Received: 18 May 2022 | Revised: 25 August 2022 | Accepted: 5 October 2022

DOI: 10.1111/jsr.13760

RESEARCH ARTICLE

Validation of an automated sleep detection algorithm using data from multiple accelerometer brands

Tatiana Plekhanova[1,2] | Alex V. Rowlands[1,2] | Melanie J. Davies[1,2] | Andrew P. Hall[3,4] | Tom Yates[1,2] | Charlotte L. Edwardson[1,2]

[1]Diabetes Research Centre, University of Leicester, Leicester, UK

[2]NIHR Leicester Biomedical Research Centre, University of Leicester, Leicester, UK

[3]University Hospitals of Leicester NHS Trust, Leicester, UK

[4]Department of Health Sciences, University of Leicester, Leicester, UK

Correspondence
Charlotte L. Edwardson, Diabetes Research Centre, Leicester General Hospital, University of Leicester, Leicester, UK.
Email: ce95@le.ac.uk

Funding information
NIHR Leicester Biomedical Research Centre

Summary

To evaluate the criterion validity of an automated sleep detection algorithm applied to data from three research-grade accelerometers worn on each wrist with concurrent laboratory-based polysomnography (PSG). A total of 30 healthy volunteers (mean [SD] age 31.5 [7.2] years, body mass index 25.5 [3.7] kg/m^2) wore an Axivity, GENEActiv and ActiGraph accelerometer on each wrist during a 1-night PSG assessment. Sleep estimates (sleep period time window [SPT-window], sleep duration, sleep onset and waking time, sleep efficiency, and wake after sleep onset [WASO]) were generated using the automated sleep detection algorithm within the open-source GGIR package. Agreement of sleep estimates from accelerometer data with PSG was determined using pairwise 95% equivalence tests (±10% equivalence zone), intraclass correlation coefficients (ICCs) with 95% confidence intervals and limits of agreement (LoA). Accelerometer-derived sleep estimates except for WASO were within the 10% equivalence zone of the PSG. Reliability between data from the accelerometers worn on either wrist and PSG was moderate for SPT-window duration (ICCs ≥ 0.65), sleep duration (ICCs ≥ 0.54), and sleep onset (ICCs ≥ 0.61), mostly good for waking time (ICCs ≥ 0.80), but poor for sleep efficiency (ICCs ≥ 0.08) and WASO (ICCs ≥ 0.08). The mean bias between all accelerometer-derived sleep estimates worn on either wrist and PSG were low; however, wide 95% LoA were observed for all sleep estimates, apart from waking time. The automated sleep detection algorithm applied to data from Axivity, GENEActiv and ActiGraph accelerometers, worn on either wrist, provides comparable measures to PSG for SPT-window and sleep duration, sleep onset and waking time, but a poor measure of wake during the sleep period.

KEYWORDS
ActiGraph, Axivity, GENEActiv, polysomnography, sleep duration

1 | INTRODUCTION

Numerous studies have implicated disturbed sleep with adverse health outcomes including increased risk of obesity and diabetes (Reutrakul & Van Cauter, 2018), cardiovascular disease (Cappuccio et al., 2011), and mental health conditions (João et al., 2018); however, the quality of evidence depends on the validity of the measurement of sleep parameters. Most epidemiological studies linking poor

sleep with negative health outcomes rely on a single, retrospective self-report of sleep (Girschik et al., 2012), which is prone to recall bias (Lauderdale et al., 2008). Polysomnography (PSG), the 'gold standard' for sleep assessment, is expensive, labour-intensive (Sadeh, 2015), and hence not feasible for large scale studies. The use of wrist actigraphy devices such as Actiwatch is cost-effective, non-intrusive, and allows continuous recording of activity over days or weeks under free-living conditions (Martin & Hakim, 2011).

Wrist-worn accelerometers that collect raw data have become increasingly used for physical activity assessment in large population-based studies (e.g., National Health and Nutrition Examination Survey [Loprinzi & Cardinal, 2011], UK Biobank [Doherty et al., 2017]). The three most widely used research-grade raw data accelerometer brands used in epidemiological studies are the Axivity (Axivity Ltd., Newcastle, UK), ActiGraph (ActiGraph LLC, Pensacola, FL, USA), and GENEActiv (Actvinsights Ltd., Kimbolton, Cambridgeshire, UK). Implementation of a 24-h protocol in these studies means that they could be used to assess sleep and physical activity with a single device in a free-living setting. Availability of raw data also potentially improves comparability among different accelerometer brands and enables the development and application of novel algorithms. In the early devices, due to limited memory and battery capacity, data were pre-processed onboard the device into the form of 'counts' through proprietary algorithms that complicates comparability of data from different devices. Some of the most widely used count-based sleep detection algorithms include those developed by Sadeh et al. (1994) and Cole et al. (1992).

Recently, van Hees et al. (2015) developed a sleep detection algorithm based on angular wrist rotation measured with raw acceleration data. Sleep is defined as sustained inactivity, determined as the absence of change in wrist rotation greater than 5 ° for 5 min, or user-defined duration within a defined sleep window (starting at sleep onset and ending when waking up) (van Hees et al., 2015). The sleep window can be obtained from sleep onset and offset timings participants record in their sleep log (sleep log method) or from an automated sleep window detection (the Heuristic Algorithm looking at Change of Z-Angle, HDCZA [van Hees et al., 2018]). This algorithm is available as part of an open-source software which combines sleep and activity data over the 24-h day, can be applied to raw accelerometer data irrespective of accelerometer brand, and is now widely used in the research community (Jones et al., 2019; Wendt et al., 2020). However, the sleep algorithm has undergone limited validation to date and parameters such as wake after sleep onset (WASO) have not been validated. Additionally, the impact of handedness on the accelerometer-derived sleep estimates compared to PSG has not been assessed.

