physical activity and

sedentary behaviour. *Br J Sports Med* 2020;**54**:1451–62. https://doi.org/10.1136/bjsports-2020-102955
2. Jike M, Itani O, Watanabe N, Buysse DJ, Kaneita Y. Long sleep duration and health outcomes: A systematic review, meta-analysis and meta-regression. *Sleep Med Rev* 2018;**39**: 25–36. https://doi.org/10.1016/j.smrv.2017.06.011
3. Chen Y, Chan S, Bennett D, *et al.*; China Kadoorie Biobank Collaborative Group. Device-measured movement behaviours in over 20,000 China Kadoorie Biobank participants. *Int J Behav Nutr Phys Act* 2023;**20**:138. https://doi.org/10.1186/s12966-023-01537-8
4. Doherty A, Jackson D, Hammerla N *et al.* Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study. *PLoS One* 2017;**12**:e0169649. https://doi.org/10.1371/journal.pone.0169649
5. Weber A, van Hees VT, Stein MJ *et al.* Large-scale assessment of physical activity in a population using high-resolution hip-worn accelerometry: the German National Cohort (NAKO). *Sci Rep* 2024;**14**:7927. https://doi.org/10.1038/s41598-024-58461-5
6. Wasfy MM, Lee IM. Examining the dose–response relationship between physical activity and health outcomes. *NEJM Evid* 2022;**1**:EVIDra2200190. https://doi.org/10.1056/EVIDra2200190
7. Yuan H, Plekhanova T, Walmsley R *et al.* Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality. *NPJ Digit Med* 2024; **7**:86. https://doi.org/10.1038/s41746-024-01065-0
8. Saint-Maurice PF, Freeman JR, Russ D *et al.* Associations between actigraphy-measured sleep duration, continuity, and timing with mortality in the UK Biobank. *Sleep* 2024;**47**: zsad312. https://doi.org/10.1093/sleep/zsad312
9. Pulsford RM, Brocklebank L, Fenton SAM *et al.* The impact of selected methodological factors on data collection outcomes in observational studies of device-measured physical behaviour in adults: a systematic review. *Int J Behav Nutr Phys Act* 2023;**20**:26. https://doi.org/10.1186/s12966-022-01388-9
10. Keadle SK, Shiroma EJ, Kamada M, Matthews CE, Harris TB, Lee IM. Reproducibility of accelerometer-assessed physical activity and sedentary time. *Am J Prev Med* 2017;**52**:541–8. https://doi.org/10.1016/j.amepre.2016.11.010
11. Jaeschke L, Steinbrecher A, Jeran S, Konigorski S, Pischon T. Variability and reliability study of overall physical activity and activity intensity levels using 24 h-accelerometry-assessed data. *BMC Public Health* 2018;**18**:530. https://doi.org/10.1186/s12889-018-5415-8
12. Saint-Maurice PF, Sampson JN, Keadle SK, Willis EA, Troiano RP, Matthews CE. Reproducibility of accelerometer and posture-derived measures of physical activity. *Med Sci Sports Exerc* 2020;**52**:876–83. https://doi.org/10.1249/mss.0000000000002206
13. Al-Shaar L, Pernar CH, Chomistek AK *et al.* Reproducibility, validity, and relative validity of self-report methods for assessing physical activity in epidemiologic studies: findings from the women's lifestyle validation study. *Am J Epidemiol* 2022;**191**:696–710. https://doi.org/10.1093/aje/kwab294
14. Pernar CH, Chomistek AK, Barnett JB *et al.* Validity and relative validity of alternative methods of assessing physical activity in epidemiologic studies: findings from the men's lifestyle validation study. *Am J Epidemiol* 2022;**191**:1307–22. https://doi.org/10.1093/aje/kwac051
15. Clarke R, Shipley M, Lewington S *et al.* Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. *Am J Epidemiol* 1999;**150**: 341–53. https://doi.org/10.1093/oxfordjournals.aje.a010013
16. Frost C, Thompson SG. Correcting for regression dilution bias: comparison of methods for a single predictor variable. *J R Stat Soc A: Stat Soc* 2000;**163**:173–89. https://doi.org/10.1111/1467-985x.00164
17. Prospective Studies C. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* 2002;**360**:1903–13. https://doi.org/10.1016/S0140-6736(02)11911-8
18. Prospective Studies C. Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55 000 vascular deaths. *Lancet* 2007;**370**:1829–39. https://doi.org/10.1016/S0140-6736(07)61778-4
19. Bennett DA, Du H, Clarke R, *et al.*; China Kadoorie Biobank Study Collaborative Group. Association of physical activity with risk of major cardiovascular diseases in Chinese men and women. *JAMA Cardiol* 2017;**2**:1349–58. https://doi.org/10.1001/jamacardio.2017.4069
20. Sudlow C, Gallacher J, Allen N *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**: e1001779. https://doi.org/10.1371/journal.pmed.1001779
21. UK Biobank. *UK biobank: protocol for a large-scale prospective epidemiological resource*. https://www.ukbiobank.ac.uk/wp-content/uploads/2025/01/Main-study-protocol.pdf (21 January 2026, date last accessed).
22. Small SR, Chan S, Walmsley R *et al.* Self-supervised machine learning to characterise step counts from wrist-worn accelerometers in the UK biobank. *Med Sci Sports Exerc* 2024;**56**: 1945–53. https://doi.org/10.1249/Mss.0000000000003478
23. Walmsley R, Chan S, Smith-Byrne K *et al.* Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease. *Br J Sports Med* 2021; **56**:1008–17. https://doi.org/10.1136/bjsports-2021-104050
24. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;**15**:155–63. https://doi.org/10.1016/j.jcm.2016.02.012
25. Paluch AE, Bajpai S, Ballin M, *et al.*; Steps for Health Collaborative. Prospective association of daily steps with cardiovascular disease: a harmonized meta-analysis. *Circulation* 2023;**147**:122–31. https://doi.org/10.1161/CIRCULATIONAHA.122.061288
26. Stens NA, Bakker EA, Mañas A *et al.* Relationship of daily step counts to all-cause mortality and cardiovascular events. *J Am Coll Cardiol* 2023;**82**:1483–94. https://doi.org/10.1016/j.jacc.2023.07.029
27. Fry A, Littlejohns TJ, Sudlow C *et al.* Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *Am J*

