Comparing two wrist-worn accelerometers (Axivity AX3 and Matrix 003) for measuring movement behaviours in British and Chinese older adults

Abstract

Introduction: Two nationally representative cohorts, the English Longitudinal Study of Ageing (ELSA) and the China Health and Retirement Longitudinal Study (CHARLS), recently introduced wrist-worn accelerometry to measure movement behaviours. However, the use of different brands (Axivity AX3 and Matrix 003) may hinder data harmonisation. This study assessed whether the raw acceleration data and machine learning-derived physical activity and sleep outcomes were equivalent between these two accelerometers in both British and Chinese adults.

Methods: Eighty-five British and 117 Chinese adults aged ≥50 years wore both accelerometers in a random positional order on their dominant wrist for up to eight days. Data were processed using open-source machine learning algorithms, developed in the UK, to generate outcomes such as average acceleration (mg), time in 24-hour movement behaviours (hours/day), daily step count, and sleep duration (hours/night). Equivalency was assessed using 95% equivalence tests (±10% equivalence zone).

Results: In both British and Chinese adults, average acceleration, sedentary time, time in bed, and sleep duration were equivalent between the two accelerometers, while time in moderate-vigorous physical activity (MVPA) was not (within ±17.7% in British and ±28.2% in Chinese adults). Time in light physical activity (LPA) was equivalent in British (±6.2%) but borderline in Chinese adults (±10.3%), whereas the opposite was observed for daily step count (±10.5% in British and ±2.9% in Chinese adults).

Conclusion: Average acceleration – a widely used measure of overall physical activity – was comparable between the Axivity AX3 and the Matrix 003 in both British and Chinese adults.

Machine learning-derived physical activity and sleep outcomes were also largely comparable; however, the cross-nationality differences observed highlight the need for further population-specific algorithm development and validation.

Word count: 250 words

Keywords: raw acceleration; machine learning; physical activity; sleep; data harmonisation; cross-nationality comparison.

Highlights

1) The harmonisation of accelerometry data across large-scale epidemiological cohorts is currently being hindered by different brands and processing methods.
2) Average acceleration – a widely used measure of overall physical activity – was comparable between the Axivity AX3 and the Matrix 003 in both British and Chinese adults aged 50 years or older; this suggests that data from these two different brands of wrist-worn accelerometer can be pooled and/or compared.
3) Machine learning-derived physical activity and sleep outcomes were also largely comparable; however, the cross-nationality differences observed highlight the need for further population-specific algorithm development and validation.

Introduction

Less time spent being physically active, more time spent being sedentary, and suboptimal sleep duration (outside 6-8 hours per night) are associated with premature mortality and a higher risk of many chronic diseases (Bull et al., 2020; Chaput et al., 2020). However, the majority of evidence to date is based on self-reported measures of these movement behaviours. Although self-report has been sufficient to demonstrate the importance of movement behaviours for health, and can provide valuable contextual information, it lacks the precision to provide specific quantitative recommendations about the optimal amounts and/or composition of movement behaviours for health. For example, studies using accelerometry to examine the association between physical activity and premature mortality have reported effect sizes almost double the size of studies using self-report (Ekelund et al., 2019; Wasfy & Lee, 2022).

An increasing number of large-scale epidemiological cohorts are implementing wrist-worn accelerometry to measure movement behaviours. However, differences in accelerometer brand and processing method create challenges for data harmonisation. The UK Biobank (Doherty et al., 2017), the China Kadoorie Biobank (Chen et al., 2023), and more recently, the English Longitudinal Study of Ageing (ELSA; 2021-23) (Steptoe et al., 2013) all used the Axivity AX3 wrist-worn accelerometer. In contrast, the US National Health and Nutrition Examination Survey (NHANES; 2011-14) used the ActiGraph GT3X+, the British Whitehall II Study (Menai et al., 2017) used the GENEActiv, and the China Health and Retirement Longitudinal Study (CHARLS; 2021-23) (Zhao et al., 2014) developed their own wrist-worn device, the Matrix 003, in collaboration with a Chinese manufacturer. Using open-source software (van Hees et al., 2013; 2014), it is possible to process the raw acceleration data from different accelerometer brands and directly compare them (Rowlands et al., 2019). In addition, there is now the availability of open-source machine learning algorithms to estimate the following: 1) time spent in 24-hour movement behaviours – i.e., moderate-vigorous physical activity (MVPA), light physical activity (LPA), sedentary behaviour (SB), and time in bed (Walmsley et al., 2021); 2) sleep duration and efficiency (Yuan et al., 2024); and 3) step count and cadence (Small et al., 2024). However, it remains to be seen whether these machine learning algorithms can be applied to different accelerometer brands in different populations.

