Protocol

Generation of a Free-Living Ground-Truth Validation Dataset for Wearable Measures of Physical Activity, Sedentary Behavior, Sleep, and Heart Rate in Adults (OxWEARS): Protocol for a Cross-Sectional Study

Benjamin D Maylor[1,2], MSc, PhD; Scott R Small[1], MSc, DPhil; Tatiana Plekhanova[1,2], MSc, PhD; Laura Brocklebank[1,2], MSc, PhD; Stefan van Duijvenboden[1,2], MSc, PhD; Rachel Sharman[3], PhD; Elizabeth A Hill[3,4], PhD; Fredrik Karpe[5,6], MBBS, PhD, CCST; Simon D Kyle[3], MA, PhD; Aiden Doherty[1,2], PhD

[1]Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom

[2]Li Ka Shing Centre for Health Information and Discovery, Big Data Institute, University of Oxford, Oxford, United Kingdom

[3]Nuffield Department of Clinical Neurosciences, Sir Jules Thorn Sleep and Circadian Neuroscience Institute (SCNi), University of Oxford, Oxford, United Kingdom

[4]School of Applied Sciences, University of the West of England, Bristol, United Kingdom

[5]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, United Kingdom

[6]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Trust, Oxford, United Kingdom

Corresponding Author:

Aiden Doherty, PhD
Nuffield Department of Population Health
University of Oxford
Old Road Campus
Oxford OX3 7LF
United Kingdom
Phone: 44 1865 617794
Email: aiden.doherty@ndph.ox.ac.uk

Abstract

Background: Wearable devices enable continuous measurement of physical activity, sedentary behavior, sleep, and heart rate under free-living conditions. However, most validation studies rely on small, homogeneous samples; are conducted under laboratory conditions; or lack gold standard ground-truth measurements, limiting the generalizability and accuracy of derived metrics. There is a pressing need for open-access, large-scale, free-living validation datasets that include multisensor data from diverse body locations and participant demographics to aid in model development.

Objective: The Oxford Wearable ECG, Activity, Circadian Rhythm, and Sleep Validation Study (OxWEARS) aims to (1) validate accelerometer-based measurement of physical behaviors across 5 body sites against annotated camera data; (2) validate measurements of sleep and sleep staging from 5 different body sites against polysomnography; (3) validate wrist-worn photoplethysmography heart rate measurements against chest-worn electrocardiogram; and (4) generate a comprehensive, annotated, and anonymized dataset for open-access research use.

Methods: This cross-sectional study will recruit approximately 160 adults (aged ≥40 years) stratified by age, sex, and BMI from the Oxford BioBank. Over 3 days and 4 nights, participants will wear sensors on the wrists, chest, hip, thigh, and ankle. Ground-truth measures will be obtained from a chest electrocardiogram patch for heart rate, a first-person camera for activity annotation, an ankle-worn accelerometer for step count, and at-home polysomnography for sleep. An under-mattress sensor will collect measures of sleep, respiration rate, and bedtime, and a subjective sleep diary will also be obtained. Signals from different wear locations will be compared against the ground truth using precision, recall, F_1-score, κ, and agreement metrics.

Results: Recruitment commenced in November 2024, with 15 participants enrolled by May 2025. Overall, 50% of eligible adults contacted were happy to consent to the study, with excellent compliance with the protocol observed to date. Data collection is ongoing and expected to conclude in 2026, with the final annotated dataset made publicly available as soon as possible thereafter.

Conclusions: The OxWEARS study will generate an openly accessible dataset containing more than 10,000 annotated hours from a stratified sample of adults. This will directly support scalable, generalizable human activity recognition efforts, while also enabling robust development and benchmarking of wearable-derived health metrics.

International Registered Report Identifier (IRRID): DERR1-10.2196/78779

JMIR Res Protoc 2025;14:e78779; doi: 10.2196/78779

Keywords: validation; wearables; machine learning; physical activity; sedentary behavior; sleep; heart rate

Introduction

The use of wearable devices for the assessment of physical activity, sedentary behavior, and sleep allows for the quantification of 24-hour human physical behaviors without the recall and social desirability biases associated with self-reported measures [1-3]. While these behaviors are commonly measured using commercial- and research-grade devices within epidemiological studies, the accuracy of the derived metrics against gold standard ground truth is currently uncertain. To date, most activity and sleep validation studies of wearables have been of poor quality—conducted in small, homogeneous cohorts; under laboratory conditions for a single day; and often without a gold standard reference [4,5]. Additionally, most studies focus on either sleep or physical activity, resulting in a lack of ground-truth reference measures for all behaviors across the 24-hour day.

While some accelerometer models have been trained using free-living validation data, significant room for improvement exists in model generalizability and specifically in improved performance in analyzing physical activity, sedentary behavior, and sleep. For example, some validation studies have collected wrist-based accelerometer data paired with first-person camera images for behavior classification [6,7], and these methods have subsequently been adopted in large epidemiological studies for health research [8,9]. However, these data were collected in a convenience sample over a single day without deference to age, sex, and BMI [10]. Additionally, more multisensor datasets are needed to compare data collected concurrently from different wear locations, as reflected in the existing literature. For example, large datasets exist with accelerometers worn on the wrist [8,11], hip [12], and thigh [13-15]; however, there is limited evidence regarding the harmonization of physical behavior phenotypes generated from different wear locations for subsequent statistical analyses. Furthermore, combining different signals from newer devices, such as photoplethysmography (PPG) and accelerometry, may improve the detection of sleep, physical activity, and sedentary behavior by providing additional data streams for model development. Finally, in most validation studies, raw data are not published, hampering external validation efforts and method development to improve performance [4,16].

Therefore, our overarching aim is to collect a free-living validation dataset, annotating wearable sensor data with gold standard, ground-truth labeled data, for application to large-scale accelerometer datasets with linked health records.