The primary aim of the present study was to validate an automated sleep detection algorithm (HDCZA) when applied to data from the ActiGraph, Axivity, and GENEActiv accelerometers worn simultaneously on both wrists against PSG in a healthy adult population. The secondary aim was to compare the automated sleep window detection and the sleep log method.

2 | METHODS

2.1 | Study participants

Adults aged 18–65 years inclusive without known sleep disorder (self-reported by the participants) were recruited through the University of Leicester, University Hospitals of Leicester, and word of mouth between January and December 2020. This study was approved by the Nottingham 1 Research Ethics Committee (19/EM/0275) and registered on the ClinicalTrials.gov Protocol Registration (NCT04288557). Informed consent was sought for all participants. Individuals unwilling or unable to give informed consent and without a good command of the English language were excluded from the study.

2.2 | Procedure

Participants underwent informed consent on a separate day prior to the PSG assessment. Demographic details such as age, sex, ethnicity, medical history, and habitual sleep duration were recorded. Participants' height and weight were measured and recorded to the nearest 0.5 cm and 0.5 kg, respectively. Each participant self-reported their handedness.

The PSG assessment took place on a weeknight. Participants arrived early in the evening and were fitted with three accelerometers on each wrist: the GENEActiv Original (ActivInsights Ltd, Cambridgeshire, UK), Axivity AX3 (Axivity Ltd, Newcastle, UK), and ActiGraph GT9X Link (ActiGraph LLC, Pensacola, FL, USA). For comfort reasons, two of the devices (the Axivity and ActiGraph) were taped together to reduce the number of straps on each wrist. The relative position of the three devices on a given wrist was randomised between the participants but consistent between wrists for each participant. After fitting the accelerometers, participants were prepared for the PSG assessment in accordance with the American Academy of Sleep Medicine (AASM) guidelines (Berry et al., 2012) by a trained technician. The recording began when participants expressed willingness to go to bed and ended the following morning (usually between 6:00 and 7:00 a.m.). In the morning, participants were asked about time they tried to sleep, fell asleep and woke up. To assess the extent to which sleep duration in the laboratory setting was representative of habitual sleep duration all participants were fitted with a GENEActiv accelerometer on their non-dominant wrist for 8 days. Participants completed a sleep log for the days they wore the device.

2.3 | Measures

2.3.1 | Polysomnography

Digital PSG was conducted using an Alice-5 unit recorder (Philips Respironics) in conjunction with Philips Respironics proprietary software. A total of seven channels were utilised: two bipolar and two

common reference electroencephalogram (EEG) channels (sampled at 100 Hz, gain factor 1, high pass [30 Hz] and low pass [0.3 Hz] filters applied), two electro-oculogram channels (EOG; sampled at 100 Hz) and one bilateral electromyogram submentalis (EMG; sampled at 200 Hz, gain factor 1, high pass [90 Hz] and low pass [0.3 Hz]) channel. The EEG array was attached in accordance with the 10:20 system of electrode placement (Jasper, 1958); the final montage included C3-A2, C4-A1, fP1-F3, and O1-P3 placements.

The sleep PSG data were processed using Philips Respironics software. The beginning of sleep scoring was determined by 'lights out' time and ended at 'lights on'. Sleep parameters were scored using the AASM criteria (Berry et al., 2012) by a single specialist sleep technologist. This included use of 30-s epochs for sleep staging, assigning epochs a state of sleep or wake, documenting and generating indices of the frequency limb movement events.

2.3.2 | Accelerometers

The GENEActiv Original, Axivity AX3, and ActiGraph GT9X Link are triaxial accelerometers with a dynamic range of ±8 ***g***, where ***g*** is equal to the Earth's gravity. All accelerometers were configured to record at a frequency of 100 Hz and initialised using the same personal computer (PC). However, the accelerometers could not be precisely time synchronised with the PSG (accelerometers initialised on a different PC). GENEActiv devices were initialised, and data were downloaded and saved in raw format as .bin files using GENEActiv PC software version 3.2. Axivity devices were set up and data downloaded with OmGui software version 1.0.0.30 (Open Movement, Newcastle, UK). ActiGraph Link GT9X devices were initialised and downloaded using ActiLife version 6.13.3, saved in raw format as .gt3x, then converted to .csv format for data processing.

All accelerometer files were processed using R-package GGIR version 2.5 (https://cran.r-project.org/web/packages/GGIR/) (Migueles et al., 2019). Signal processing in GGIR includes autocalibration using local gravity as a reference, detection of sustained abnormally high values, detection of non-wear, calculation of the average magnitude of dynamic acceleration (i.e., the vector magnitude of acceleration corrected for gravity [Euclidean Norm – 1 ***g***] in milli-gravitational units averaged over 5-s epochs).

The automated sleep window detection algorithm calculates wrist rotation (changes in the z-angle) for each 5-min rolling window and values under the 10th percentile over an individual day (noon-to-noon). The algorithm then detects blocks lasting >30 min, with gaps <60 min counted towards the identified blocks. The longest block in the day between noon–noon represents the sleep window (van Hees et al., 2018).

When a sleep log is used to guide the algorithm in GGIR, it is also possible to calculate sleep onset latency (SOL) if onset reported by participants corresponds to intention to fall asleep as opposed to the timing of sleep onset. In this study, the results reported for the sleep log condition used timings of sleep onset to guide the algorithm. In addition, accelerometer data were also processed with onset

TABLE 1 Descriptive statistics of sleep outcomes (mean [SD] or median [IQR], where not normally distributed, per night) by each method ($n = 30$)