This study aimed to test the first two components of the V3 framework for evaluating Biometric Monitoring Technologies (BioMeTs) such as accelerometers (Goldsack et al., 2020). The primary aim (verification) was to establish whether the raw acceleration data – i.e., the volume and intensity distribution of physical activity – from two different brands of accelerometer (the well-established Axivity AX3 and the newly developed Matrix 003) worn on the dominant wrist can be considered equivalent in both British and Chinese older adults (aged 50 years or older). The secondary aim (analytical validation) was to establish whether the machine learning-derived physical activity and sleep outcomes can also be considered equivalent in both British and Chinese older adults.

Methods

Free-living validation study in British adults

A convenience sample of 85 ambulant adults aged 50 years or older living in the UK was recruited by email and word of mouth. All participants provided written informed consent, and the study was approved by the University College London (UCL) Research Ethics Committee. Data were collected between October 2021 and November 2022.

Participants self-reported their height and weight, to calculate body mass index (BMI) in kg/m^2, and any long-standing (≥3 months) mobility limitations (Supplementary Table S1). Participants were asked to start wearing both accelerometers immediately after receiving them in the post and wear them on their dominant wrist 24 hours per day for eight consecutive days. The positional order of the two accelerometers on the wrist was randomised between participants (Supplementary Figure S1). After eight days, the UCL Courier Service collected the accelerometers. Participants were informed that they could wear the waterproof accelerometers when bathing or swimming but not in extremely high temperature or pressure environments (e.g., in a sauna or when diving). Participants were asked to carry on with their normal activities whilst wearing the accelerometers and did not receive feedback on their activity levels until after they were returned.

Free-living validation study in Chinese adults

A convenience sample of 117 ambulant adults aged 50 years or older living in China was recruited by visiting two urban and two rural communities in or near Beijing. All participants provided written informed consent, and the study was approved by the Peking University Ethics Review Committee. Data were collected between May and November 2023.

Participants self-reported their height and weight, to calculate BMI in kg/m^2, and any long-standing (≥3 months) mobility limitations (Supplementary Table S1). The researchers placed both accelerometers on the participants' dominant wrist in a randomised positional order (Supplementary Figure S1). Participants were asked to wear the accelerometers 24 hours per day for two nights and one day. After the second night, the researchers collected the accelerometers. Participants were informed that they could wear the waterproof accelerometers when bathing or swimming but not in extremely high temperature or pressure environments (e.g., in a sauna or when diving). Participants were asked to carry on with their normal activities whilst wearing the accelerometers and did not receive feedback on their activity levels until after they were collected.

The Axivity AX3

The Axivity (Axivity Ltd, Newcastle, UK) is a wrist-worn triaxial accelerometer that has been used to measure movement behaviours in large-scale epidemiological cohorts such as the UK Biobank (Doherty et al., 2017), the China Kadoorie Biobank (Chen et al., 2023), and the tenth wave of ELSA (Steptoe et al., 2013). In British adults, the Axivity was set to start recording at 10am two working days after postal dispatch and stop recording seven full days later, whereas in Chinese adults it was set to start recording immediately before placement by one of the researchers and stop recording immediately after collection. In both British and Chinese adults, the Axivity was set to capture triaxial acceleration data at 100 Hz with a dynamic range of ±8 *g*.

The Matrix 003

The Matrix (Beijing XMatrix Tech. Co., Ltd, Beijing, China) was developed in 2021 to measure movement behaviours in the fifth wave of CHARLS (Zhao et al., 2014). It is a wrist-worn triaxial accelerometer similar to the Axivity, but it also has a gyroscope and a heart rate monitor. It is not possible to configure Matrix accelerometers to start and stop recording at specific times; its only options are to start immediately, after 24 hours, or after 48 hours. Therefore, in British adults, the Matrix was set to start recording just before 10am two working days after postal dispatch, whereas in Chinese adults it was set to start recording immediately before placement by one of the researchers. In both British and Chinese adults, the Matrix was set to capture triaxial acceleration data at 50 Hz with a dynamic range of ±8 *g*.