The main aims of this study are as follows:

1. To validate the measurement of physical activity and sedentary behavior from accelerometers worn on the wrist, chest, waist, hip, thigh, and ankle compared with ground-truth annotations from wearable camera data.
2. To validate the measurement of sleep and sleep staging from accelerometers worn on the wrist, chest, waist, hip, and thigh compared with gold standard polysomnography (PSG) sleep metrics.
3. To validate the measurement of heart rate and heart rate variability from wrist-worn PPG sensors against a ground-truth chest-worn electrocardiogram (ECG) patch.
4. To generate an anonymized dataset of annotated physical activity, sedentary behavior, sleep, and cardiac monitoring data for unrestricted use by the wider scientific community.

Methods

Ethical Considerations

Ethics approval was received from the University of Oxford Central University Research Ethics Committee on August 1, 2023 (R74559). Before the initiation of any study procedures, all participants were provided with an information sheet (Multimedia Appendix 1) and gave written informed consent using a digital signature platform compliant with ISO 27001 standards (E-Sign; E-Sign Ltd). Participants' privacy is of the utmost importance to us; therefore, no identifying information will be published, and individual data will be deidentified before any data release. Participants will receive no compensation for their participation but will be provided with a nondiagnostic report on their sleep and physical behaviors during the course of the monitoring period. Conforming to current ethical frameworks [17,18], written consent to be recorded by the video camera is not required from bystanders; however, participants will be instructed to obtain verbal permission from and provide an explanation of the device to family members, cohabitants, workplace managers or supervisors, or other people in settings where a reasonable expectation of privacy exists. This approach was considered reasonable. A script and information card will be provided for use if participants are questioned by members of the public. While we obtain consent, we will also ask participants to sign a privacy agreement for the recording of video for this study in public and private settings (Multimedia Appendix 2). Participants will be instructed on how to cover the camera or pause data collection at any time when they need privacy or feel uncomfortable or unsafe. Examples of this include using the bathroom (in public or private), using

public changing rooms, interacting with unrelated children, or working in contexts where intellectual property must be protected. In line with the ethical framework established by Kelly et al [17], participants can request the removal of any unwanted or sensitive footage potentially caught on camera. Privacy concerns related to the wearable camera data are of critical importance. Therefore, raw video footage will never be released, and any examples used (such as for annotator training) will be confined to the University of Oxford research team. All data will remain on University of Oxford servers.

Data Deidentification

In line with previous research, no wearable camera data will be shared outside of the University of Oxford [7]. Each participant will be assigned a unique study identifier, which will be used to label all associated raw data files. Because participants could potentially be deidentified based on combined age, sex, and BMI, we will report only the subgroup to which each participant belongs, as described earlier. Finally, the start dates of the raw data will be randomized, and time stamps will also be shifted by a small random amount so that they do not reflect the true dates and times of the data. Similar steps have been used previously to ensure deidentification of participants taking part in studies [7].

Recruitment

OxWEARS is a cross-sectional study and will recruit a target sample of 160 participants from the Oxford BioBank [19]. A sample size of 160 will provide 80% power to detect a difference of 0.05 in Cohen $\varkappa$ score (effect size=0.25, assuming an SD of 0.25 from previous studies [20]) between 3 age groups using a fixed effects one-way ANOVA test (α=.05). The Oxford BioBank is a population-based cohort of approximately 9000 participants aged 25 to 55 years recruited in Oxfordshire, United Kingdom, from 1999 onward. Inclusion and exclusion criteria for this study are presented in Textbox 1. Recruitment will target the creation of a final study cohort evenly balanced by sex, age (40-54 years, 55-64 years, and ≥65 years), and BMI (18.5-25 kg/m^2, 25-30 kg/m^2, and >30 kg/m^2). The distribution of invitations will be regularly rebalanced to achieve 13 to 14 participants in each subgroup by the end of the study.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Enrolled participant of the Oxford BioBank
- Healthy adults aged 40 years and older
- Able to ambulate (with or without a walking aid)
- Willing to wear multiple sensors over 3 days and 4 nights of data collection

Exclusion criteria

- Unable to speak or understand English
- Self-reported neurological conditions or diagnosed sleep impairment
- Self-reported clinicians, teachers, caregivers, or anyone working in environments where image capture would be inappropriate and who cannot commit to 3 consecutive days of image capture outside of these sensitive environments
- Unwilling to wear all the monitors according to the study protocol

Study Timeline

A study timeline is presented in Figure 1, outlining the overall data collection period. Data collection will be conducted in the free-living environment for 3 full days and 4 consecutive nights. Following receipt of consent, an in-person study setup visit is arranged, wherein 2 members of the research team will meet the participant at their home to set up and deploy the wearable sensors, camera, and under-mattress sleep analyzer for the data monitoring period, in addition to the PSG sleep assessment for the first night only. One night of PSG was chosen to balance participant and researcher burden with the inclusion of a gold standard sleep assessment. The distribution of ambulatory wearable sensors across the body is presented in Figure 2, with further description in the following methodological subsections.

Figure 1. Overview flowchart of study activities during the data collection window. Green activities are conducted by the OxWEARS research team, and blue activities are conducted or collected by the study participant. ECG: electrocardiography; PSG: polysomnography.

