frontiers in Neuroscience

TYPE Original Research
PUBLISHED 17 February 2026
DOI 10.3389/fnins.2026.1693860

OPEN ACCESS

EDITED BY
Craig Canapari,
Yale University, United States

REVIEWED BY
Daniel M. Roberts,
The Pennsylvania State University (PSU), United States
Robert LeMoyne,
Northern Arizona University, United States

*CORRESPONDENCE
Shaun Davidson
✉ shaun.davidson@eng.ox.ac.uk

†These authors have contributed equally to this work

RECEIVED 27 August 2025
REVISED 06 January 2026
ACCEPTED 27 January 2026
PUBLISHED 17 February 2026

CITATION
Davidson S, Carter JF, Stanyer EC, Sharman R, Roman C, Kyle SD and Tarassenko L (2026) Longitudinal cardiorespiratory wearable sleep staging in the home.
Front. Neurosci. 20:1693860.
doi: 10.3389/fnins.2026.1693860

Longitudinal cardiorespiratory wearable sleep staging in the home

Shaun Davidson[1]*, Jonathan F. Carter[1†], Emily C. Stanyer[2†], Rachel Sharman[2], Cristian Roman[1], Simon D. Kyle[2] and Lionel Tarassenko[1]

[1]Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford, United Kingdom, [2]Sir Jules Thorn Sleep and Circadian Neuroscience Institute, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom

Introduction: There is a growing interest in performing automated, longitudinal tracking of sleep in the home using wearables and machine learning (ML). Wearables such as smart watches or chest patches can be comfortably worn for long periods, and cardiorespiratory waveforms measured by these wearables combined with ML models to estimate sleep state. However, the performance of these ML models is typically assessed using retrospective data from polysomnography, which is traditionally performed in a sleep lab. Importantly, polysomnography involves monitoring cardiorespiratory waveforms with bulky, specialized equipment, typically chest bands placed around the circumference of the upper toro and diaphragm, rather than with wearables. The performance of ML models evaluated on cardiorespiratory data from polysomnography may therefore not be representative of model performance when the cardiorespiratory input signals are acquired using modern wearable devices, where signal quality and available signal modalities may be more limited. Further work is needed to validate these ML models in their intended scenario of use: longitudinal, wearable sleep monitoring in the home.

Methods: This paper establishes and validates a pipeline for longitudinal cardiorespiratory sleep monitoring in the home using data from the RESTORE study. In RESTORE, 17 participants with a sleep-related condition (insomnia and depressive symptoms) underwent a sleep-related clinical intervention (sleep restriction therapy). Participants simultaneously wore a low-density home electroencephalogram device, allowing for expert, manual sleep staging using brain activity, as well as a wearable peel-and-stick chest patch, allowing for wearable monitoring of cardiorespiratory waveforms. Both devices were worn by participants for 10 nights while undergoing treatment at home. A state-of-the-art cardiorespiratory sleep staging model, combining transformer and convolutional neural networks, was then tuned and tested on the wearable data using leave-one-subject-out-cross-validation.

Results: After transfer learning, the cardiorespiratory sleep staging model had an accuracy of 77.1% and Cohen's Kappa of 0.679 for four-class sleep staging. Further, the model was able to accurately track sleep and sleep-derived metrics longitudinally while participants underwent sleep restriction therapy.

Discussion: These results represent one of the first direct demonstrations of the potential for wearable, cardiorespiratory sleep staging to track longitudinal, clinically relevant changes in sleep in individuals undergoing a sleep-related intervention in the home.

KEYWORDS
home monitoring, machine learning, sleep restriction therapy, sleep staging, validation study, wearable data

1 Introduction

There is a growing body of literature on the association between sleep and health. Chronic sleep deprivation is known to result in neural, molecular, and immune changes that contribute to cardiovascular disease development (Luyster et al., 2012). Further, the timing and relative amount of particular sleep states is known to be associated with susceptibility to, and progression of, a diverse range of conditions including dementia, diabetes, and depression (Zavecz et al., 2023; Tasali et al., 2008; Himali et al., 2023). The continuous monitoring of sleep patterns has the potential to provide early indication of disease progression and treatment response, as well as potential targets for intervention (Himali et al., 2023; Steiger and Pawlowski, 2019).

Gold-standard sleep monitoring is performed using polysomnography (PSG), which typically includes monitoring an individual's brain activity via electroencephalography (EEG), eye movement via electrooculography (EOG), muscle movement via electromyography (EMG), heart activity via electrocardiography (ECG), and breathing via Impedance Pneumography (IP) or nasal airflow in a dedicated sleep lab. Each 30 seconds of sleep, typically referred to as an "epoch," is then manually labeled by an expert scorer as wake, Rapid Eye Movement (REM), Non-REM 1 (N1), Non-REM 2 (N2), or Non-REM 3 (N3 or slow wave) sleep (Ferber et al., 1994). PSG is expensive and demanding of expert time for manual labeling of sleep stages (Silber et al., 2007). PSG also conventionally requires a subject to sleep in an unfamiliar lab environment while wearing obtrusive monitoring equipment, potentially disrupting a participant's sleep (sometimes referred to as the "first night effect") (Byun et al., 2019). As such, while sleep research using PSG remains the gold standard and has produced compelling findings in small, targeted cohorts (Himali et al., 2023; Steiger and Pawlowski, 2019), it is not a feasible approach for large-scale, longitudinal monitoring of sleep.

Recent advances in wearable technology and machine learning have the potential to address this issue (Chee et al., 2025). There are now a variety of wearable devices that can provide a subset of PSG or PSG-adjacent waveforms in a continuous and minimally disruptive fashion. These include wearable EEG-based devices such as the Dreem headband, which exhibit good performance but have trade-offs in terms of price, tolerability, and ease-of-use (Ong et al., 2023). A less disruptive approach to wearable sleep monitoring is to analyse cardiorespiratory waveforms acquired with wrist- or chest-worn wearables, which are typically unobtrusive and can be worn continuously for multiple days (Areia et al., 2020). While manual sleep staging is performed primarily using the EEG, EOG, and EMG, cardiorespiratory waveforms are associated with sleep behavior via the autonomic nervous system, and machine learning models can be trained to leverage this association for automated sleep staging (Davidson et al., 2023; Bakker et al., 2021). Although photoplethysmography (PPG) measured at the wrist is a sensing technology used extensively in commercial devices, an advantage of chest-worn wearables is that they can provide a direct measurement of not only cardiac but also respiratory signals, a combination that has been shown to improve sleep staging performance (Carter and Tarassenko, 2024).