The Matrix was set to have a lower sampling rate than the Axivity (50 Hz versus 100 Hz) to ensure it had at least seven days of battery life whilst also capturing triaxial gyroscope data at 50 Hz and heart rate data every 15 minutes. However, during data processing, both the Axivity and Matrix datasets were resampled using nearest neighbour interpolation at a rate of 50 Hz. Nearest neighbour interpolation resampling, recommended by Small et al. (2021), was used to avoid unintended smoothing of slower sampled data – specifically, the Matrix data – which has been shown to result in lower overall physical activity.

Data processing

Each participant's Matrix dataset was clipped to match their Axivity start and end times. Periods of Axivity non-wear time were then removed from the Matrix time series data and vice versa. Finally, gravity was calibrated in both accelerometers to ensure that at rest, the average magnitude of acceleration was 1 *g* (9.81 m/s^2).

Participants were excluded if, for at least one of the accelerometers, the data could not be parsed, the device could not be calibrated, more than 1% of readings were 'clipped' (fell outside ±8 *g* for *biobankAccelerometerAnalysis*, ±3 *g* for *asleep,* and ±2 *g* for *stepcount*) before or after calibration, or the average acceleration was implausibly high (>100 m*g* for *biobankAccelerometerAnalysis* and *stepcount*, and >200 m*g* for *asleep*).

Primary outcomes (verification)

Volume and intensity distribution of physical activity

The volume and intensity distribution of physical activity were derived from the Biobank Accelerometer Analysis Tool (github.com/OxWearables/biobankAccelerometerAnalysis, v7.1.1), which was developed and validated by the Oxford Wearables Group (Walmsley et al., 2021). Participants were excluded if they did not have sufficient wear time (≥3 days in British and ≥1 day in Chinese adults, and data in each one-hour period of the 24-hour cycle), with non-wear time defined as unbroken episodes of at least 60 minutes during which the standard deviation (SD) of each axis of acceleration was less than 13 m*g*. To account for potential wear time diurnal bias, recording interruptions and non-wear time were imputed using the average values from the corresponding minute of the day on the remaining days of worn data.

The *biobankAccelerometerAnalysis* algorithm produced the following primary outcomes for each accelerometer separately: total wear time (days), average acceleration (m*g* per day), and time (hours per day) accumulated above incremental acceleration thresholds (>25-200 m*g* in 25-m*g* increments). Average acceleration refers to the Euclidean Norm Minus One (ENMO), a widely used measure of overall physical activity due to its correlation with physical activity energy expenditure (PAEE) (van Hees et al., 2013).

Secondary outcomes (analytical validation)

Time spent in 24-hour movement behaviours

The *biobankAccelerometerAnalysis* algorithm also produced the following secondary outcomes for each accelerometer separately: time (hours per day) spent in MVPA, in LPA, being sedentary, and in bed.

Sleep duration and efficiency

Sleep duration and efficiency were derived from a sleep staging algorithm (github.com/OxWearables/asleep, v0.4.13), which was also developed and validated by the Oxford Wearables Group (Yuan et al., 2024). The *asleep* algorithm classifies each 30-second epoch of acceleration data into one of the three sleep stages: 1) wake; 2) rapid eye movement sleep (REM); and 3) non-rapid eye movement sleep (NREM). Participants were excluded if they did not have sufficient wear time (≥22 hours per day for ≥3 days [including ≥1 weekend day] in British and ≥1 day in Chinese adults), with non-wear time defined as unbroken episodes of at least 90 minutes during which the SD of each axis of acceleration was less than 13 m*g*.

For each accelerometer separately, the *asleep* algorithm calculated the following sleep parameters for the longest sleep window over a noon-to-noon interval, with up to 60 minutes of sleep discontinuity allowed: overnight sleep duration (hours per night) and sleep efficiency (percentage of time in bed spent asleep).

Step count and cadence

Step count and cadence were derived from a hybrid machine learning and peak detection step counting algorithm (github.com/OxWearables/stepcount, v3.8.0), which was also developed and validated by the Oxford Wearables Group (Small et al., 2024). Participants were excluded if they did not have sufficient wear time (≥3 days in British and ≥1 day in Chinese adults, and data in each one-hour period of the 24-hour cycle), with non-wear time defined as unbroken episodes of at least 90 minutes during which the SD of each axis of acceleration was less than 13 m*g*. To account for potential wear time diurnal bias, recording interruptions and non-wear time were imputed using the average values from the corresponding minute of the day on the remaining days of worn data.