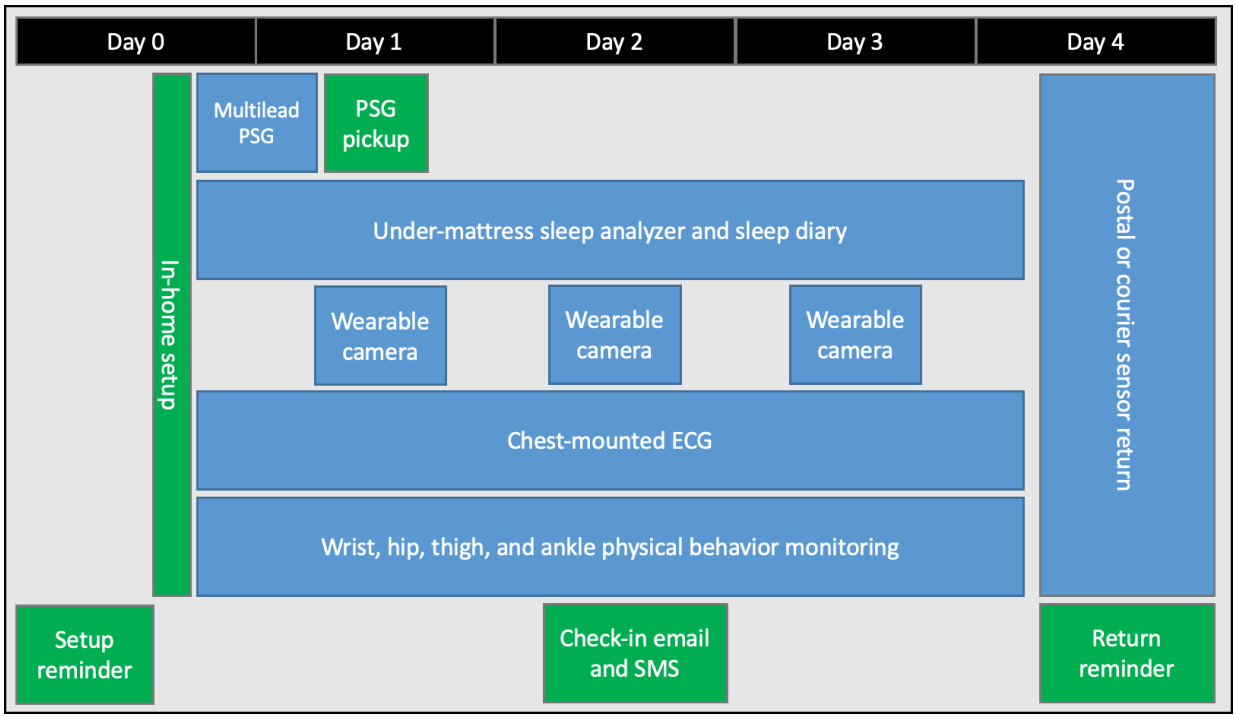
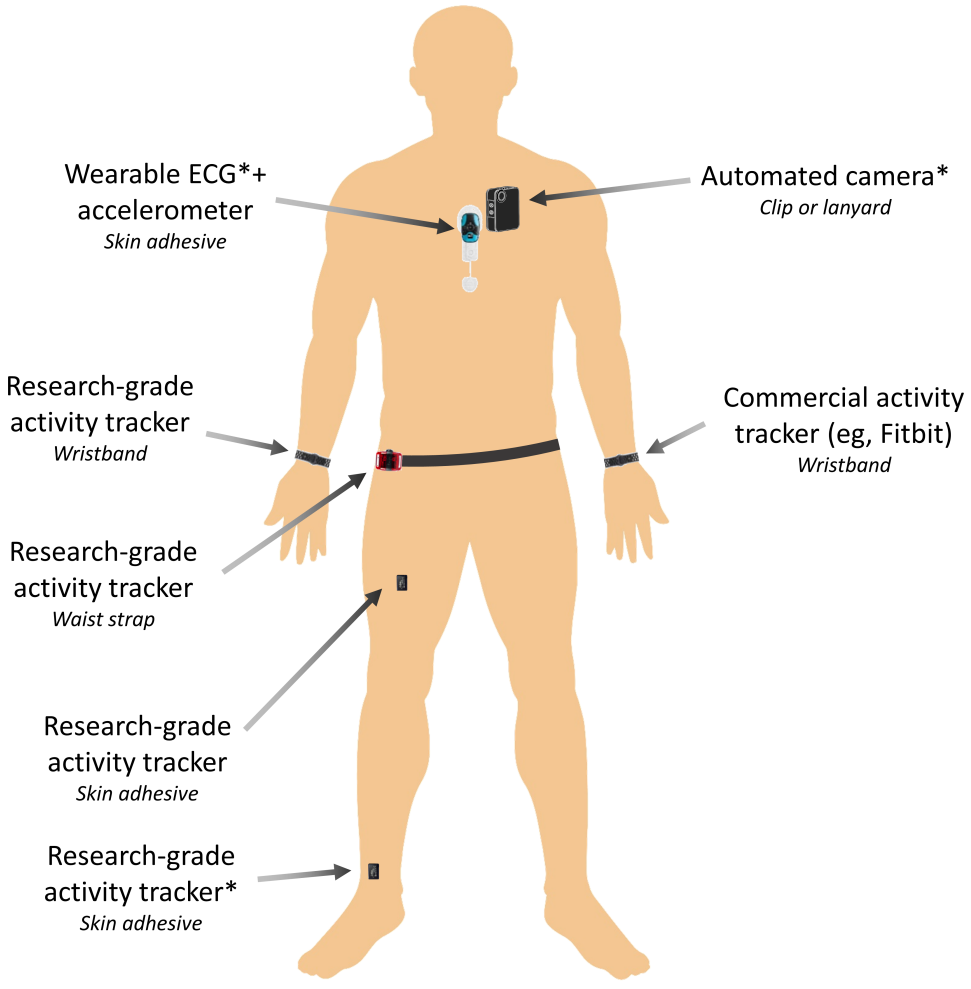


Figure 2. Distribution of wearable sensors for 3-day and 4-night ambulatory monitoring, including wrist, waist, thigh, and ankle accelerometers; a chest-mounted ECG and accelerometer patch; and an automated camera. *Denotes ground-truth or reference measurements (image designed by Freepik).



Ground-Truth ECG Data Collection

Ground-truth cardiac monitoring will be conducted using a clinical-grade single-lead wearable ECG patch (Bittium Faros 180; Bittium) affixed in a vertical orientation directly to the skin on the chest immediately inferior to the jugular notch and superficial to the sternum. Before adhesion, participants with chest hair will be instructed to shave the designated area to ensure adequate adhesion as per manufacturer guidelines. Continuous single-lead ECG data will be collected at 250 Hz and logged directly onto the ECG device for the entirety of the monitoring period (24 hours/day). In addition to raw ECG data, the monitor will be set to record continuous raw triaxial accelerometer data at the chest at a sampling rate of 100 Hz and with a dynamic range of ±4*g*.