Non-dominant wrist				Dominant wrist			
				ActiGraph			
Sleep parameter	PSG Mean (SD) or median [IQR]	Axivity	GENEActiv		Axivity	GENEActiv	ActiGraph
SPT-window, min	424.6 (46.7)	426.9 [376.6–464.9]	427.3 [376.4–463.3]	417.6 [376.6–461.3]	419.0 (55.4)	430.8 [382.0–460.1]	427.4 (44.8)
Sleep onset, clock time (min)	11:00 p.m. (47.0)	11:16 p.m. (60.0)	11:14 p.m. (60.0)	11:14 p.m. (60.0)	11:08 p.m. (48.0)	11:14 p.m. (59.0)	11:05 p.m. (48.0)
Waking time, clock time (min)	6:05 a.m. (26.0)	6:13 a.m. (27.0)	6:14 a.m. (24.5)	6:12 a.m. (28.1)	6:17 a.m. [5:52–6:29 a.m.]	6:17 a.m. [5:55–6:32 a.m.]	6:13 a.m. (28.7)
Sleep duration, min	382.6 (49.7)	387.8 [351.6–429.1]	388.9 [352.5–427.4]	389.5 [353.2–431.8]	387.2 [355.3–424.6]	390.0 [359.4–417.5]	388.3 (41.5)
Sleep efficiency, %	91.0 [87.3–95.8]	91.8 (4.5)	92.0 (4.5)	91.9 (4.4)	91.3 (4.6)	92.2 (4.7)	90.9 (4.4)
WASO, min	42.0 [14.5–53.6]	32.1 [20.6–40.6]	34.6 (20.9)	34.5 (20.0)	35.0 [17.8–50.5]	33.8 (21.8)	39.1 (20.6)

Abbreviations: IQR, interquartile range; PSG, polysomnography; SPT-window, sleep period time window; WASO, wake after sleep onset.

TABLE 2 Intraclass correlations, agreement and equivalence zones between accelerometers worn on the non-dominant wrist and polysomnography ($n = 30$)

Agreement (Bland–Altman)[a]											
	ICC[b] (95% CI)			Mean bias			95% limits of agreement			Equivalence zones, %[c]		
Non-dominant wrist	AX	GEN	ActiG	AX	GEN	ActiG	AX	GEN	ActiG	AX	GEN	ActiG
SPT-window (min)	0.67 (0.41, 0.83)	0.68 (0.43, 0.83)	0.68 (0.42, 0.83)	7.6	5.5	7.0	90.6	90.0	90.4	7.1	6.7	6.9
Sleep onset (min)	0.61 (0.33, 0.79)	0.61 (0.33, 0.79)	0.61 (0.34, 0.79)	−15.6	−14.4	−13.9	94.1	94.1	94.1	2.1	2.0	2.0
Waking time (min)	0.88 (0.56, 0.95)	0.89 (0.38, 0.96)	0.90 (0.69, 0.96)	−8.5	−9.0	−7.1	21.1	17.3	20.9	0.7	0.7	0.6
Sleep duration (min)	0.73 (0.51, 0.86)	0.74 (0.52, 0.87)	0.73 (0.51, 0.86)	0.6	−1.9	−0.5	79.6	76.0	79.4	4.4	3.9	4.3
Sleep efficiency (%)	0.08 (−0.29, 0.42)	0.16 (−0.20, 0.49)	0.12 (−0.25, 0.45)	−1.6	−1.8	−1.7	15.9	15.1	15.5	5.1	5.1	4.9
Wake after sleep onset (min)	0.08 (−0.28, 0.42)	0.17 (−0.19, 0.50)	0.12 (−0.25, 0.45)	7.0	7.4	7.5	69.4	65.7	67.2	36.8	39.9	37.3

Abbreviations: ActiG, ActiGraph GT9X; AX, Axivity, GEN, GENEActiv.

[a]Bland and Altman: bias calculated as polysomnography − accelerometer.

[b]Single measure, absolute agreement. Sleep efficiency and wake after sleep onset $p > 0.05$. All other outcomes $p < 0.05$.

[c]The equivalence zone required for the accelerometers and polysomnography sleep estimates to be deemed equivalent.

indicating intention to fall asleep. These data were used to demonstrate agreement between accelerometer SOL and PSG SOL, thus only results for SOL are presented from these outputs.

To demonstrate the agreement between PSG and count-based sleep detection, the zero-crossing algorithm developed by Sadeh et al. (1994) also available in GGIR was applied to data.

2.4 | Statistical methods

The variables of interest included: sleep period time window duration (SPT-window; time difference between falling asleep and waking up), sleep duration (accumulated nocturnal sustained inactivity bouts [time spent sleeping] within SPT-window), sleep efficiency (ratio of sleep duration compared to SPT-window), WASO (total minutes of wake in the sleep window), timing of sleep onset, waking time, and SOL (time it took to fall asleep; only in the sleep log method).

Only those participants who provided all six valid accelerometer files (three devices on both wrists) and valid PSG recording output were included in analyses. Descriptive characteristics were calculated for the study population, as well as the sleep variables estimated by each method mean (SD) or median (25–75th percentile) after testing for normality. Equivalence testing was performed using the confidence interval (CI) approach (95% equivalence testing) (Dixon et al., 2018), with PSG as the reference method. A 90% CI was obtained for the difference in means of sleep estimates from PSG and accelerometers. This CI was then used to test equivalence at $\alpha = 5\%$ (Dixon et al., 2018). To be deemed equivalent, the 90% CI of for the mean of one accelerometer should fall entirely within the predefined equivalence zone of ±10% of the mean of PSG. The ±10% equivalence zone of the reference mean was selected based on previous validation studies of physical behaviours measures. However, as presented by O'Brien (2021), the selection of an equivalence zone (i.e., narrow versus broad) influences equivalence test outcomes. Therefore, as there is no evidence of a standardised equivalence criteria, this study also reported a minimum equivalence zone for the PSG estimates that include 90% CI of the accelerometer sleep estimates. Log transformation was applied to data where data were not normally distributed. The agreement between sleep estimates was examined using intraclass correlation coefficients (ICCs, single measures, absolute agreement) with 95% CI and limits of agreement (LoA) (Bland & Altman, 1986). The level of reliability was classified as 'poor' (ICC > 0.5), 'moderate' (ICC 0.5–0.75), 'good' (ICC > 0.75–0.9) and 'excellent' (ICC > 0.9) (Koo & Li, 2016). Bland–Altman plots were created to assess the pattern and magnitude of differences in sleep estimates between PSG and each accelerometer. Paired *t* tests were conducted to assess whether habitual sleep duration measured using 7-day accelerometry with and without use of a sleep log was different from sleep duration defined by 1-night PSG.