The *stepcount* algorithm produced the following secondary outcomes for each accelerometer separately: overall daily step count (steps per day) and peak cadence (steps per minute). Overall daily step count was reported as the median number of steps taken per day across the monitoring period. One-minute peak cadence was calculated as previously described by Saint-Maurice et al. (2020).

Statistical analyses

Descriptive statistics (mean [SD] where data were normally distributed, otherwise median [25th-75th percentile]) were calculated for all outcomes, with differences between British and Chinese adults being examined using the independent *t*-test or the Chi-squared test, and differences between the two accelerometer brands being examined using the paired *t*-test or the Wilcoxon signed-rank test.

We used 95% equivalence tests with a 10% equivalence zone to determine whether the 95% confidence interval (95% CI) for the mean of one accelerometer fell within ±10% of the mean of the other accelerometer (Wellek, 2003). This was based on a previous study comparing three different accelerometers (Axivity AX3, ActiGraph GT9X, and GENEActiv) worn on both wrists (Rowlands et al., 2019), as well as many other validity studies of physical behaviour measures (O'Brien, 2021). Where data were not normally distributed, the log transformation of the original data were used for the equivalency analyses. As neither accelerometer can be considered the gold standard, equivalency analyses were carried out twice – i.e., with each accelerometer as the reference monitor. In all cases, equivalency was consistent regardless of the reference monitor. Therefore, results are presented with the Axivity as the reference monitor because it is more established.

The level of agreement between outputs from the two accelerometers was determined using intraclass correlation coefficients (ICCs; two-way mixed effects, absolute agreement, single measures) with 95% CI, and mean bias with 95% limits of agreement (LoA) (Bland & Altman, 1986). The ICC was classified as 'poor', 'moderate', 'good', or 'excellent' reliability if the lower band of the 95% CI of the ICC was <0.5, 0.5-0.75, >0.75-0.9, or >0.9, respectively (Koo & Li, 2016).

Separate analyses were conducted for British and Chinese adults, and all analyses were conducted in RStudio (R version 4.4.1). Statistical significance was defined as $p < 0.05$.

Results

Of the 85 participants who were recruited in the UK, 82 had valid data for all outcomes from both accelerometers (52 [63.4%] female, mean [SD] age: 65.6 [10.2] years, 54 [65.9%] living in an urban area, mean [SD] BMI: 24.4 [3.6] kg/m^2, 21 [25.6%] with mobility limitations). Of the 117 participants who were recruited in China, 106 had valid data for all outcomes from both accelerometers. The Chinese participants had similar characteristics to the British participants, except they were more likely to have mobility limitations (64 [61.0%] female, mean [SD] age: 66.7 [10.3] years, 58 [55.2%] living in an urban area, mean [SD] BMI: 24.4 [3.2] kg/m^2, 49 [46.7%] with mobility limitations). Descriptive statistics for all outcomes by nationality (British or Chinese) and accelerometer brand (Axivity or Matrix) are presented in Table 1.

In both British and Chinese adults, time spent in bed and sleeping were higher for the Matrix than the Axivity, whilst sleep efficiency and sedentary time were lower. In British adults, overall daily step count was lower for the Matrix than the Axivity, whereas the opposite was observed in Chinese adults. In British adults only, total wear time was higher for the Matrix than the Axivity, whilst overall physical activity (i.e., average acceleration), time accumulated above the lowest threshold of acceleration (25 m*g*), and time spent in LPA were lower. In Chinese adults only, time accumulated above higher thresholds of acceleration (>100 m*g*) were higher for the Matrix than the Axivity.

Verification results

Equivalency results for the volume and intensity distribution of physical activity by nationality are shown in Figure 1 and Supplementary Table S2. In Figure 1, markers are denoted in solid squares if the 95% CI for the mean of the Matrix 003 fell within ±10% of the mean of the Axivity AX3, otherwise markers are denoted in hollow squares. In both British and Chinese adults, total wear time, average acceleration, and time accumulated above the two lowest thresholds of acceleration (25 and 50 m*g*) were equivalent between the two accelerometer brands (within ±10% equivalence zone). Time accumulated above 75 m*g* was borderline equivalent in British (±10.33%) but not equivalent in Chinese adults (±14.33%), and time accumulated above the five highest thresholds of acceleration (100-200 mg) were not equivalent in either British or Chinese adults.