Ground-Truth Physical Behavior Data Collection

To obtain ground-truth data on everyday activities, participants will wear a first-person perspective video camera (MIUFLY) during waking hours for 3 days, mirroring procedures from prior studies using first-person image capture to annotate physical activity and sedentary behavior [7,21]. In this study, a wearable camera is worn by the participant via a chest harness, neck mount, clip, or lapel and continuously records video during participant waking hours. Previous automated camera studies of this nature have reported wear time compliance of more than 80% [22,23]. No audio is captured by the camera, and all footage is encrypted on the device so that only the research team can download and view the video. Participants will be instructed on how to cover the camera or pause data collection at any time when they desire privacy or feel uncomfortable or unsafe. Examples of this include using the bathroom (public or private), public changing rooms, interacting with unrelated children, or working in contexts where intellectual property must be protected.

Ground-Truth Sleep Assessment

On the first study night, ground-truth sleep data will be collected with an ambulatory diagnostic multichannel PSG system (SOMNO HD Eco; S-Med Limited). The electroencephalogram (EEG) array will be attached in accordance with the 10-20 system of electrode placement [24]. The montage and recording settings (30-second epochs, 256 Hz sampling rate for EEG channels) will conform to the American Academy of Sleep Medicine (AASM) recommendations [25]. The following channels will be recorded: scalp EEG (F3, F4, C3, C4, O1, and O2), bilateral references on the mastoid processes (M1 and M2), grounding electrode (placed on FpZ), a common scalp reference electrode (Cz), bilateral electro-oculogram, 2-lead ECG, and 3-lead submental electromyography. Participants will be asked to change into sleeping clothes prior to sensor setup to minimize the risk of electrode disturbance. Prior to recording, researchers will check signal quality via a tablet and replace electrodes with high impedance values (>5 kΩ). The in-home sleep study will be conducted for 1 night only. Participants will be instructed to blink several times to indicate lights off at the time of their choosing. Participants will be instructed on how to remove the electrodes themselves following sleep, and a member of the research team will make a return visit the morning following the at-home visit to collect the PSG equipment and answer any further questions regarding the study.

Wearable-Based Physical Behavior Monitoring

Participants will be asked to wear 5 stand-alone wearable devices continuously (24 hours/day) throughout the course of the study period (Figure 2). The wear locations for the sensors were selected based on their popularity in previous validation studies [4,5]. The recording specifications of each device and sensor are detailed in Table 1. Wrist-measured physical behavior data will be collected using a combination of research-grade and consumer-grade devices. Specifically, a multimodal research-grade device (ActiGraph LEAP: ActiGraph LLC) consisting of triaxial accelerometer, gyroscope, and PPG sensors will be worn on the participant's dominant wrist. Similarly, a consumer-grade device (Fitbit Sense 2; Fitbit by Google) with comparable multimodal sensors to the ActiGraph LEAP will be worn on the participant's nondominant wrist, according to manufacturer's instructions.

Table 1. Wearable device data capture.

Device	Location	Sensors and recording settings	Heart rate
Bittium Faros 180	Chest	Accelerometer: 100 Hz, ±4*g*	1-channel ECG[a]: 250 Hz
ActiGraph LEAP	Dominant wrist	Accelerometer: 32 Hz, ±8*g*; gyroscope: 128 Hz and ±2000 dps[b]	PPG[c]: green (530 nm), 128 Hz
Fitbit Sense 2	Nondominant wrist	60 s epoch proprietary phenotypes	60 s epoch heart rate average
ActiGraph GT3X-BT	Waist or hip	Accelerometer: 100 Hz, ±8*g*	[d]—
Axivity AX6	Dominant thigh	Accelerometer: 100 Hz, ±8*g*; gyroscope: 100 Hz, ±2000 dps	—
Axivity AX6	Dominant ankle	Accelerometer: 100 Hz, ±8*g*; gyroscope: 100 Hz, ±2000 dps	—

[a]ECG: electrocardiogram.
[b]dps: degree per second.
[c]PPG: photoplethysmography.
[d]Not applicable.

Participants will wear an ActiGraph GT3X-BT monitor (ActiGraph LLC) on their hip continuously throughout the data collection period, removing it only for water-based activities. The device will be worn using an adjustable belt strap, with a buckle for easier removal.

For thigh monitoring, a medical-grade hypoallergenic adhesive dressing will be used to affix an accelerometer (Axivity AX6; Axivity Ltd) on the anterior aspect of the participant's dominant thigh, wrapped in a nitrile sleeve for protection. The Axivity AX6 houses a triaxial accelerometer and triaxial gyroscope, which will both record at a sampling frequency of 100 Hz and dynamic sensor ranges of ±8*g* and ±2000 degrees per second, respectively.

As an additional measurement point for physical activity, a further Axivity AX6 device (with matching recording characteristics) will be mounted on the dominant-side ankle, wrapped in a nitrile sleeve using medical adhesive tape. This is of specific interest for potential step count validation [26], for which a separate study including video-captured daily steps with concurrent ankle- and wrist-worn Axivity AX6 devices will be conducted to assess the validity of the ankle as a suitable ground-truth measure for step count in 50 adults, incorporating machine learning methods to estimate step count compared with proprietary algorithms assessed previously [26]. This model will then be applied to OxWEARS to compare step count from the ankle with each of the other wear locations (chest, both wrists, hip, and thigh).

Nearable-Based Sleep Assessment (Under-Mattress Sensor)

In addition to the ground-truth PSG sleep assessment, a popular consumer-grade sleep device (Withings Sleep Analyzer; Withings) will be set up by the researcher for the full 4 nights of the study. The Withings Sleep Analyzer will be placed under the participant's mattress in line with where the heart would be while the participant lies in bed. It uses ballistocardiography and a built-in microphone to estimate sleep duration, efficiency, sleep onset latency, sleep staging (awake, light, deep, or rapid eye movement sleep), as well as breathing rate and snoring events, using proprietary algorithms. It has shown some early promise as a potential "nearable" device for individuals to monitor long-term sleeping patterns without the need to physically wear a device [27].