Additionally, sensitivity, specificity, and accuracy of the binary classification of sleep (any sleep stage) and wakefulness, were derived from epoch-by-epoch comparison to PSG. However, it should be noted that it was not possible to precisely synchronise the data and raw PSG data was not available for one participant. Accelerometer 5-s

TABLE 3 Intraclass correlations, agreement and equivalence zones between accelerometers worn on the dominant wrist and polysomnography ($n = 30$)

Agreement (Bland–Altman)[a]											
	ICC[b] (95% CI)			Mean bias			95% limits of agreement			Equivalence zones, %[c]		
Dominant wrist	AX	GEN	ActiG	AX	GEN	ActiG	AX	GEN	ActiG	AX	GEN	ActiG
SPT-window (min)	0.66 (0.40, 0.82)	0.65 (0.38, 0.82)	0.86 (0.72, 0.93)	5.6	7.1	−2.8	82.9	88.0	48.8	5.1	6.4	2.7
Sleep onset (min)	0.87 (0.74, 0.94)	0.62 (0.34, 0.80)	0.87 (0.75, 0.94)	−7.2	−13.8	−5.4	46.8	94.1	47.0	1.0	2.0	1.0
Waking time (min)	0.40 (0.05, 0.66)	0.82 (0.63, 0.91)	0.80 (0.57, 0.91)	−2.8	−6.9	−8.5	74.5	30.5	30.9	0.8	0.7	0.7
Sleep duration (min)	0.54 (0.23, 0.76)	0.69 (0.44, 0.84)	0.62 (0.35, 0.80)	0.5	−1.1	−5.6	96.0	80.0	78.2	4.5	3.7	5.1
Sleep efficiency (%)	0.28 (−0.08, 0.58)	0.09 (−0.26, 0.43)	0.29 (−0.08, 0.59)	−1.1	−2.0	−0.7	14.1	16.1	13.9	4.1	5.3	3.7
Wake after sleep onset (min)	0.27 (−0.10, 0.57)	0.10 (−0.25, 0.44)	0.30 (−0.07, 0.59)	5.0	8.2	2.9	62.3	69.8	60.8	28.0	40.0	35.5

Abbreviations: ActiG, ActiGraph GT9X; AX, Axivity, GEN, GENEActiv.

[a]Single measure, absolute agreement. Sleep efficiency and wake after sleep onset $p > 0.05$. All other outcomes $p < 0.05$.

[b]Bland and Altman: bias calculated as polysomnography – accelerometer.

[c]The equivalence zone required for the accelerometers and polysomnography sleep estimates to be deemed equivalent.

epoch data were converted into 30-s epochs by taking the mode for each 12 sets of data. If 50% of epochs were classified as 'sleep' and 50% as 'wake', then the transformed 30-s epoch was classified as 'wake'. The SPT-window was defined from sleep onset and waking time from PSG recordings.

Descriptive statistics, *t* tests and ICCs were performed using the IBM Statistical Package for the Social Sciences (SPSS) version 26 and LoA using GraphPad Prism 7. Equivalence tests were conducted in Minitab (version 21.1). Alpha was set at 0.05.

3 | RESULTS

3.1 | Study participants

In total, 31 healthy volunteers underwent a 1-night PSG assessment in the sleep laboratory. Two accelerometers malfunctioned (same participant), thus data from 30 participants (mean [SD] age 31.5 [7.2] years, body mass index 25.5 [3.7] kg/m^2, 63% females and 77% White) were available for the analyses. The description of the participants and sleep characteristics defined by PSG are shown in Table S1. The median wear duration of the accelerometer during the habitual sleep assessment was 6.8 days and 6.0 nights.

3.2 | Comparison of sleep estimates by polysomnography and automated sleep window detection

Table 1 presents descriptive statistics of sleep estimates (mean [SD] or median [interquartile range, IQR]) measured by PSG and accelerometers worn on the non-dominant and dominant wrist.

Tables 2 and 3 shows the ICCs, mean bias, 95% LoA, and equivalence zones between each accelerometer worn on (a) the non-dominant and (b) dominant wrist and PSG for all sleep estimates. Equivalence and Bland–Altman plots for all sleep estimates between each accelerometer worn on the non-dominant and dominant wrist and PSG are presented in Figure 1 and in Figure S1a–c, respectively.

Reliability between all three brands of accelerometers when worn on either wrist and PSG for SPT-window was moderate, with ICCs of 0.65–0.86. The mean differences between PSG and the three brands of accelerometers worn on either wrist were low (up to 8 min); however, 95% LoA (up to ±91 min) were wide. The size of the zone required for PSG and accelerometer estimates of the SPT-window to be deemed equivalent was ~7% for the non-dominant wrist. This was similar for the dominant wrist for the Axivity and GENEActiv, but smaller for the ActiGraph, at 2.7%.