Agreement results for the volume and intensity distribution of physical activity by nationality are shown in Figure 2 and Supplementary Table S3. Bland-Altman plots are also shown in Supplementary Figure S2. In Figure 2, markers are denoted in solid squares if the lower band of the 95% CI of the ICC was good (>0.75) or excellent (>0.9), otherwise markers are denoted in hollow squares. In both British and Chinese adults, reliability between the two accelerometer brands was excellent for total wear time, average acceleration, and time accumulated above all thresholds of acceleration (25-200 m*g*).

Analytical validation results

Time spent in 24-hour movement behaviours

Equivalency results for time spent in 24-hour movement behaviours by nationality are shown in Figure 1 and Supplementary Table S2. In both British and Chinese adults, time spent in bed and being sedentary were equivalent between the two accelerometer brands (within ±10% equivalence zone); whereas time spent in MVPA was not equivalent, particularly in Chinese adults (±17.74% in British and ±28.20% in Chinese adults). Time spent in LPA was equivalent between the two accelerometer brands in British (±6.22%) but only borderline equivalent in Chinese adults (±10.28%).

Agreement results for time spent in 24-hour movement behaviours are shown in Figure 2 and Supplementary Table S3. Bland-Altman plots are also shown in Supplementary Figure S2. In both British and Chinese adults, reliability between the two accelerometer brands was excellent for time spent in both LPA and MVPA but only moderate for time spent in bed. Reliability between the two accelerometer brands was excellent for sedentary time in British but only good in Chinese adults.

Sleep duration and efficiency

Equivalency results for sleep duration and efficiency are shown in Figure 1 and Supplementary Table S2. In both British and Chinese adults, both overnight sleep duration and sleep efficiency were equivalent between the two accelerometer brands (within ±10% equivalence zone).

Agreement results for sleep duration and efficiency are shown in Figure 2 and Supplementary Table S3. Bland-Altman plots are also shown in Supplementary Figure S2. In both British and Chinese adults, reliability between the two accelerometer brands was good for sleep efficiency.

Reliability between the two accelerometer brands was good for overnight sleep duration in British but poor in Chinese adults.

Step count and cadence

Equivalency results for step count and cadence are shown in Figure 1 and Supplementary Table S2. In both British and Chinese adults, peak cadence was equivalent between the two accelerometer brands (within ±10% equivalence zone). Overall daily step count was also equivalent between the two accelerometer brands in Chinese (±2.93%) but only borderline equivalent in British adults (±10.50%).

Agreement results for step count and cadence are shown in Figure 2 and Supplementary Table S3. Bland-Altman plots are also shown in Supplementary Figure S2. In both British and Chinese adults, reliability between the two accelerometer brands was excellent for both overall daily step count and peak cadence.

Discussion

In both British and Chinese adults, average acceleration was equivalent between the Axivity AX3 and the Matrix 003. The vast majority of machine learning-derived physical activity and sleep outcomes were also equivalent or borderline equivalent. The only exceptions were time spent above higher thresholds of acceleration (>75 m*g*), including time spent in MVPA. However, there were a few notable cross-nationality differences. In Chinese compared to British adults, equivalency between the two accelerometer brands was lower for time spent in both LPA and MVPA but higher for overall daily step count. Furthermore, the ICCs were good or excellent

(>0.75) for all outcomes, except for time spent in bed in both British and Chinese adults and overnight sleep duration in Chinese adults only.

We found that average acceleration was equivalent between the Axivity AX3 and the Matrix 003 when measured at the dominant wrist. In contrast, in 56 young British adults (mean age 24.5 years), Rowlands et al. (2019) found that average acceleration measured at the dominant wrist was approximately 10% higher for the Axivity AX3 and the GENEActiv than for the ActiGraph GT9X. Our results suggest that accelerometry data from the UK Biobank, China Kadoorie Biobank, ELSA, and CHARLS can be pooled and/or compared because they all used the Axivity AX3 or the Matrix 003 on the dominant wrist. We also found that time spent above lower thresholds of acceleration (≤75 m*g*) were equivalent between the Axivity AX3 and the Matrix 003, but time spent above higher thresholds (>75 m*g*) were not. In contrast, Rowlands et al. (2019) found that the intensity gradient – a single metric describing the distribution of acceleration intensity across the 24-hour day (Rowlands et al., 2018) – was equivalent irrespective of accelerometer brand or wrist. The lack of equivalency in our study was most likely due to the 10% equivalence zone being too strict for less common activities that tend to have low magnitudes and/or high variability (Rowlands et al., 2019). This issue may have been exacerbated by the particularly low levels of higher-intensity physical activity typically observed in older adults.