A variety of machine learning models for cardiorespiratory sleep staging have been reported in the literature (Davidson et al., 2023; Bakker et al., 2021; Kotzen et al., 2022). These models are typically trained and evaluated on large, open access, retrospective PSG datasets such as those provided in the National Sleep Research Resource (NSRR) (Zhang et al., 2018). Given the relatively low incidence and low inter-rater agreement associated with N1 sleep, it is common for these cardiorespiratory sleep staging models to combine N1 and N2 sleep into a single "light sleep" class, enabling them to perform four-class sleep staging (i.e., wake, N1/N2, N3, REM) (Carter and Tarassenko, 2024; Kotzen et al., 2022). Recent advances incorporating the transformer architecture have demonstrated near expert-level sleep scoring using ECG and IP waveforms from sleep lab PSG datasets [Cohen's Kappa of 0.78 and accuracy of 84.8% in Carter and Tarassenko (2024)]. However, these methods have yet to be directly validated in their intended scenario of use—longitudinal, chest-worn wearable data in the home environment. This environment presents several potential challenges, including differences in signal quality between wearable and PSG-derived ECG, as well as the fact that most chest wearables do not directly provide an IP waveform.

Highlighting this challenge, a recent publication (Wang et al., 2025) investigated wearable sleep staging using the Physionet DREAMT dataset (Wang et al., 2025). This dataset contains expert scored PSG data alongside PPG data from a research grade wearable smart watch (Empatica E4) for 100 participants (1 night per participant). Wang et al. (2025) report a Cohen's Kappa of 0.553 after fine tuning/transfer learning for three-class sleep staging (i.e., wake, Non-REM, and REM, with no differentiation between N1/N2 and N3 sleep), outperforming previous state-of-the art models including SleepPPGNet (Kotzen et al., 2022) when fine tuned and tested on the same dataset. It is notable here that the Cohen's Kappa reported for SleepPPGNet after fine tuning on the DREAMT dataset (0.426) is significantly lower than that reported for retrospective PSG test data [0.74 in Kotzen et al. (2022)], underlining the challenges of translating such models from sleep lab PSG waveforms [in this case, PPG from a clinical grade enclosed finger probe, as in the NSRR datasets (Zhang et al., 2018)] into the home environment (in this case, PPG from a research grade wearable smart watch in DREAMT).

In this paper, we establish a pipeline for extracting a respiratory signal from a wearable chest patch accelerometer and then performing cardiorespiratory sleep staging using a state-of-the-art transformer model. We assess the performance of this pipeline against manual, expert sleep labeling using a low-density EEG/EOG/EMG device in the home environment. Our study cohort includes longitudinal monitoring of participants with a sleep-related condition [insomnia disorder and depressive symptoms (Steiger and Pawlowski, 2019)] undergoing a sleep-related intervention [sleep restriction therapy (Tse et al., 2024; Kyle et al., 2023)]. As such, this study allows for the assessment of the model's ability to track sleep in a scenario that closely resembles a potential use case for a wearable cardiorespiratory sleep staging system.

2 Materials and methods

2.1 The RESTORE study

Wearable cardiorespiratory data and expert sleep labels were drawn from the "examining the mechanisms of sleep RESTriction therapy fOR insomnia in people with dEpressive symptoms" (RESTORE) study (Stanyer et al., 2024). The primary outcome of the RESTORE study was the feasibility of serial home EEG monitoring in participants undergoing Sleep Restriction Therapy (SRT), as assessed by the proportion of EEG recordings that were scorable and yielded usable data, as well as the feasibility of the study for participants, and acceptability of the EEG equipment both qualitatively and quantitatively. A secondary aim was to investigate the physiological mechanisms underlying SRT's potential to ameliorate depressive symptoms (Tse et al., 2024; Kyle et al., 2023). The RESTORE study was conducted in Oxfordshire, UK and was approved by the Medical Sciences Interdivisional Research Ethics Committee (MS IDREC) at the University of Oxford (Approval reference: R91701/RE001). Informed consent was obtained from all participants prior to their involvement. Full details of the study protocol are available in Stanyer et al. (2024).

Participants were recruited through community advertising and assessed for eligibility via an online questionnaire and interview (Stanyer et al., 2024). Overall, 17 participants with insomnia disorder (Sleep Condition Indicator (SCI) $\leq$16), depressive symptoms (Hospital Anxiety and Depression Scale (HADS-D) $\geq$ 8), and who were not taking medication that could affect their sleep, were recruited. Once recruited, participants underwent a 6-week study protocol (shown in Figure 1), which included a 2-week baseline period and 4-week SRT period. Participants recorded daily sleep diaries and underwent daily affect assessments [e.g., Positive and Negative Affect Schedule (PANAS)] throughout the study period.

Each participant also underwent 10 nights of EEG sleep monitoring (1 adaptation night, 1 baseline night directly prior to beginning SRT, the 7 initial nights of SRT, and 1 night at the end of the SRT protocol). During these nights, participants wore a low-density EEG device [HomeSleepTest (HST) device, Somnomedics, similar to Fietze et al. (2015) and Coon et al. (2025)], which included a forehead (FPz) EEG electrode [with the output signal filtered between 0.3 and 35 Hz, according to AASM standards (American Academy of Sleep Medicine, 2023)], left and right EOG electrodes, and EMG/reference (M1/mastoid) electrode, each sampled at 256 Hz. Simultaneously, participants also wore an adhesive 7-day chest-patch (VitalPatch, VitalConnect), part of an in-house wearable remote monitoring system (Santos et al., 2021), which directly records continuous ECG and three-axis accelerometer measurements at 125 Hz. The chest patch was initially applied to the participant by the research team prior to the baseline recording session, with subsequent reapplications of the chest patch after 7 days of monitoring performed at home by the participant. The HST device was self-applied by the participant prior to each night of EEG sleep monitoring. All monitoring occurred in the participant's home environment.

At the conclusion of the RESTORE study, sleep stages for all non-adaptation nights (i.e., up to 9 nights per participant) were labeled by a blinded expert scorer (ECS) using data from the HST and DOMINO v3.0.0.8. Scoring followed the AASM guidelines version 3 (American Academy of Sleep Medicine, 2023) with consideration to the montage (e.g., single forehead FPz EEG and mastoid EMG).