Regardless of wrist placement, reliability was moderate for sleep onset (ICCs of 0.61–0.87) and mostly good for waking time (ICCs of 0.80–0.90) between the three accelerometer brands and PSG. Sleep onset and waking time detected by accelerometers on either wrist compared to PSG was up to 16 and 9 min later, respectively. Noticeably wider 95% LoA were observed between all accelerometers and

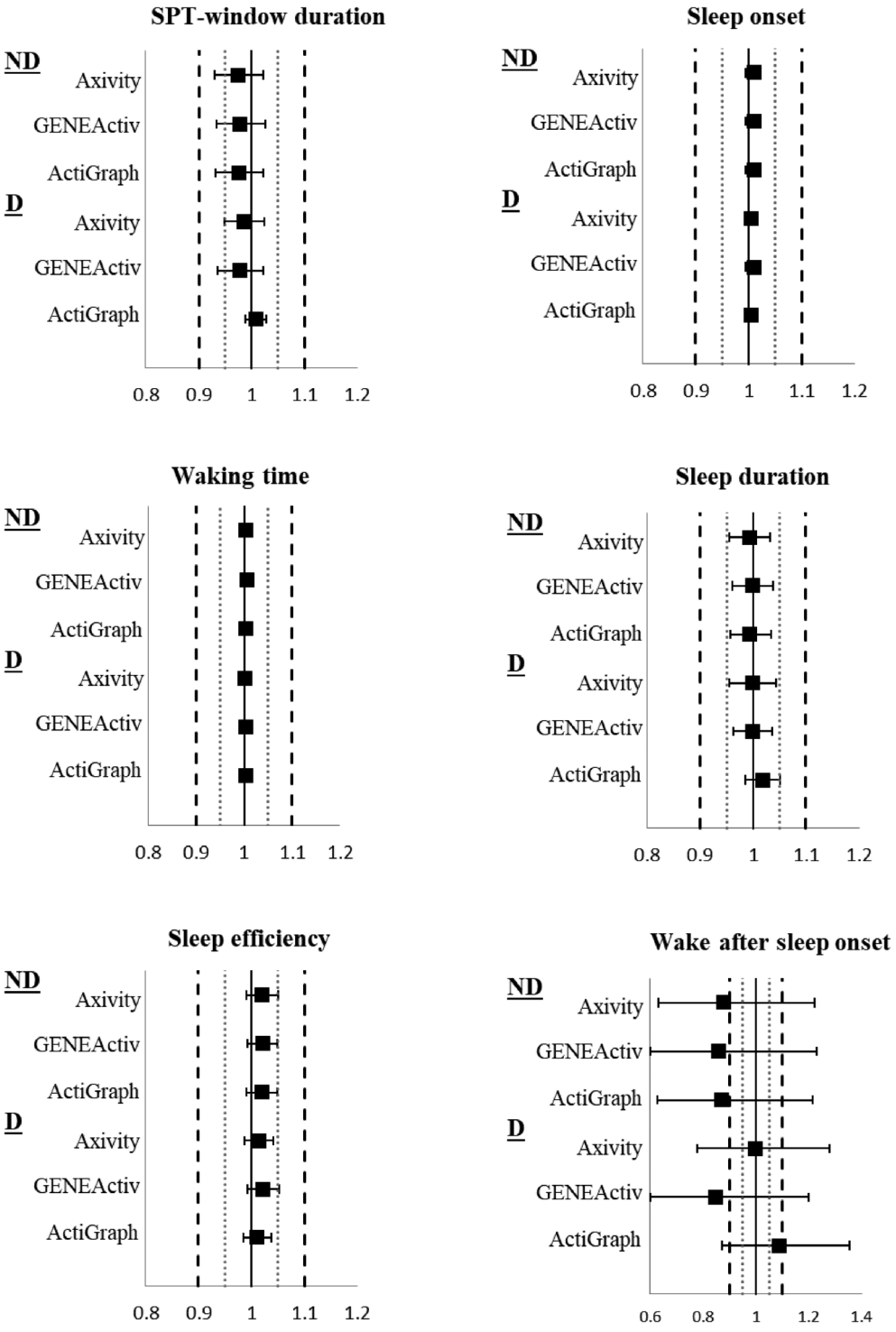

FIGURE 1 Equivalence between each accelerometer worn on the non-dominant wrist and dominant wrist and polysomnography. Error bars represent the 95% confidence interval of the ratio. Equivalence = 1 (solid line). Dashed lines indicate the 10% equivalence zone and dotted lines indicate a 5% equivalence zone. D, dominant; ND, non-dominant

PSG for sleep onset (up to ±1 h 34 min) compared to waking time (up to ±31 min; except for the Axivity on dominant wrist LoA were ± 75 min). Across the three accelerometer brands and both wrists, the equivalence zone required for PSG and accelerometer estimates of sleep onset and waking time to be deemed equivalent was similar at up to 2.1% and 0.8%, respectively.

Moderate reliability was observed for sleep duration (ICCs of 0.54–0.74). The mean differences between the three accelerometer brands worn on either wrist and PSG were low; however, 95% LoA (up to ±96 min) were high. Sleep duration was slightly overestimated (up to 6 min) by accelerometers compared to PSG worn on either wrist except for the Axivity. Across the three accelerometer brands and both wrists, the equivalence zone required for PSG and accelerometer estimates of sleep duration to be deemed equivalent was similar at up to 5.1%.

Between each accelerometer and PSG, reliability was poor for sleep efficiency (ICCs of 0.08–0.29) and WASO (ICCs of 0.08–0.30). Sleep efficiency was slightly overestimated and WASO

underestimated by accelerometers irrespective of brand or wrist compared to PSG, by up to 2% and 8 min, respectively. Wide 95% LoA for sleep efficiency (up to ±16%) and WASO (up to ±70 min) were observed. Bland–Altman plots also revealed a pattern in the bias between accelerometer-measured sleep efficiency and WASO and PSG such that accelerometer measures showed smaller differences compared to PSG when sleep efficiency increased and WASO decreased. Regardless of accelerometer brand or wrist placement, equivalence zone required of PSG for sleep efficiency was reached at up to 5.1%. The equivalence zones for WASO ranged from 28% to 40%.

3.3 | Comparison of sleep estimates by PSG and the sleep log method

The descriptive statistics of sleep estimates (mean [SD] or median [IQR] per night) measured by PSG and accelerometers worn on the non-dominant and dominant wrists using the sleep log method are shown in Table S2. Table S3a,b demonstrates ICCs, mean bias, 95% LoA, and equivalence zones between each accelerometer worn on (a) non-dominant and (b) dominant wrists and PSG for all sleep estimates.

The pattern of the results using the sleep log method was largely unaltered compared to the automated sleep window detection.