Our findings, together with those of Rowlands et al. (2019), have implications for harmonising data across large-scale epidemiological cohorts that have used wrist-worn accelerometers. Historically, however, many studies have relied on waist-worn devices. In a systematic review of observational studies measuring device-based physical behaviours in adults, Pulsford et al. (2023) reported that the waist was the most common wear location, used in 53% of study waves compared to 20% for the wrist. Nonetheless, wear time compliance was higher for wrist-worn devices than for waist-worn ones, supporting the increasing preference for wrist placement in more recent studies. Furthermore, some studies used thigh-worn accelerometers (5% of study waves), which enable more accurate determination of posture and stepping, while others used multiple wear locations (13% of study waves) to measure different dimensions of physical behaviour, such as intensity, posture, and biological state (e.g., asleep or awake).

Substantial differences in accelerometer outputs have been observed between wear locations under free-living conditions. For example, a meta-analysis by Gall, Sun, and Smuck (2022) reported that wrist-worn accelerometers recorded, on average, 3,537 more steps per day than waist-worn ones. Wrist-worn devices also captured more time in MVPA and less sedentary time compared to waist-worn ones. Similarly, Maylor et al. (2023) found that the wrist-worn Axivity AX3 recorded higher average acceleration and a wider range of acceleration values than the thigh-worn activPAL micro.

Therefore, we are currently conducting two new free-living validation studies – in adults aged 18-30 years and those aged 40 years or older – to establish whether the same machine learning algorithms can be applied to different accelerometer wear locations, or if adjustment factors or location-specific algorithms are needed for data harmonisation.

To our knowledge, this is the first study to show that the same machine learning algorithms can be applied to different brands of accelerometer in different populations. However, the cross-nationality differences observed support the findings of two previous free-living validation studies in British and Chinese adults: CAPTURE-24 (Walmsley et al., 2021) and CAPTURE-24CN (Chen et al., 2023), respectively. In CAPTURE-24, 82% of all MVPA instances involved walking or cycling because almost all of the participants were office workers. In contrast, in

CAPTURE-24CN, only 42% of MVPA instances involved walking or cycling, while the rest consisted of farm or construction work (Chen et al., 2023). Therefore, our finding that equivalency between the two accelerometer brands was lower for time spent in both LPA and MVPA but higher for overall daily step count in Chinese compared to British adults may be due to the UK-derived machine learning algorithms being better at identifying walking and cycling than farm or construction work.

Reprocessing the raw acceleration data from Chinese adults using a China-derived machine learning algorithm (Chen et al., 2023) improved equivalency between the two accelerometer brands for time spent in MVPA, bringing it closer to the level of equivalency observed in British adults (Supplementary Table S4). However, it still fell outside the predefined 10% equivalence zone, and equivalency for time spent in LPA was not improved by the China-derived algorithm. These findings highlight the need for further algorithm retraining using the additional wearable camera data collected in the current study.

It was somewhat surprising that equivalency between the two accelerometer brands for overall daily step count was higher in Chinese than in British adults, given that the *stepcount* machine learning algorithm was trained on data from British adults. Koffman and Muschelli (2024) applied five open-source and one proprietary algorithm to three publicly available datasets with ground truth step counts, including the free-living OxWalk dataset used to train our *stepcount* algorithm. They found that our algorithm was the most accurate, achieving the highest F1 score (0.89 ± 0.11) and the lowest mean absolute percentage error (8.6 ± 9.0%). However, the adults in OxWalk were much younger than those in the current study (mean age: 38.5 years vs. 65.6 and 66.7 years), highlighting the need for additional free-living validation studies across a wider range of ages and nationalities.

The sleep outcomes had the lowest ICCs in the current study, with the Matrix 003 recording more time in bed, higher overnight sleep duration, and poorer sleep efficiency than the Axivity AX3. This was particularly the case in Chinese adults, which may have been due to the shorter monitoring period of only two nights. Yuan et al. (2024) previously found that at least three days were required for stable weekly sleep parameter estimates. Further investigation using the additional sleep diary data collected is needed before pooling or comparing sleep outcomes from these two different accelerometer brands.