Epidemiol 2017;**186**:1026–34. https://doi.org/10.1093/aje/kwx246
28. Dwyer T, Pezic A, Sun C *et al.* Objectively measured daily steps and subsequent long term all-cause mortality: the Tasped prospective cohort study. *PLoS ONE* 2015;**10**: e0141274. https://doi.org/10.1371/journal.pone.0141274
29. Knutson KL, Rathouz PJ, Yan LL, Liu K, Lauderdale DS. Intra-individual daily and yearly variability in Actigraphically recorded sleep measures: the CARDIA study. *Sleep* 2007;**30**: 793–6. https://doi.org/10.1093/sleep/30.6.793
30. Pearce M, Strain T, Kim Y *et al.* Estimating physical activity from self-reported behaviours in large-scale population studies using network harmonisation: findings from UK biobank and associations with disease outcomes. *Int J Behav Nutr Phys Act* 2020;**17**:40. https://doi.org/10.1186/s12966-020-00937-4
31. Wong ATY, Heath AK, Tong TYN *et al.* Sleep duration and breast cancer incidence: results from the million women study and meta-analysis of published prospective studies. *Sleep* 2021;**44**:zsaa166. https://doi.org/10.1093/sleep/zsaa166
32. Rutter CE, Millard LAC, Borges MC, Lawlor DA. Exploring regression dilution bias using repeat measurements of 2858 variables in ≤49 000 UK Biobank participants. *Int J Epidemiol* 2023;**52**:1545–56. https://doi.org/10.1093/ije/dyad082
33. Martinez-Gomez D, Cabanas-Sanchez V, Yu T *et al.* Long-term leisure-time physical activity and risk of all-cause and cardiovascular mortality: dose–response associations in a prospective cohort study of 210 327 Taiwanese adults. *Br J Sports Med* 2022;**56**:919–26. https://doi.org/10.1136/bjsports-2021-104961
34. Lee DH, Rezende LFM, Ferrari G *et al.* Physical activity and all-cause and cause-specific mortality: assessing the impact of reverse causation and measurement error in two large prospective cohorts. *Eur J Epidemiol* 2021;**36**:275–85. https://doi.org/10.1007/s10654-020-00707-3
35. Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am J Clin Nutr* 1997;**65**:1179S–86S. https://doi.org/10.1093/ajcn/65.4.1179S

International Journal of Epidemiology, 2026, 55, 1–11
https://doi.org/10.1093/ije/dyag014
Original Article