Table S4 shows ICCs, mean bias, 95% LoA, and equivalence zones between each accelerometer worn on both wrists and PSG for SOL. Reliability was poor (ICCs of −0.24 to −0.11) between each accelerometer brand worn on either wrist and PSG. SOL was underestimated by accelerometers by up to 5 min with wide 95% LoA (up to ±49 min) observed. Regardless of accelerometer brand or wrist placement the equivalence zones for SOL ranged from 40% to 73%.

3.4 | Comparison of sleep estimates by PSG and the zero-crossing algorithm

The descriptive statistics of sleep estimates (mean [SD] or median [IQR] per night) measured by PSG and accelerometers worn on the non-dominant and dominant wrists using the zero-crossing method are shown in Table S5. Table S6 demonstrates ICCs, mean bias, 95% LoA, and equivalence zones between each accelerometer worn on (a) non-dominant and (b) dominant wrists and PSG for all sleep estimates.

The agreement between PSG and the zero-crossing algorithm was comparable to the automated sleep window detection based on angular wrist rotation for SPT-window duration, sleep onset and waking time; however, noticeably lower for sleep duration, sleep efficiency and WASO.

3.5 | Epoch-by-epoch comparison to PSG

Relative to PSG, the overall accuracy of the algorithm was 84%. Sensitivity (identifying sleep as sleep) was generally high at 92%, while specificity (identifying wake as wake) was relatively low at 20%. Results were similar for all brands of accelerometers on either wrist (Table S7).

3.6 | Comparison of 1-night PSG and 7-day accelerometry

Habitual sleep duration measured using 7-day accelerometry did not significantly differ from 1-night PSG assessment. The mean (SD) sleep duration defined by PSG was 385.7 (29) min versus accelerometer sleep duration in presence of a sleep log 373.1 (55) min ($p = 0.29$) or absence of a sleep log 395.7 (49.5) min ($p = 0.26$).

4 | DISCUSSION

Irrespective of brand of monitor or placement on the dominant or non-dominant wrist, all accelerometer sleep estimates apart from WASO, could be deemed equivalent to sleep estimates measured by PSG. This was true whether the sleep window was determined using automated detection or the sleep log method. These data provide further evidence that GGIR can be used to provide valid and monitor-agnostic estimates of sleep quantity and timing, as well as physical activity, in the large studies worldwide employing research-grade raw data accelerometers. Overall, the results suggests that these accelerometers can be used to capture an accurate picture of the whole 24-h physical behaviour profile. However, poor detection of wakefulness during sleep by the algorithm tested should be considered in the future studies.

An important finding of this study was that overall agreement between PSG and accelerometers for all sleep estimates, irrespective of the accelerometer brand or placement, was similar. Given that handedness had little impact on accelerometer-derived sleep estimates, researchers could decide on wrist placement based on other outcomes of interest. For instance, in physical activity research the non-dominant wrist is commonly used for device placement (Dieu et al., 2017). Of note, although until recently hip placement has been most widely used for physical activity assessment (Ainsworth et al., 2015), hip-mounted accelerometers (GTX3+) compared to wrist-worn demonstrate lower agreement for sleep duration, sleep efficiency, WASO, and SOL when compared to PSG (Full et al., 2018; Slater et al., 2015; Zinkhan et al., 2014).

The sleep detection algorithm used in the present study has been previously developed and validated in a sample of clinical patients ($n = 28$) and healthy young adults ($n = 22$) (van Hees et al., 2018). In healthy sleepers, sleep parameters were assessed by the Axivity accelerometer worn on the non-dominant wrist and compared to PSG. The findings from this study are consistent with the study by van Hees et al. (2018) reporting small differences between accelerometer and PSG for SPT-window and sleep duration, sleep onset and waking time. The findings from the equivalence tests in this study confirm that these sleep estimates derived from the HDCZA algorithm are

comparable to PSG. Nonetheless, healthy sleepers usually have high sleep efficiencies (in this study >90% on average), therefore, it should be noted that an algorithm that always classified 'sleep' within the SPT-window would also have high sensitivity for sleep. It is worth noting, that while mean bias for sleep efficiency were low, the reliability was poor irrespective of accelerometer brand or wrist placement compared to PSG. Furthermore, a difference of 5% is considered clinically significant in sleep efficiency (Smith et al., 2018). As the equivalence tests revealed that accelerometer measures of sleep efficiency were around the 5% equivalence zone, this should be considered when interpreting data.

In this study, the HDCZA algorithm poorly detected periods of wakefulness in SPT-window compared to PSG. WASO was underestimated by accelerometers, and in two participants it was underestimated by ~1.5 h. These findings were not surprising, as van Hees et al. (2015) previously reported moderate specificity (45%) of the algorithm to detect wake during sleep period in a sample of sleep clinic patients. Similar specificity of 34%–46% was reported for count-based sleep detection using algorithms developed by the Sadeh et al. (1994) or Cole et al. (1992) (Quante et al., 2018; Slater et al., 2015). In this study specificity for WASO was lower (20%) than in previous studies. One explanation for low specificity could be that a few participants had very little wake during the sleep period. In accelerometry-based assessment, wrist movement indicates wakefulness and immobility indicates sleep. However, immobility is possible during periods of wakefulness and as such can be mistakenly identified as sleep periods by accelerometers. Bland–Altman plots also revealed a pattern in the bias between accelerometer-measured WASO and PSG such that accelerometer measures showed smaller differences compared to PSG when WASO decreased. Therefore, more wakefulness in the sleep period will likely result in misclassification of WASO.

Implementing the algorithm developed by Sadeh et al. (1994) based on zero-crossing in GGIR to generate sleep estimates demonstrated poor detection of sleep episodes with the SPT-window resulting in substantial underestimation of WASO.

Further, this study compared sleep estimates from accelerometers with PSG using sleep onset and offset timings to guide the algorithm. Overall, the use of a sleep log did not improve the level of agreement of sleep estimates between accelerometers and PSG. Although the participants were asked about their sleep onset and waking times not long after awakening, it appears that estimating these timings by self-report is challenging, particularly estimating timing of sleep onset.