2.2 Extracting a respiratory signal

While the chest patch directly records ECG sampled at 125 Hz and reports an estimated respiratory rate every 4 seconds, it does not output a respiratory waveform. Instead, we extract a respiratory waveform (ACC-RESP) from the three-axis accelerometer signal (which is sampled at 125 Hz) provided by the chest patch. To achieve this, we use a modified version of the recursive method described in Schipper et al. (2021). With this method, a respiratory signal is extracted by projecting the accelerometer signal onto the plane perpendicular to gravity and then using Principal Component Analysis (PCA) to identify the major axis of chest respiratory movement. This process effectively extracts a respiratory waveform, but emphasizes different components of respiratory chest wall movement (i.e., those perpendicular to gravity) depending on whether the individual being monitored is lying supine or in a lateral position. In contrast, a thoracic IP waveform, such as that typically available in PSG studies, employs an impedance band which detects overall chest wall movement, regardless of body position. To extract a respiratory waveform from the accelerometer data that more closely resembles a thoracic IP waveform and captures respiratory chest wall movement, regardless of whether an individual is lying supine, prone, or in a lateral position, we developed a modified approach which avoids assuming that the direction of chest wall respiratory movement is perpendicular to gravity. This approach was implemented as described in the Supplementary material.

2.3 wav2sleep model

For determining sleep states from the continuous ECG and ACC-RESP signals, we use the state-of-the-art deep learning architecture developed in Carter and Tarassenko (2024), which combines transformer and Convolutional Neural Network (CNN) components. This architecture has been shown to outperform existing cardiorespiratory sleep staging methods across a variety of signal input modalities, achieving a Cohen's Kappa of 0.78 and accuracy of 84.8% when using ECG and respiratory (thoracic IP) input waveforms. wav2sleep is designed to simultaneously classify sleep state for periods of up to 10 h (i.e., 1,200 epochs), and consists of three components:

- **Signal encoders**: These consist of a series of residual blocks (He et al., 2016) which extract features from the raw input signals (ECG and IP) using a configuration based on SleepPPGNet (Kotzen et al., 2022). Each block consists of three convolutional layers followed by a pooling operation, as shown in Figure 2. Separate encoders are used for each signal modality, facilitating different input sample rates (parameters shown in Table 1). Once features are extracted, they are

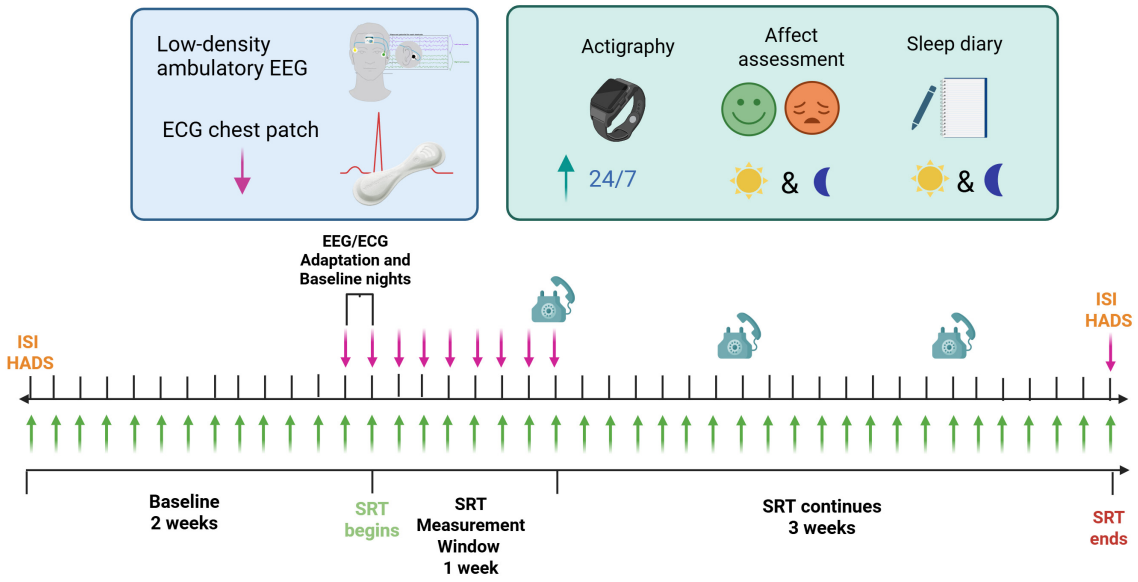

FIGURE 1
Summary of the RESTORE study protocol. ISI, Insomnia Severity Index; HADS, Hospital Anxiety and Depression Scale.

grouped per epoch by performing a reshape operation, and then fused into a reduced set of 128 features per epoch using a time-distributed dense operation (parameters in Table 1).

- **Epoch mixer**: This consists of two transformer encoder blocks (Vaswani et al., 2017) that fuse the features extracted by the signal encoders for each signal modality into a single representation. This encoder attends over the entire night's (i.e., 1,200 epochs) worth of extracted features, allowing for the modeling of long-range feature interactions and dependencies across the entire night of sleep using positional encoding. A simplified representation of an epoch mixer encoder block is shown in Figure 2 and supplemental parameters are provided in Table 1.
- **Sequence Mixer**: This is a dilated CNN that takes the unified representation output by the epoch mixer and generates sleep state labels for each epoch. A sequence mixer block is shown in Figure 2 and parameters are provided in Table 1. Dilated convolutions increase the receptive field of a CNN, allowing for long-range feature interactions and influences to be modeled. These long-range interactions are important to consider when labeling sleep state, where long-range factors such as time since the beginning of sleep or position within a sleep cycle have known influences on sleep behavior.

2.4 Transfer learning

Due to the domain shift between PSG and wearable ECG, and the ACC-RESP signal extracted from the chest patch accelerometer not directly corresponding to the thoracic IP signal in the labeled PSG data used to train wav2sleep, we perform transfer learning to enable the model to adapt to the changed signal modalities. As a starting point, we use the pre-trained wav2sleep model checkpoint provided by Carter and Tarassenko (2024), which was trained on over 10,000 expert-labeled overnight PSG recordings across 7 of the constituent databases in the NSRR (Zhang et al., 2018). Each participant in RESTORE was included in the analysis if good quality HST (i.e., containing any data considered scoreable by expert annotators) and chest patch data (i.e., no overnight gaps in data greater than 30 min in duration) were available for at least 75% of their study nights.