Subjective Sleep Assessment

As a final complementary measure of sleep duration and quality, participants will additionally be asked to complete a standardized self-report sleep diary on each of the 4 study nights [28], which captures times the participant got into bed, went to sleep, woke up, and got out of bed. We have further adapted this to capture information on perceived sleep quality and whether any naps were taken outside of the primary sleep window (Multimedia Appendix 3).

Data Processing and Analysis

Physiological reference metrics derived from the ground-truth clinical-grade cardiac monitor will include participant heart rate and heart rate variability. There is ongoing work to develop and validate an ECG algorithm, which will be made publicly available in the future. During this process, clinical experts with extensive experience will review ECGs for RR intervals to determine algorithm performance. Derived heart rate and variability will be calculated on a per-epoch basis in 10-second windows and will serve as the ground-truth labels of heart rate for the study. Comparisons will then be made between these cardiac monitor–derived metrics and PPG-derived metrics [29] from both wrist-worn devices in terms of epoch-level accuracy and agreement across participant age, sex, and BMI subgroups.

Physical activity behaviors derived from the camera will take the form of video annotation of labeled activities based on posture (eg, lying, sitting, and standing) and whole-body movements (eg, walking, running, and cycling), in line with previous research [30]. Each participant's camera data will be annotated by an annotator who has completed a rigorous training regimen and is certified for annotation after consistently reaching a κ agreement of >0.8 against a series of day-long reference participants. If multiple annotators are required, we will additionally assess for interrater reliability scores. Further annotation and assessment will be conducted against machine learning–based automated image and video annotations [31]. These annotations will serve as the basis for retraining machine learning behavior classification models to better identify periods of sedentary, light, moderate, and vigorous physical activity [9,32].

Ground-truth sleep metrics, including time of sleep onset, waking time, and sleep staging, will be scored according to the AASM 2023 guidelines [25] by a single AASM-accredited sleep scientist with additional European Sleep Research Society accreditation and more than 15 years of experience working with EEG data. Sleep stages (non–rapid eye movement sleep [stages N1, N2, and N3], rapid eye movement, and awake) will be assigned for every epoch (30 seconds). A subsample of recordings (10%-20%) will be scored by a secondary accredited researcher to assess interrater reliability. As per recommendations in the field, scoring will be adjusted until interrater agreement reaches >80%. Similar to the physical activity behavioral ground-truth annotations, sleep annotations derived from the in-home PSG will serve as the basis for comparison of sleep time, sleep efficiency, and sleep staging derived from other accelerometers and the under-mattress sleep nearable, in addition to current accelerometer-based machine learning sleep models [33,34].

Data quality will be assessed regularly throughout the data collection period to check protocol adherence. This will involve processing the raw accelerometer data from the research-grade wearables through existing analytic software, such as Biobank AccelerometerAnalysis (University of Oxford) [8] and Stepcount (University of Oxford) [35], to assess device recording duration and nonwear detection

based on accelerometer thresholds. The Fitbit- and Withings-derived data will be checked at the summary level to flag possible poor compliance. PSG data quality will be assessed within the manufacturers' software using an automated sleep-stage classification system. We will report on the technical validation of the data when we publish the complete dataset, including the number of participants and the volume of camera-labeled data, both separately and when combined with concurrent data from other devices.

The primary aim of this study is to collect a high-quality validation dataset to facilitate new ways of analyzing time-series wearable data. As such, there will be many future imaginative analyses of these data that we do not anticipate at present. At present, we anticipate that mean precision, recall, F_1-score, Cohen $\varkappa$, and accuracy will be used to evaluate model performance for all comparisons with ground-truth labels. Additionally, summary metrics will be assessed against the ground-truth measure using mean absolute bias, mean amplitude percentage error, Cohen $\varkappa$ [36], and Bland-Altman plots [37].

Metadata Description

Our complete dataset and metadata will be hosted by the Oxford University Research Archive under the Creative Commons "Attribution 4.0 International (CC BY 4.0)" license. Raw accelerometry, ECG, and PPG data will be provided as compressed CSV files in folders for each participant. We will provide separate CSV files containing a dictionary of annotation labels for scored PSG data, a full annotation schema, labels for wearable camera data, and participant characteristics (age category, sex, and BMI category).

Results

To date, 150 participant information sheets have been sent to members of the Oxford BioBank database. Of these, 30 (20%) have expressed interest, and 15 (50% of those expressing interest) have provided consent. Common reasons for not progressing to consenting so far include ineligibility due to occupation (teacher or health care worker), requests to be contacted at a later date, and inability to make contact with the participant. The first participant consented and completed data collection in November 2024. As of May 2025, we had enrolled 15 participants, with 12 (80%) completing the monitoring period with excellent adherence to the protocol. Data collection is expected to be completed in 2026.

Discussion

This protocol details the design of a new free-living validation dataset to more accurately characterize physical behaviors and heart rate using wearable-based sensors on the wrist, chest, waist, hip, thigh, and ankle. By establishing ground-truth metrics for physical activity, sedentary behavior, sleep, and heart rate in a cohort of middle-aged and older adults, we attempt to create the largest resource for the validation of current and future methods for deriving physical behaviors and heart rate that is freely accessible to the wider research community. Furthermore, the methods that we implement in this study overlap with past research [7,38]; ongoing studies at the University of Oxford (S van Duijvenboden, unpublished data, December 2025); and future planned studies in low- and middle-income countries, such as India, South Africa, and Malaysia. This will foster efforts to improve the generalizability of machine learning models by training on diverse datasets, in addition to providing suitable datasets for external validation and pooled analyses. Large cohort studies currently hold rich datasets of unlabeled data. The creation of this curated dataset, comprising labeled behaviors for all popular sensor wear sites [39], will support the development of more accurate and novel phenotyping. The application of these phenotypes into epidemiological research will then be used to provide novel insights into human health, with respect to risk prediction, discovery of target mechanisms, and new methods to prioritize and assess the impact of potential treatments on day-to-day physical activity, sedentary behavior, and sleep.