Strengths and limitations

A strength of this study was the concurrent wear of two different brands of accelerometer (the well-established Axivity AX3 and the newly developed Matrix 003) in a random positional order on the dominant wrist 24 hours per day for up to eight days. Furthermore, the raw acceleration data from both accelerometer brands were processed identically using three well-validated, open-source machine learning algorithms to extract meaningful physical activity and sleep outcomes (Walmsley et al., 2021; Small et al., 2024; Yuan et al., 2024). Another strength of this study was that it was conducted in both the UK and China, allowing us to identify some important cross-nationality differences. The same methods were used in both countries, except the accelerometers were distributed and returned via post in the UK but in person in China. The monitoring period was also shorter in China than in the UK (two nights and one day versus eight days), meaning the findings may be less robust. Another limitation of this study was the use of self-selected cohorts, which may limit the generalisability of the findings. Moreover, older adults may exhibit substantially different physical activity and sleep patterns compared to middle-aged and younger adults.

Conclusions

In both British and Chinese adults, average acceleration – a widely used measure of overall physical activity – was equivalent between the Axivity AX3 and the Matrix 003 worn in a random positional order on the dominant wrist. This finding suggests that data from these two different brands of accelerometer can be pooled and/or compared, including across large-scale epidemiological cohorts such as the UK Biobank, China Kadoorie Biobank, ELSA, and CHARLS. The vast majority of machine learning-derived physical activity and sleep outcomes were also equivalent or borderline equivalent. However, the cross-nationality differences observed likely stem from the UK-developed algorithms being better at identifying walking and cycling than farm or construction work. This highlights the need for further population-specific algorithm development and validation.

References

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Bull, F. C., Al-Ansari, S. S., Biddle, S., Borodulin, K., Buman, M. P., Cardon, G., … Willumsen, J. F. (2020). World Health Organization 2020 guidelines on physical activity and sedentary behaviour. *British Journal of Sports Medicine*, *54*(24), 1451–1462. https://doi.org/10.1136/bjsports-2020-102955

Chaput, J. P., Dutil, C., Featherstone, R., Ross, R., Giangregorio, L., Saunders, T. J., … Carrier, J. (2020). Sleep duration and health in adults: an overview of systematic reviews. *Applied Physiology, Nutrition, and Metabolism*, *45*(10 Suppl. 2), S218–S231. https://doi.org/10.1139/apnm-2020-0034

Chen, Y., Chan, S., Bennett, D., Chen, X., Wu, X., Ke, Y., … Doherty, A. (2023). Device-measured movement behaviours in over 20,000 China Kadoorie Biobank participants. *International Journal of Behavioral Nutrition and Physical Activity*, *20*, 138. https://doi.org/10.1186/s12966-023-01537-8

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., … Wareham, N. J. (2017). Large scale population assessment of physical activity using wrist worn accelerometers: The UK Biobank Study. *PLoS ONE*, *12*(2), e0169649. https://doi.org/10.1371/journal.pone.0169649

Ekelund, U., Tarp, J., Steene-Johannessen, J., Hansen, B. H., Jefferis, B., Fagerland, M. W., … Lee, I. M. (2019). Dose-response associations between accelerometry measured physical activity and sedentary time and all cause mortality: systematic review and harmonised meta-analysis. *BMJ*, *366*, l4570. https://doi.org/10.1136/bmj.l4570

Gall, N., Sun, R., & Smuck, M. (2022). A comparison of wrist- versus hip-worn ActiGraph sensors for assessing physical activity in adults: A systematic review. *Journal for the Measurement of Physical Behaviour*, *5*(4), 252-262. https://doi.org/10.1123/jmpb.2021-0045

Goldsack, J. C., Coravos, A., Bakker, J. P., Bent, B., Dowling, A. V., Fitzer-Attas, C., … Dunn, J. (2020). Verification, analytical validation, and clinical validation (V3): The foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). *NPJ Digital Medicine*, *3*, 55. https://doi.org/10.1038/s41746-020-0260-4

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Maylor, B. D., Edwardson, C. L., Clarke-Cornwell, A. M., Davies, M. J., Dawkins, N. P., Dunstan, D. W., … Rowlands, A. V. (2023). Physical activity assessed by wrist and thigh worn accelerometry and associations with cardiometabolic health. *Sensors*, *23*(17), 7353. https://doi.org/10.3390/s23177353

Menai, M., van Hees, V. T., Elbaz, A., Kivimäki, M., Singh-Manoux, A., & Sabia, S. (2017). Accelerometer assessed moderate-to-vigorous physical activity and successful ageing: Results from the Whitehall II study. *Scientific Reports, 8*, 45772. https://doi.org/10.1038/srep45772