Given the relatively small number of participants in the RESTORE study, we implement Leave-One-Subject-Out-Cross-Validation (LOSOCV) when assessing model performance. For each participant in RESTORE, this entails making a given participant the test subject, splitting data from the other participants into training and validation cohorts, performing transfer learning, and then evaluating model performance on the test individual's data. In line with Carter and Tarassenko (2024), the model was trained using the AdamW optimiser and Cross Entropy loss with a batch size of 16. Transfer learning was performed using a fixed learning rate of 10^{-5}, with training stopped if the validation loss did not decrease for 5 consecutive epochs. The model weights were then reverted to those from the model state with the lowest validation loss for evaluation. Overall performance metrics are reported for the accuracy of sleep staging across all test subjects, with each participant serving as the test subject in turn.

Summary performance metrics (accuracy and Cohen's Kappa) were computed across all epochs in the dataset. Changes in performance metrics due to transfer learning were assessed using a paired, two-tailed Student's t-test. To account for the clustering of nights within participants, each participant contributed a single metric value for this t-test (i.e., averaged across all nights of monitoring for that participant). Differences in inter-participant deviation in model accuracy before and after transfer learning were assessed using a paired, two-tailed Student's t-test applied to

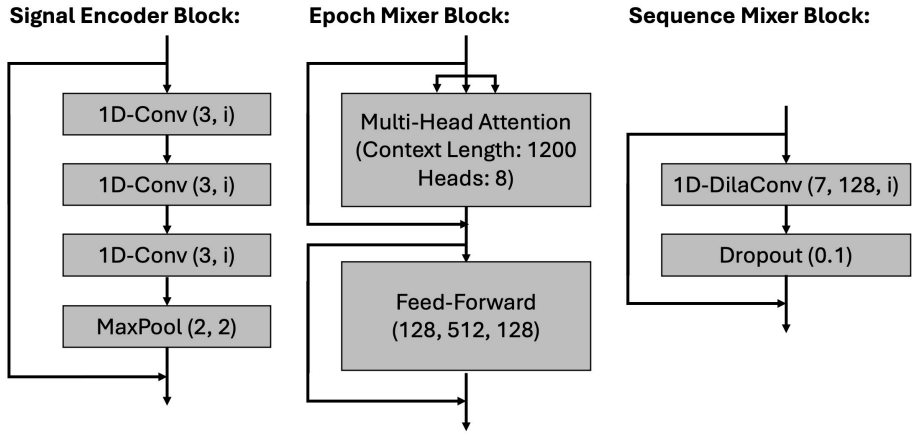

FIGURE 2
Example signal encoder, epoch mixer, and sequence mixer blocks. For the signal encoder, each 1D-Conv layer has a kernel size of 3 and number of features ("i") specified in Table 1. For the sequence mixer, each 1D-Dila(ted) Conv layer has a kernel size of 7, 128 features, and dilation factor specified in Table 1.

TABLE 1 wav2sleep model parameters and layer input/output dimensions from Carter and Tarassenko (2024).

Model component	Parameters	Output dim
Raw input, ECG	34.13 Hz	1,228,800, 1
Raw input, IP	8.53 Hz	307,200, 1
Signal encoder, ECG	i in [16, 16, 32, 32, 64, 64, 128, 128]	4,800, 256
Signal encoder, IP	i in [16, 32, 64, 64, 128, 128]	4,800, 256
Reshape (ECG)	Group features per epoch	1,200, 1,024
Reshape (IP)	Group features per epoch	1,200, 1,024
Time-distributed dense (ECG)	Dense (1,024 in, 128 out)	1,200, 128
Time-distributed dense (IP)	Dense (1,024 in, 128 out)	1,200, 128
Epoch mixer (ECG+IP)	2 Encoder blocks, dropout 0.1	1,200, 128
Sequence mixer	i in [1, 2, 4, 8, 16, 32] × 2 dilation	1,200, 128
Sleep state prediction	Dense (128 in, 4 out)	1,200, 4

Block architectures are shown in Figure 2.

TABLE 2 Demographics of the RESTORE cohort.

Demographic	Value
Age (years), mean (SD)	38.2 (13.9)
Age range	19–62
Female, *n* (%)	13 (76.5)
Ethnicity, *n* (%)	
White, British	12 (70.6)
White, other	1 (5.9)
Asian, any	0 (0.0)
Black, any	1 (5.9)
Mixed	2 (11.8)
Other	1 (5.9)
Baseline sleep, mean (SD)	
Total sleep time (mins)	414.9 (70.9)
Time in bed (mins)	461.2 (72.6)
Sleep efficiency (%)	89.8 (4.5)
Sleep onset latency (mins)	16.7 (14.1)
Insomnia & depression, mean (SD)	
ISI	16.8 (3.7)
HADS	10.0 (2.8)

the squared deviation between per-participant and cohort mean model accuracy.

3 Results

3.1 Study cohort

Overall, 17 participants were recruited for the RESTORE study, the demographics of whom are shown in Table 2. Participants were mostly female (76.5%) and of white British ethnicity (70.6%). Table 2 also summarizes cohort sleep metrics computed at baseline from the HST device, as well as baseline depression (HADS) and insomnia (ISI) scores.

Of the 17 participants recruited for the study, 13 had good quality HST and chest patch data available for at least 75% of their study nights, giving a total of 105 nights of data for analysis (with each participant contributing 7–9 nights of data). As shown in Table 3, this corresponds to 84,209 epochs sleep data labeled by a single expert sleep scorer, with an average of 6,478 labeled epochs per participant. Note that any epochs manually labeled as "artefact" by the expert scorer were excluded when computing performance

TABLE 3 Distribution of expert-labeled epochs in the RESTORE dataset.

Label	# (%) of epochs
Wake	9,973 (11.8)
N1/N2	26,626 (31.6)
N3	29,289 (34.8)
REM	16,402 (19.5)
Artifact	1,919 (2.3)
Total	84,209 (100.0)

metrics. To perform LOSOCV over each of the 13 participants, data from the remaining 12 participants was randomly split into training and validation sets, as described in Section 2.4. In each case, data from 10 participants was used for training and data from the other 2 participants for validation.