A potential value of a sleep log is that when the timing of intention to fall asleep is recorded, it is possible to calculate SOL. This estimate generated in GGIR has not been validated against PSG previously. In this study, SOL was not comparable to PSG as indicated by poor reliability and wide equivalence zones. A systematic review found that in healthy participants actigraphy (usually using the algorithms developed by Sadeh et al. [1994] and Cole et al. [1992]) generally underestimated SOL and showed poor to moderate agreement (ICCs range −0.07 to 0.56) when compared to PSG (Scott et al., 2020). It should be noted, that in these studies SOL is usually defined as 'lights off' to the first epoch scored as sleep from PSG recording not 'intention to fall asleep' reported by the participants. Hence, generally reporting somewhat better agreement between PSG and accelerometers. A sleep log can also be valuable to guide the algorithm to identify the main sleep episode if someone is very inactive or has multiple episodes of sleep throughout the day, which can result in misclassification of the SPT-window. However, the consistency of the algorithm to identify the SPT-window when the sleep estimates were compared with and without a sleep log in free-living setting has been previously demonstrated (Plekhanova et al., 2020).

5 | STRENGTHS AND LIMITATIONS

The strengths of this study include the simultaneous comparison of three research-grade accelerometers worn on each wrist with PSG. Importantly, sleep data were generated using an automated sleep detection algorithm that can be applied to raw accelerometer data irrespective of accelerometer brand and is available as part of the open-source software. This method allows identical data handling and facilitates comparability of the results. The limitations of this study include the small sample size, which makes it hard to generalise findings beyond the specific population recruited for this study. The sample consisted of healthy volunteers; thus, the findings cannot be generalised to individuals with sleep disorders. While the comparison of wrist accelerometers to the 'gold standard' PSG is a strength, 1 night of PSG assessment may not represent participants' habitual sleep due to the first-night effect. However, the habitual sleep duration of the participants was similar to that of 1 night of PSG as indicated by 7-day accelerometry. The influence of the first-night effect should also be a minor issue when evaluating the agreement between the measurement methods. Also, the GENEActiv was worn adjacent to the Axivity, which was taped on top of the ActiGraph, this may have impacted on the agreement between the three devices. Future studies should randomise the positioning of all devices and/or consider different set-ups of accelerometers to establish the impact on agreement between the devices and PSG.

6 | CONCLUSION

This study suggests that the automated sleep detection algorithm HDCZA applied to the Axivity, GENEActiv and ActiGraph accelerometers, worn on either wrist, provides comparable measures with PSG of SPT-window and sleep duration, sleep onset and waking time. Accelerometry should be used cautiously in studies where estimates of sleep efficiency and wakefulness during sleep period are important.

AUTHOR CONTRIBUTIONS

All authors developed the study concept and contributed to the study design. Tatiana Plekhanova collected and analysed the data and

drafted the manuscript. Tatiana Plekhanova, Alex V. Rowlands, Charlotte L. Edwardson contributed to data interpretation. All authors contributed to critical review of the manuscript and approved the final version for submission.

ACKNOWLEDGMENTS

The authors would like to acknowledge Somaya Turk for assisting with data collection and thank all participants who volunteered to take part in this study. The research was supported by the National Institute for Health Research (NIHR) Leicester Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. No external sources of funding were accessed.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Tatiana Plekhanova https://orcid.org/0000-0002-8270-4655

REFERENCES

Ainsworth, B., Cahalin, L., Buman, M., & Ross, R. (2015). The current state of physical activity assessment tools. *Progress in Cardiovascular Diseases*, *57*(4), 387–395. https://doi.org/10.1016/j.pcad.2014.10.005

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Marcus, C. L., & Vaughn, B. V. (2012). *The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications*. American Academy of Sleep Medicine. www.aasmnet.org

Bland, J. M., & Altman, D. (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

Cappuccio, F. P., Cooper, D., D'elia, L., Strazzullo, P., & Miller, M. A. (2011). Sleep duration predicts cardiovascular outcomes: A systematic review and meta-analysis of prospective studies. *European Heart Journal*, *32*(12), 1484–1492. https://doi.org/10.1093/eurheartj/ehr007

Cole, R. J., Kripke, D. F., Gruen, W., Mullaney, D. J., & Gillin, J. C. (1992). Automatic sleep/wake identification from wrist activity. *Sleep*, *15*(5), 461–469. https://doi.org/10.1093/sleep/15.5.461

Dieu, O., Mikulovic, J., Fardy, P. S., Bui-Xuan, G., Béghin, L., & Vanhelst, J. (2017). Physical activity using wrist-worn accelerometers: Comparison of dominant and non-dominant wrist. *Clinical Physiology and Functional Imaging*, *37*(5), 525–529. https://doi.org/10.1111/cpf.12337

Dixon, P. M., Saint-Maurice, P. F., Kim, Y., Hibbing, P., Bai, Y., & Welk, G. J. (2018). A primer on the use of equivalence testing for evaluating measurement agreement. *Medicine and Science in Sports and Exercise*, *50*(4), 837. https://doi.org/10.1249/MSS.0000000000001449

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., van Hees, V. T., Trenell, M. I., & Owen, C. G. (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

Full, K. M., Kerr, J., Grandner, M. A., Malhotra, A., Moran, K., Godoble, S., Natarajan, L., & Soler, X. (2018). Validation of a physical activity accelerometer device worn on the hip and wrist against polysomnography. *Sleep Health: Journal of the National Sleep Foundation*, *4*(2), 209–216. https://doi.org/10.1016/j.sleh.2017.12.007

Girschik, J., Fritschi, L., Heyworth, J., & Waters, F. (2012). Validation of self-reported sleep against Actigraphy. *Journal of Epidemiology*, *22*(5), 462–468. https://doi.org/10.2188/jea.JE20120012

Jasper, H. H. (1958). Report of committee on methods of clinical examination in EEG: The ten-twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology*, *10*, 371.