O'Brien, M. W. (2021). Implications and recommendations for equivalence testing in measures of movement behaviours: a scoping review. *Journal for the Measurement of Physical Behaviour*, *4*(4), 353-362. https://doi.org/10.1123/jmpb.2021-0021

Pulsford, R. M., Brocklebank, L., Fenton, S. A. M., Bakker, E., Mielke, G. I., Tsai, L., … Stamatakis, E. (2023). The impact of selected methodological factors on data collection outcomes in observational studies of device-measured physical behaviour in adults: A systematic review. *International Journal of Behavioural Nutrition and Physical Activity*, *20*(26). https://doi.org/10.1186/s12966-022-01388-9

Rowlands, A. V. (2018). Moving forward with accelerometer-assessed physical activity: Two strategies to ensure meaningful, interpretable, and comparable measures. *Pediatric Exercise Science*, *30*(4), 450–456. https://doi.org/10.1123/pes.2018-0201

Rowlands, A. V., Plekhanova, T., Yates, T., Mirkes, E. M., Davies, M., Khunti, K., & Edwardson, C. L. (2019). Providing a basis for harmonization of accelerometer-assessed physical activity outcomes across epidemiological datasets. *Journal for the Measurement of Physical Behaviour*, *2*(3), 131–142. https://doi.org/10.1123/jmpb.2018-0073

Saint-Maurice, P. F., Troiano, R. P., Bassett Jr, D. R., Graubard, B. I., Carlson, S. A., Shiroma, E. J., … Matthews, C. E. (2020). Association of daily step count and step intensity with mortality among US adults. *JAMA*, *323*(12), 1151–1160. https://doi.org/10.1001/jama.2020.1382

Small, S., Khalid, S., Dhiman, P., Chan, S., Jackson, D., Doherty, A., & Price, A. (2021). Impact of Reduced Sampling Rate on Accelerometer-Based Physical Activity Monitoring and Machine Learning Activity Classification. *Journal for the Measurement of Physical Behaviour*, *4*(4), 298–310. https://doi.org/10.1123/jmpb.2020-0061

Small, S. R., Chan, S., Walmsley, R., von Fritsch, L., Acquah, A., Mertes, G., … Doherty, A. (2024). Self-supervised machine learning to characterize step counts from wrist-worn accelerometers in the UK Biobank. *Med Sci Sports Exerc*, *56*(10), 1945-1953. https://doi.org/10.1249/MSS.0000000000003478

Steptoe, A., Breeze, E., Banks, J., & Nazroo, J. (2013). Cohort profile: the English longitudinal study of ageing. *International Journal of Epidemiology*, *42*(6), 1640–1648. https://doi.org/10.1093/ije/dys168

van Hees, V. T., Gorzelniak, L., Dean León, E. C., Eder, M., Pias, M., Taherian, S., ... Brage, S. (2013). Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PLoS One, 8*(4), e61691. https://doi.org/10.1371/journal.pone.0061691

van Hees, V. T., Fang, Z., Langford, J., Assah, F., Mohammad, A., da Silva, I. C. M., ... Brage, S. (2014). Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: An evaluation on four continents. *J Appl Physiol (1985), 117*(7), 738–744. https://doi.org/10.1152/japplphysiol.00421.2014

Walmsley, R., Chan, S., Smith-Byrne, K., Ramakrishnan, R., Woodward, M., Rahimi, K., … Doherty, A. (2021). Reallocation of time between device-measured movement behaviours and risk of incident cardiovascular disease. *British Journal of Sports Medicine*, *56*(18), 1008-1017. https://doi.org/10.1136/bjsports-2021-104050

Wasfy, M. M., & Lee, I. M. (2022). Examining the dose-response relationship between physical activity and health outcomes. *NEJM Evidence*, *1*(12), EVIDra2200190. https://doi.org/10.1056/EVIDra2200190

Wellek, S. (2003). *Testing statistical hypotheses of equivalence*. Chapman & Hall/CRC

Yuan, H., Plekhanova, T., Walmsley, R., Reynolds, A. C., Maddison, K. J., Bucan, M., … Doherty, A. (2024). Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality. *NPJ Digital Medicine*, *7*(1), 86. https://doi.org/10.1038/s41746-024-01065-0

Zhao, Y., Hu, Y., Smith, J. P., Strauss, J., & Yang, G. (2014). Cohort profile: the China Health and Retirement Longitudinal Study (CHARLS). *International Journal of Epidemiology*, *43*(1), 61–68. https://doi.org/10.1093/ije/dys203