3.2 wav2sleep model performance—Without transfer learning

Figures 3A, B show confusion matrices summarizing the LOSOCV performance of the wav2sleep model for cardiorespiratory sleep staging without transfer learning on the 105 nights of data from 13 RESTORE participants included in this analysis. It is notable here that, even without transfer learning, both models show reasonably good agreement with manual expert sleep scoring (accuracy 68.0% and Cohen's Kappa 0.554 for the model using ECG+ACC-RESP as inputs, and accuracy 66.9% and Cohen's Kappa 0.542 for the model using only ECG as an input). This is despite considerable domain shift between the PSG data used to train the model and the wearable sensor data used in this study. Further, performance is still improved by adding our accelerometer-derived respiratory waveform despite substantial differences in data acquisition and pre-processing compared to a thoracic IP signal [as employed in the training data for the wav2sleep model (Carter and Tarassenko, 2024)]. Across both models, agreement between expert scoring and model-based cardiorespiratory sleep staging is generally good for wake (82.32%–85.45%), N1/N2 sleep (81.21%–86.61%), and REM sleep (77.25%–78.45%), but less so for N3 sleep (40.25%–41.75%).

3.3 wav2sleep model performance—With transfer learning

Figures 4A, B show confusion matrices summarizing the performance of the wav2sleep model for cardiorespiratory sleep staging after transfer learning is performed. Overall model performance is significantly improved ($p < 0.01$, Table 4) compared to before transfer learning (Accuracy from 66.9% to 75.3%, Cohen's Kappa from 0.542 to 0.655 for ECG only, and Accuracy from 68.0% to 77.1%, Cohen's Kappa from 0.554 to 0.679 for ECG+ACC-RESP). Once again, the model incorporating ECG+ACC-RESP as inputs outperforms the model using only ECG as an input. For both models, accuracy across classes is considerably more consistent compared to before transfer learning, ranging from 74.09%–83.00% for the ECG+ACC-RESP model and 71.20%–80.42% for the ECG only model. Table 4 also shows that model accuracy is significantly more consistent ($p < 0.01$) across participants after transfer learning: standard deviation in accuracy across participants decreases from 11.1% to 6.9% for ECG only, and from 11.3% to 5.9% for ECG+ACC-RESP with transfer learning.

Figure 5 shows an example comparison of model-derived hypnograms before and after transfer learning with manual expert EEG sleep staging. In Figure 5 it can be observed that the major shift the transfer learning process introduces is to increase the model detected incidence of N3 sleep (as compared to N1/N2 sleep) to better align with manual expert EEG labels.

Figure 6 shows an example of longitudinal trends in total sleep time, %N1/N2, %N3, and %REM sleep for a participant with good quality data for all 9 recorded nights during the RESTORE protocol. In Figure 6 it can be observed that model trends in total sleep time align well with expert labeling regardless of transfer learning, but that the transfer learning process notably improves the ability of the model to accurately track longitudinal trends in the relative incidence of different sleep states while participants are undergoing the RESTORE protocol. Figure 7 highlights that transfer learning improves the ability of the model to longitudinally track changes in N1/N2 sleep and N3 sleep across the entire RESTORE cohort.

4 Discussion

Prior to transfer learning, the transformer model exhibits reasonable agreement with manual, expert sleep staging using a low-density HST device, with an accuracy 66.9% and Cohen's Kappa of 0.542 (using only ECG as an input, Figure 3A) or an accuracy of 68.0% and Cohen's Kappa of 0.554 (using ECG and ACC-RESP as inputs, Figure 3B). The model exhibits very good performance for classifying wake, N1/N2, and REM sleep (accuracies of 77.25%–86.61%), but notably worse performance for classifying N3 sleep (accuracies of 40.25%–41.75%), with N3 sleep primarily misclassified as N1/N2 sleep (55.40%–58.64%). Both overall and per-class sleep staging accuracies are lower than those reported for the retrospective PSG test set in Carter and Tarassenko (2024), with a slight reduction in wake accuracy (90.78% to 83.32%–85.45%) and N1/N2 accuracy (88.96% to 81.21%–86.61%), a moderate reduction in REM accuracy (89.53% to 77.25%–78.45%), and a notable reduction N3 accuracy (72.24% to 40.25%–41.75%). The slight reduction in accuracy for wake and N1/N2 sleep is likely attributable to differences in signal quality between PSG and wearable cardiorespiratory sensors, as well as the shift from a respiratory waveform derived from a thoracic impedance band to one derived from a 3-axis accelerometer, without any transfer learning to account for this change in modality.

While these factors are also likely to contribute to the greater reductions in accuracy observed for REM and N3 sleep, there are other factors to consider regarding these two sleep states. With regards to the moderate decrease in REM accuracy compared to that reported in Carter and Tarassenko (2024), it is well established that ocular artifacts can contaminate frontal EEG signals, and

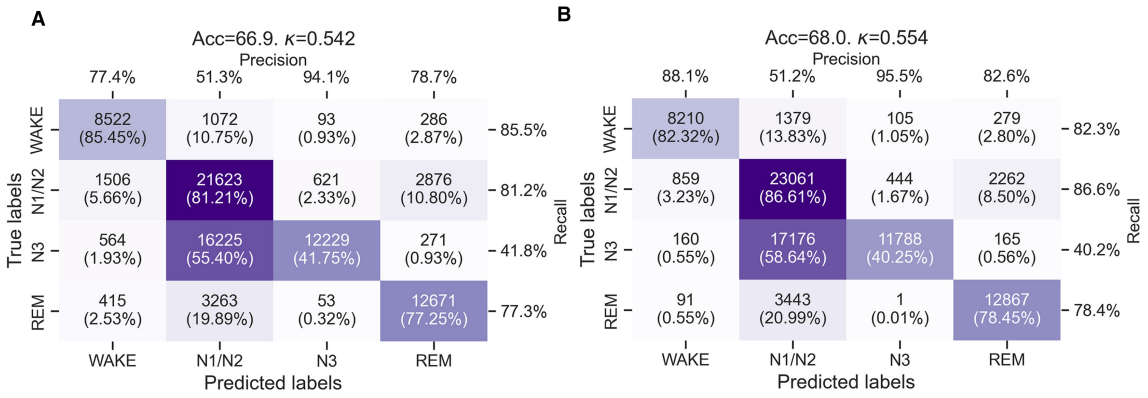

FIGURE 3
(A) Confusion matrices for the sleep staging performance of the wav2sleep transformer model without transfer learning using only ECG from the chest patch as an input. **(B)** Confusion matrices for the sleep staging performance of the wav2sleep transformer model without transfer learning using ECG and the accelerometer-derived respiratory waveform from the chest patch as inputs.