João, K. A. D. R., Neves, S., de Jesus, C. C., & Pinto, P. (2018). The impact of sleep quality on the mental health of a non-clinical population. *Sleep Medicine*, *46*, 69–73. https://doi.org/10.1016/j.sleep.2018.02.010

Jones, S. E., van Hees, V. T., Mazzotti, D. R., Marques-Vidal, P., Sabia, S., van der Spek, A., … Wood, A. R. (2019). Genetic studies of accelerometer-based sleep measures yield new insights into human sleep behaviour. *Nature Communications*, *10*(1), 1–12. https://doi.org/10.1038/s41467-019-09576-1

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

Lauderdale, D. S., Knutson, K. L., Yan, L. L., Liu, K., & Rathouz, P. J. (2008). Sleep duration: How well do self-reports reflect objective measures? The Cardia sleep study. *Epidemiology*, *19*(6), 838–845. https://doi.org/10.1097/EDE.0b013e318187a7b0

Loprinzi, P. D., & Cardinal, B. J. (2011). Association between objectively-measured physical activity and sleep, NHANES 2005–2006. *Mental Health and Physical Activity*, *4*(2), 65–69. https://doi.org/10.1016/j.mhpa.2011.08.001

Martin, J. L., & Hakim, A. D. (2011). Wrist actigraphy. *Chest*, *139*(6), 1514–1527. https://doi.org/10.1378/chest.10-1872

Migueles, J. H., Rowlands, A. V., Huber, F., Sabia, S., & van Hees, V. T. (2019). GGIR: A research community–driven open source R package for generating physical activity and sleep outcomes from multi-day raw accelerometer data. *Journal for the Measurement of Physical Behaviour*, *2*(3), 188–196. https://doi.org/10.1123/jmpb.2018-0063

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

Plekhanova, T., Rowlands, A. V., Yates, T., Hall, A., Brady, E. M., Davies, M., Khunti, K., & Edwardson, C. L. (2020). Equivalency of sleep estimates: Comparison of three research-grade accelerometers. *Journal for the Measurement of Physical Behaviour*, *3*(4), 294–303. https://doi.org/10.1123/jmpb.2019-0047

Quante, M., Kaplan, E. R., Cailler, M., Rueschman, M., Wang, R., Weng, J., Taveras, E. M., & Redline, S. (2018). Actigraphy-based sleep estimation in adolescents and adults: A comparison with polysomnography using two scoring algorithms. *Nature and Science of Sleep*, *10*, 13–20. https://doi.org/10.2147/NSS.S151085

Reutrakul, S., & Van Cauter, E. (2018). Sleep influences on obesity, insulin resistance, and risk of type 2 diabetes. *Metabolism*, *84*, 56–66. https://doi.org/10.1016/j.metabol.2018.02.010

Sadeh, A. (2015). III. Sleep assessment methods. *Monographs of the Society for Research in Child Development*, *80*(1), 33–48. https://doi.org/10.1111/mono.12143

Sadeh, A., Sharkey, M., & Carskadon, M. A. (1994). Activity-based sleep-wake identification: An empirical test of methodological issues. *Sleep*, *17*(3), 201–207. https://doi.org/10.1093/sleep/17.3.201

Scott, H., Lack, L., & Lovato, N. (2020). A systematic review of the accuracy of sleep wearable devices for estimating sleep onset. *Sleep Medicine Reviews*, *49*, 101227. https://doi.org/10.1016/j.smrv.2019.101227

Slater, J. A., Botsis, T., Walsh, J., King, S., Straker, L. M., … Eastwood, P. R. (2015). Assessing sleep using hip and wrist actigraphy. *Sleep and Biological Rhythms*, *13*(2), 172–180. https://doi.org/10.1111/sbr.12103

Smith, M., McCrae, C. S., Cheung, J., Martin, J. L., Harrod, C., Heald, J., & Carden, K. (2018). Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: An American Academy of sleep medicine systematic review, meta-analysis, and Grade Assessment. *Journal of Clinical Sleep Medicine*, *14*(7), 1209–1230. https://doi.org/10.5664/jcsm.7228

van Hees, V. T., Sabia, S., Anderson, K. N., Denton, S. J., Oliver, J., Catt, M., Abell, J. G., Kivimäki, M., Trenell, M. I., & Singh-Manoux, A. (2015). A novel, open access method to assess sleep duration using a wrist-worn accelerometer. *PLoS One*, *10*(11), e0142533. https://doi.org/10.1371/journal.pone.0142533

van Hees, V. T., Sabia, S., Jones, S. E., Wood, A. R., Anderson, K. N., Kivimaki, M., Frayling, T. M., Pack, A. I., Bucan, M., & Mazzotti, D. R. (2018). Estimating sleep parameters using an accelerometer without sleep diary. *Scientific Reports*, *8*(1), 12975. https://doi.org/10.1038/s41598-018-31266-z

Wendt, A., da Silva, I. C. M., Gonçalves, H., Menezes, A., Barros, F., & Wehrmeister, F. C. (2020). Short-term effect of physical activity on sleep health: A population-based study using accelerometry. *Journal of Sport and Health Science*, *00*, 1–9. https://doi.org/10.1016/j.jshs.2020.04.007

Zinkhan, M., Berger, K., Hense, S., Nagel, M., Obst, A., Koch, B., … Young, P. (2014). Agreement of different methods for assessing sleep characteristics: A comparison of two actigraphs, wrist and hip placement, and self-report with polysomnography. *Sleep Medicine*, *15*(9), 1107–1114. https://doi.org/10.1016/j.pcad.2014.10.005

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Plekhanova, T., Rowlands, A. V., Davies, M. J., Hall, A. P., Yates, T., & Edwardson, C. L. (2022). Validation of an automated sleep detection algorithm using data from multiple accelerometer brands. *Journal of Sleep Research*, e13760. https://doi.org/10.1111/jsr.13760