Chest-worn ECGs are increasingly being used in cohort and clinical trials, where long-term monitoring of ECG data during free living can provide a more comprehensive understanding of heart function during daily activities, exercise, and sleep. The validation of ECG data from chest-worn wearables supports ongoing and future long-term cardiac monitoring trials such as the UK BioBank cardiac monitoring substudy [40]. At present, there is promising, but very limited, evidence on the validity of accelerometer-derived physical behavior phenotypes from chest-worn accelerometers based on simulated free living in small samples [41,42]. Incorporating physiological parameters, such as heart rate and other ECG parameters—key markers of cardiovascular health—could provide deeper insights into the intensity and impact of physical activity. For example, several activities performed at moderate-to-vigorous intensity, such as cycling, resistance exercises, or walking on an incline, are assessed with limited certainty with accelerometry [43] and are likely to be better captured by including heart rate monitoring, as it is not affected by the modality biases inherent to single-site accelerometry. Kuo et al [44] demonstrated an increase in reliability of 10% to 15% when including heart rate on top of accelerometry at 5 different wear sites to estimate energy expenditure during treadmill exercise at different speeds and gradients. This study was limited to a small sample of 16 young, healthy men and a treadmill protocol. Free-living behaviors captured in a larger, stratified sample, such as in this study, will further enhance our understanding of the added value of combining accelerometer data with heart rate, as these measures become increasingly available in research and consumer markets.

Identification of sleep stages that involve minimal movement may also benefit from additional physiological metrics to enhance their accuracy and reliability. This has been demonstrated previously in 31 adults who wore an Apple watch during a PSG assessment [45]. The combination of accelerometer with additional PPG sensor did not improve overall wake-sleep classification. However, sleep

staging classification accuracy improved by 15% to 25% when heart rate was added as a feature on top of accelerometry signal alone and showed similar performance when applied to a large heterogeneous sample of approximately 6800 US adults from the National Sleep Research Resource [46]. This study will generate the largest accessible free-living dataset with ground-truth annotations for model development to improve human activity recognition at the chest for immediate application to clinical trial and prospective cohort datasets.

Strengths of this study include the large, stratified sample to create a more heterogeneous, multimodal dataset; the inclusion of ground-truth measures of physical activity, sedentary behavior, sleep, and heart rate; and the planned release of the data to the wider research community. One limitation of this study is the lack of dietary or cognitive data, meaning certain behaviors cannot be linked to physiological metrics, for example, postmeal sedentariness and heart rate variability. Future work could explore the capture of these behaviors to provide additional context to physiological responses. Furthermore, despite recruiting a large sample size in comparison to existing device validation studies, subgroup analyses may be challenging when assessing algorithm performance. However, due to our methodology, it would be possible for these data to be combined with other datasets (eg, Capture-24) to increase the statistical power of these comparisons in the future. Finally, although we recruit a sample stratified by age, sex, and BMI, our sample does not include younger adults aged <40 years, those with chronic illnesses, or non-English speakers. Therefore, the results will only be applicable to healthy adults with a demographic profile similar to the recruited participants. Future work should strive to include representativeness of these characteristics in addition to improving representativeness from low- and middle-income countries to assess generalizability of machine learning algorithms trained on this dataset. Despite these limitations, we anticipate that the study's strengths will make the resulting dataset the largest and most comprehensive open-access validation dataset worldwide.

Acknowledgments

The authors express their gratitude to all current and future participants for their time in taking part in the OxWEARS study. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

Funding

This work is supported by the Wellcome Trust (grant 223100/Z/21/Z). Additional institutional funding was received from Swiss Re Institute. In-kind support for this project is being provided by Google United Kingdom. BDM and AD are supported by SwissRe. BDM and AD are further supported by the Wellcome Trust (grant 223100/Z/21/Z). AD and SRS are supported by NovoNordisk. AD's research team is supported by a range of grants from the Wellcome Trust (grants 223100/Z/21/Z and 227093/Z/23/Z), Novo Nordisk, Swiss Re, Boehringer Ingelheim, National Institutes of Health's Oxford Cambridge Scholars Program, Engineering and Physical Sciences Research CouncilCentre for Doctoral Training in Health Data Science (grant EP/S02428X/1), British Heart Foundation Centre of Research Excellence (grant RE/18/3/34214), and funding administered by the Danish National Research Foundation in support of the Pioneer Centre for Statistical and computational Methods for Advanced Research to Transform Biomedicine. SvD is supported by an Oxford British Heart FoundationCentre of Research Excellence Basic Science Intermediate Transition Fellowship. SDK reports current grant support from the Wellcome Trust (grants 227684/Z/22/Z and 227093/Z/23/Z), the National Institute for Health and Care Research (NIHR) (grants EME131789 and NIHR203667), and the Oxford Health NIHR Biomedical Research Centre (grants NIHR203316). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. The funding organizations had no role in design or conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. The views expressed are those of the authors and not necessarily those of the aforementioned funders.

Authors' Contributions

BDM and SRS drafted the original manuscript with support from LB. AD and SRS conceptualized the study. BDM, TP, SvD, and LB contributed to data curation. All authors contributed to the development of the methodology and analysis plan and provided feedback on manuscript drafts.

Conflicts of Interest

SRS is currently an employee of Novo Nordisk, but work was completed under association with the University of Oxford only. AD's team creates new methods based on wearable device validation dataset that are made available via academic-use software licenses. If used for commercial use, commercial entities pay his institution to purchase a license. In such circumstances, AD receives personal payment via the University of Oxford for sold software licenses.