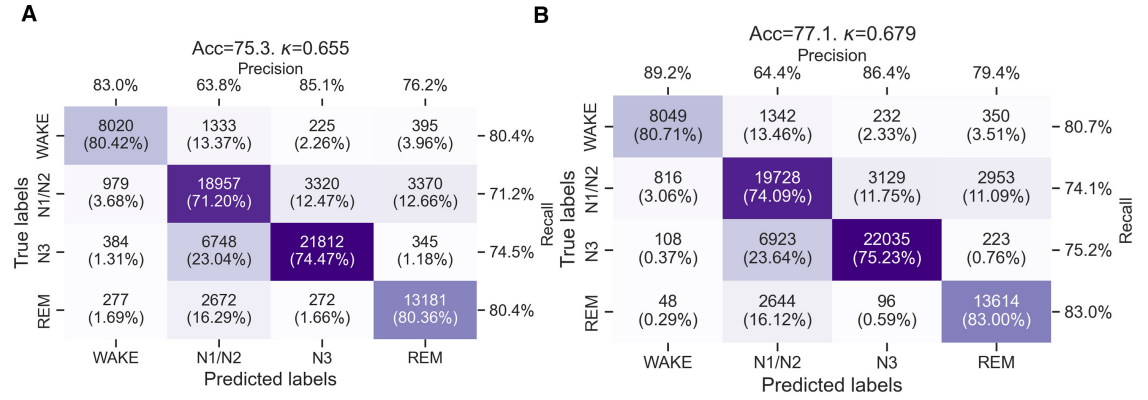

FIGURE 4
(A) Confusion matrices for the sleep staging performance of the wav2sleep transformer model with transfer learning using only ECG from the chest patch as an input. **(B)** Confusion matrices for the sleep staging performance of the wav2sleep transformer model with transfer learning using ECG and the accelerometer-derived respiratory waveform from the chest patch as inputs.

that this effect can be particularly significant for the FPz forehead electrode due to its proximity to the eyes (Lins et al., 1993). With regards to the notable drop in N3 accuracy compared to that reported in Carter and Tarassenko (2024), it is important to note that the NSRR databases used to train and test the transformer model label N3 sleep using the C4/A1 or C3/A2 central electrodes (the "acceptable" AASM electrode configuration), as opposed to the frontal F4/A1 or F3/A2 electrodes (the "ideal" AASM electrode configuration) (Zhang et al., 2018; American Academy of Sleep Medicine, 2023). It is additionally important to note that slow waves are known to exhibit greater amplitude in frontal EEG leads due to their association with the frontal regions of the brain (Carrier et al., 2011). Given that the FPz forehead electrode montage utilized by the HST device is even further forward than the "ideal" F4/A1 or F3/A2 electrodes, it is unsurprising that the application of the AASM 75 μV threshold [which has well-established limitations (Davidson et al., 2025)] by expert scorers to identify slow waves from this electrode montage leads to a relatively increased scoring of N3 sleep, especially when compared to a model trained using sleep stage labels from central electrodes. Notably, the incidence of EEG-labeled N3 sleep reported in this study cohort is also significantly greater than that reported in prior meta-analyses for individuals with insomnia and depression (Baglioni et al., 2014).

TABLE 4 Comparison of summary performance metrics before and after transfer learning.

Metric	Before TFL	After TFL	p
ECG model			
Accuracy, mean (SD)	66.9 (11.1)	75.0 (6.9)	5.1E-4
Cohen's Kappa, mean (SD)	0.547 (0.132)	0.651 (0.092)	2.2E-4
ECG+ACC-RESP model			
Accuracy, mean (SD)	68.0 (11.3)	76.8 (5.9)	8.8E-4
Cohen's Kappa, mean (SD)	0.559 (0.135)	0.674 (0.078)	4.3E-4

Transfer learning can be used to address differences both in input signal modality (i.e., the shift in respiratory signal from thoracic IP to ACC-RESP) or class labeling [i.e., the shift from N3 sleep labeled using the 75 μV threshold on a central (C4/A1)

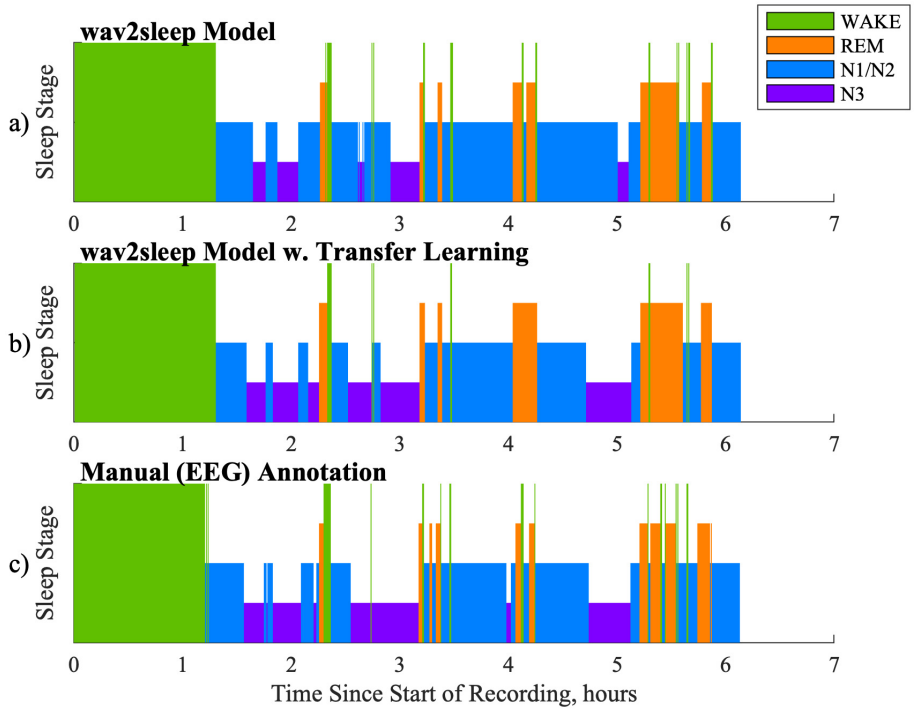

FIGURE 5
Example comparison of hypnograms between: **(a)** wav2sleep model before transfer learning; **(b)** wav2sleep model after transfer learning; **(c)** Manual Expert (EEG) annotation of sleep.

electrode to N3 sleep identified using a higher-amplitude forehead (FPz) electrode montage]. Figures 4A, B show the impact of transfer learning for models using ECG only and ECG+ACC-RESP and as inputs, respectively. Overall model performance is significantly improved ($p < 0.01$, Table 4), with accuracy increasing from 68.0% to 77.1% and Cohen's Kappa from 0.554 to 0.679 for the model using ECG+ACC-RESP as inputs, and accuracy increasing from 66.9% to 75.3% and Cohen's Kappa from 0.542 to 0.655 for the model using only ECG as an input. After transfer learning, the additional accuracy gained by including ACC-RESP as a model input increases from 1.1% to 1.8%, potentially reflecting the model adapting to the ACC-RESP waveform, having been trained on the thoracic IP waveform present in the NSRR datasets. Accuracy across sleep stages is considerably more consistent after transfer learning, ranging from 74.09% (for N1/N2 sleep) to 83.00% (for REM sleep) for the ECG+ACC-RESP model. Model accuracies for N3 and REM sleep increase at the expense of a slight decrease in accuracy for wake and a moderate decrease in accuracy for N1/N2 sleep. In particular, the accuracy for scoring N3 sleep increases from 40.25% to 75.23%, with N3 misclassification as N1/N2 decreasing from 58.64% to 23.64%.

When compared to the results in the retrospective PSG dataset in Carter and Tarassenko (2024), model accuracies for wake and REM sleep are similar, while model accuracy for N1/N2 sleep is notably lower [88.96% in Carter and Tarassenko (2024) compared to 74.09% in Figure 4B] and N3 sleep accuracy slightly greater [72.24% in Carter and Tarassenko (2024) compared to 75.23% in Figure 4B]. The likely explanation for this shift is the relatively greater incidence of N3 sleep in the RESTORE dataset (Table 3) due to scoring N3 sleep, with the accompanying trade-off in N1/N2 sleep scoring, using the frontal FPz electrode montage. This, in turn, results in N3 sleep having a relatively greater contribution to the cross-entropy loss used to train the model, resulting in the model placing an increased emphasis on identifying N3 sleep at the expense of N1/N2 sleep after transfer learning.

The example hypnograms in Figure 5 support these points, showing the model's ability to classify wake and REM accurately even prior to transfer learning [importantly capturing the short initial REM cycle at 2.25 h—the timing of which is relevant for the monitoring of conditions such as depression (Steiger and Pawlowski, 2019), as present in the RESTORE cohort]. The effect of transfer learning here is the increase in the relative incidence of model labeled N3 sleep as compared to N1/N2 sleep, better aligning hypnograms with expert sleep scoring using the HST device's FPz electrode.

The example longitudinal trends for a RESTORE participant in Figure 6 highlight the future potential of this work. In Figure 6, B1 reflects the baseline night, I1-I7 the intervention nights [i.e., Sleep Restriction Therapy (SRT)], and F1 the follow-up night. The wav2sleep model tracks total sleep time equally well before and after transfer learning, capturing the decrease in total sleep time when beginning SRT, followed up a gradual increase as the participant's sleep schedule adjusts during the first week of SRT, and finally an increase in total sleep time during follow-up after 4-weeks of SRT. After transfer learning, the model is similarly able to capture trends in N1/N2 sleep (slight initial decrease followed by gradual increase during intervention), N3 sleep (gradual decrease), and REM sleep (initial decrease, increase during intervention,

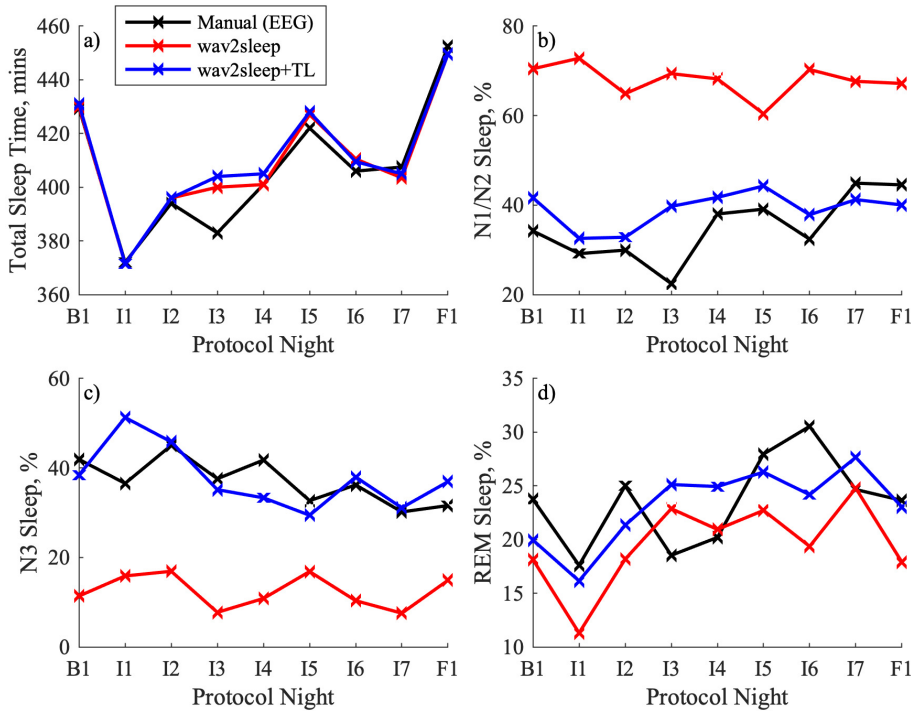

FIGURE 6
Example comparison of longitudinal trends for a RESTORE participant in: **(a)** Total sleep time; **(b)** %N1/N2 sleep; **(c)** %N3 Sleep; **(d)** %REM Sleep. B1 denotes baseline, I1-I7 denote intervention, and F1 denotes follow up nights.