Multimedia Appendix 1

Participant information sheet.
[DOCX File (Microsoft Word File), 1070 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Equipment and privacy agreement.
[DOCX File (Microsoft Word File), 32 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Sleep diary.
[DOCX File (Microsoft Word File), 32 KB-Multimedia Appendix 3]

References

1. Prince SA, Adamo KB, Hamel ME, Hardt J, Connor Gorber S, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. Int J Behav Nutr Phys Act. Nov 6, 2008;5(1):56. [doi: 10.1186/1479-5868-5-56] [Medline: 18990237]
2. Prince SA, Cardilli L, Reed JL, et al. A comparison of self-reported and device measured sedentary behaviour in adults: a systematic review and meta-analysis. Int J Behav Nutr Phys Act. Mar 4, 2020;17(1):31. [doi: 10.1186/s12966-020-00938-3] [Medline: 32131845]
3. Benz F, Riemann D, Domschke K, et al. How many hours do you sleep? A comparison of subjective and objective sleep duration measures in a sample of insomnia patients and good sleepers. J Sleep Res. Apr 2023;32(2):e13802. [doi: 10.1111/jsr.13802] [Medline: 36529876]
4. Giurgiu M, Timm I, Becker M, et al. Quality evaluation of free-living validation studies for the assessment of 24-hour physical behavior in adults via wearables: systematic review. JMIR Mhealth Uhealth. Jun 9, 2022;10(6):e36377. [doi: 10.2196/36377] [Medline: 35679106]
5. Giurgiu M, von Haaren-Mack B, Fiedler J, et al. The wearable landscape: issues pertaining to the validation of the measurement of 24-h physical activity, sedentary, and sleep behavior assessment. J Sport Health Sci. Dec 2025;14:101006. [doi: 10.1016/j.jshs.2024.101006] [Medline: 39491744]
6. Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. Sci Rep. May 21, 2018;8(1):7961. [doi: 10.1038/s41598-018-26174-1] [Medline: 29784928]
7. Chan S, Hang Y, Tong C, et al. CAPTURE-24: a large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. Sci Data. Oct 16, 2024;11(1):1135. [doi: 10.1038/s41597-024-03960-3] [Medline: 39414802]
8. 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(2):e0169649. [doi: 10.1371/journal.pone.0169649] [Medline: 28146576]
9. 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. Sep 6, 2021;56(18):1008-1017. [doi: 10.1136/bjsports-2021-104050] [Medline: 34489241]
10. Gershuny J, Harms T, Doherty A, et al. Testing self-report time-use diaries against objective instruments in real time. Sociol Methodol. Aug 2020;50(1):318-349. [doi: 10.1177/0081175019884591]
11. Leroux A, Cui E, Smirnova E, Muschelli J, Schrack JA, Crainiceanu CM. NHANES 2011-2014: objective physical activity is the strongest predictor of all-cause mortality. Med Sci Sports Exerc. Oct 1, 2024;56(10):1926-1934. [doi: 10.1249/MSS.0000000000003497] [Medline: 38949152]
12. Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc. Jan 2008;40(1):181-188. [doi: 10.1249/mss.0b013e31815a51b3] [Medline: 18091006]
13. Åsvold BO, Langhammer A, Rehn TA, et al. Cohort profile update: the HUNT study, Norway. Int J Epidemiol. Feb 8, 2023;52(1):e80-e91. [doi: 10.1093/ije/dyac095] [Medline: 35578897]
14. Gallibois M, Hennah C, Sénéchal M, Fuentes Diaz MF, Leadbetter B, Bouchard DR. Sedentary behaviour and fall-related injuries in aging adults: results from the Canadian Longitudinal Study on Aging (CLSA). JAR Life. 2024;13:93-98. [doi: 10.14283/jarlife.2024.14] [Medline: 39035110]
15. Sullivan A, Brown M, Hamer M, Ploubidis GB. Cohort profile update: the 1970 British Cohort Study (BCS70). Int J Epidemiol. Jun 6, 2023;52(3):e179-e186. [doi: 10.1093/ije/dyac148] [Medline: 35849349]
16. Keadle SK, Lyden KA, Strath SJ, Staudenmayer JW, Freedson PS. A framework to evaluate devices that assess physical behavior. Exerc Sport Sci Rev. Oct 2019;47(4):206-214. [doi: 10.1249/JES.0000000000000206] [Medline: 31524786]
17. Kelly P, Marshall SJ, Badland H, et al. An ethical framework for automated, wearable cameras in health behavior research. Am J Prev Med. Mar 2013;44(3):314-319. [doi: 10.1016/j.amepre.2012.11.006] [Medline: 23415131]

18. Behnke M, Saganowski S, Kunc D, Kazienko P. Ethical considerations and checklist for affective research with wearables. IEEE Trans Affective Comput. 2024;15(1):50-62. [doi: 10.1109/TAFFC.2022.3222524]
19. Karpe F, Vasan SK, Humphreys SM, et al. Cohort profile: the Oxford Biobank. Int J Epidemiol. Feb 1, 2018;47(1):21-21g. [doi: 10.1093/ije/dyx132] [Medline: 29040543]
20. Doherty A, Smith-Byrne K, Ferreira T, et al. GWAS identifies 14 loci for device-measured physical activity and sleep duration. Nat Commun. Dec 10, 2018;9(1):5257. [doi: 10.1038/s41467-018-07743-4] [Medline: 30531941]
21. Kelly P, Thomas E, Doherty A, et al. Developing a method to test the validity of 24 hour time use diaries using wearable cameras: a feasibility pilot. PLoS ONE. 2015;10(12):e0142198. [doi: 10.1371/journal.pone.0142198] [Medline: 26633807]
22. Harms T, Gershuny J, Doherty A, Thomas E, Milton K, Foster C. A validation study of the Eurostat harmonised European time use study (HETUS) diary using wearable technology. BMC Public Health. Jun 3, 2019;19(Suppl 2):455. [doi: 10.1186/s12889-019-6761-x] [Medline: 31159770]
23. Doherty AR, Kelly P, Kerr J, et al. Use of wearable cameras to assess population physical activity behaviours: an observational study. The Lancet. Nov 2012;380:S35. [doi: 10.1016/S0140-6736(13)60391-8]
24. Klem GH, Lüders HO, Jasper HH, Elger C. The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. Electroencephalogr Clin Neurophysiol Suppl. 1999;52:3-6. [Medline: 10590970]
25. Troester MM. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. American Academy of Sleep Medicine. 2023. URL: https://pulmo-ua.com/wp-content/uploads/2021/12/AASM-sleep-scoring-2017.pdf [Accessed 2025-05-05]
26. Toth LP, Park S, Springer CM, Feyerabend MD, Steeves JA, Bassett DR. Video-recorded validation of wearable step counters under free-living conditions. Med Sci Sports Exerc. Jun 2018;50(6):1315-1322. [doi: 10.1249/MSS.0000000000001569] [Medline: 29381649]
27. Manners J, Kemps E, Lechat B, Catcheside P, Eckert D, Scott H. Performance evaluation of an under-mattress sleep sensor versus polysomnography in ≥ 400 nights with healthy and unhealthy sleep. Health Informatics. 2024;34(6):e14480. [doi: 10.1101/2024.09.09.24312921] [Medline: 40017337]
28. Carney CE, Buysse DJ, Ancoli-Israel S, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep. Feb 1, 2012;35(2):287-302. [doi: 10.5665/sleep.1642] [Medline: 22294820]
29. van Gent P, Farah H, van Nes N, van Arem B. HeartPy: a novel heart rate algorithm for the analysis of noisy signals. Transp Res F Traffic Psychol Behav. Oct 2019;66:368-378. [doi: 10.1016/j.trf.2019.09.015]
30. Keadle SK, Martinez J, Strath SJ, et al. Evaluation of within- and between-site agreement for direct observation of physical behavior across four research groups. J Meas Phys Behav. 2023;6(3):176-184. [doi: 10.1123/jmpb.2022-0048]
31. Schönfeldt A, Maylor B, Chen X, Clark R, Doherty A. Reducing annotation burden in physical activity research using vision language models. Sci Rep. Oct 24, 2025;15(1):37253. [doi: 10.1038/s41598-025-21350-6] [Medline: 41136519]
32. Yuan H, Chan S, Creagh AP, et al. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. NPJ Digit Med. Apr 12, 2024;7(1):91. [doi: 10.1038/s41746-024-01062-3] [Medline: 38609437]
33. Yuan H, Hill EA, Kyle SD, Doherty A. A systematic review of the performance of actigraphy in measuring sleep stages. Health Informatics. 2023;33(5):e14143. [doi: 10.1101/2023.08.01.23293507] [Medline: 38384163]
34. 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. May 20, 2024;7(1):86. [doi: 10.1038/s41746-024-01065-0] [Medline: 38769347]
35. Small SR, Chan S, Walmsley R, et al. Self-supervised machine learning to characterize step counts from wrist-worn accelerometers in the UK Biobank. Med Sci Sports Exerc. Oct 1, 2024;56(10):1945-1953. [doi: 10.1249/MSS.0000000000003478] [Medline: 38768076]
36. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Erlbaum Associates; 1988. ISBN: 9780805802832
37. Martin Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. Feb 1986;327(8476):307-310. [doi: 10.1016/S0140-6736(86)90837-8]
38. Clark BK, Winkler EA, Brakenridge CL, Trost SG, Healy GN. Using Bluetooth proximity sensing to determine where office workers spend time at work. PLoS ONE. 2018;13(3):e0193971. [doi: 10.1371/journal.pone.0193971] [Medline: 29513754]
39. Pulsford RM, Brocklebank L, Fenton SA, 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. Mar 8, 2023;20(1):26. [doi: 10.1186/s12966-022-01388-9] [Medline: 36890553]

40. UK biobank cardiac monitoring project. UK Biobank Limited. URL: https://www.ukbiobank.ac.uk/taking-part/participant-opportunities/imaging-project/heart-monitor [Accessed 2025-04-11]
41. Camerlingo N, Cai X, Adamowicz L, et al. Measuring gait parameters from a single chest-worn accelerometer in healthy individuals: a validation study. Sci Rep. Jun 17, 2024;14(1):13897. [doi: 10.1038/s41598-024-62330-6] [Medline: 38886358]
42. Luckhurst J, Hughes C, Shelley B. Classifying physical activity levels using mean amplitude deviation in adults using a chest worn accelerometer: validation of the Vivalink ECG patch. BMC Sports Sci Med Rehabil. Oct 10, 2024;16(1):212. [doi: 10.1186/s13102-024-00991-6] [Medline: 39390591]
43. Liu F, Wanigatunga AA, Schrack JA. Assessment of physical activity in adults using wrist accelerometers. Epidemiol Rev. Jan 14, 2022;43(1):65-93. [doi: 10.1093/epirev/mxab004] [Medline: 34215874]
44. Kuo TB, Li JY, Chen CY, et al. Influence of accelerometer placement and/or heart rate on energy expenditure prediction during uphill exercise. J Mot Behav. 2018;50(2):127-133. [doi: 10.1080/00222895.2017.1306481] [Medline: 28850303]
45. Walch O, Huang Y, Forger D, Goldstein C. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. Sleep. Dec 24, 2019;42(12):zsz180. [doi: 10.1093/sleep/zsz180] [Medline: 31579900]
46. Zhang GQ, Cui L, Mueller R, et al. The National Sleep Research Resource: towards a sleep data commons. J Am Med Inform Assoc. Oct 1, 2018;25(10):1351-1358. [doi: 10.1093/jamia/ocy064] [Medline: 29860441]

Abbreviations

AASM: American Academy of Sleep Medicine
ECG: electrocardiogram
EEG: electroencephalogram
PSG: polysomnography

Edited by Javad Sarvestan; peer-reviewed by Andrea Caroppo, Annelinde Lettink; submitted 10.Jun.2025; final revised version received 01.Dec.2025; accepted 02.Dec.2025; published 29.Dec.2025

Please cite as:
Maylor BD, Small SR, Plekhanova T, Brocklebank L, van Duijvenboden S, Sharman R, Hill EA, Karpe F, Kyle SD, Doherty A
Generation of a Free-Living Ground-Truth Validation Dataset for Wearable Measures of Physical Activity, Sedentary Behavior, Sleep, and Heart Rate in Adults (OxWEARS): Protocol for a Cross-Sectional Study
JMIR Res Protoc 2025;14:e78779
URL: https://www.researchprotocols.org/2025/1/e78779
doi: 10.2196/78779