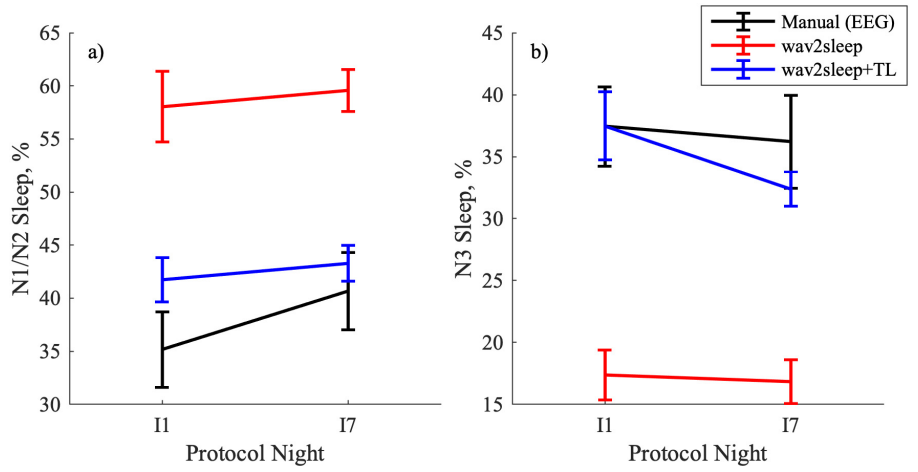

FIGURE 7
Comparison of sleep data from two nights across the RESTORE cohort for: **(a)** %N1/N2 sleep; **(b)** %N3 sleep. I1 and I7 denote the first and last nights of intervention, respectively. Error bars show standard error of the mean.

decrease during follow-up). These results are complemented by those in Figure 7, which show that transfer learning improves the ability of the model to track changes in N1/N2 and N3 sleep during SRT across the RESTORE cohort, as well as by those in Table 4, which show that transfer learning significantly reduces the variability in model performance across participants. Importantly, these results represent one of the first direct demonstrations of the potential for wearable, cardiorespiratory sleep staging to track longitudinal, clinically relevant changes in sleep in individuals undergoing a sleep intervention in the home, rather than in a sleep lab.

This study is, to our knowledge, the first to investigate longitudinal, wearable cardiorespiratory sleep staging in the home environment. This investigation is further strengthened by a comparison with manual, expert EEG sleep scoring while participants with a sleep-related condition (insomnia and depressive symptoms) undergo an intervention directly targeting sleep (sleep restriction therapy). As previously discussed, existing

literature on wearable cardiorespiratory sleep staging typically uses waveforms from retrospective, open access PSG datasets as a surrogate for wearable data (Bakker et al., 2021; Ebrahimi and Alizadeh, 2022). While this provides a valuable resource for initial model training and benchmarking, such PSG-derived cardiorespiratory data are not fully representative of wearable cardiorespiratory data gathered in the home, as highlighted by the reduction in performance of state-of-the-art models such as SleepPPGNet reported in Wang et al. (2025) when transitioned from PSG to wearable datasets.

A further important takeaway from the results in Wang et al. (2025) is the advantages of performing sleep staging using chest-based (as compared to wrist-based) wearable waveforms. PPG from wrist-worn wearables is a popular signal modality for sleep staging in commercial devices due to its low burden. However, the Cohen's Kappa of 0.553 for three-class sleep staging using smart watch PPG reported in Wang et al. (2025), outperforming existing state-of-the-art methods using the same input signal, is notably less than the Cohen's Kappa of 0.679 for four-class sleep staging reported here using the ECG and an accelerometer-derived respiratory waveform in a cohort of individuals undergoing a sleep intervention. Importantly, chest worn wearable waveforms contain significantly more respiratory information than wrist-worn PPG, both indirectly through respiratory modulation of the ECG and directly through the accelerometer-derived respiratory waveform. Four-class sleep staging has significant clinical advantages over three-class sleep staging as it allows the differentiation of N3 sleep, which has a variety of important clinical associations (Zavecz et al., 2023), from N1/N2 sleep.

4.1 Limitations

An important limitation of this study is the limited number of participants (13) and diversity (mostly white, female participants) in the RESTORE cohort. This limitation should be, however, be considered in the context of the longitudinal nature of the RESTORE study, which encompassed a total of 105 nights (82,290 epochs) of manually scored sleep epochs [i.e., similar in size to the 100 nights of data in the Physionet DREAMT dataset (Wang et al., 2025)]. Nevertheless, a future study incorporating a larger, more diverse cohort of participants would provide broader validation for the methodology developed in this manuscript.

A further limitation was the use of a limited montage, low-density EEG device (Somno-HST), as opposed to 19-lead PSG recommended by AASM scoring guidelines, for manual EEG sleep staging. The use of this low-density device was important in facilitating the study being performed in a longitudinal fashion on consecutive nights in the home environment, which would be difficult to accomplish using gold standard PSG due to the high cost and disruption associated with such an approach (Silber et al., 2007). It is worth noting that the low-density device does provide AASM standard left and right EOG, as well as single-lead EEG from the FPz electrode and EMG from the M1 electrode, and that sleep staging was performed manually by an expert sleep scorer. Further, a similar EEG/EOG/EMG device from Somnomedics was previously validated against gold standard PSG in Fietze et al. (2015), and general performance of home sleep monitoring devices using the FPz electrode in Coon et al. (2025). Regardless, direct validation of the cardiorespiratory sleep staging pipeline in a non-longitudinal study against gold-standard PSG would certainly complement the results presented in this paper.

With regards to the RESTORE study methodology, it is worth noting that no exclusion criteria were set for participants taking medication which might affect their cardiovascular system. Such medication could potentially affect the coupling between the autonomic nervous system and sleep states that cardiorespiratory sleep staging relies upon. Further, it is worth noting that participants self-applied the HST device and chest patch (after being initially shown how to apply this), which could lead to variable signal quality depending on placement. This self application does, however, mean that the dataset more closely resembles a true 'use case' scenario for this technology.

4.2 Conclusion

Overall, this paper develops a pipeline for performing wearable cardiorespiratory sleep staging in the home environment. This pipeline involves deriving a respiratory waveform from a chest worn accelerometer (as a surrogate for thoracic IP acquired from a chest band) and combining it with wearable ECG to perform cardiorespiratory sleep staging using a state-of-the-art deep learning model integrating transformer and CNN components. This pipeline is validated in a cohort of participants with a sleep related condition (insomnia and depressive symptoms) undergoing a sleep-related clinical intervention (sleep restriction therapy) in the home environment. This work thus represents one of the first validations, and an important step toward the realization, of this promising new technology in its intended use case - wearable, longitudinal sleep monitoring in the home environment.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, on reasonable request.

Ethics statement

The studies involving humans were approved by the Medical Sciences Interdivisional Research Ethics Committee (MS IDREC) at the University of Oxford (Approval reference: R91701/RE001). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

SD: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. JC: Formal analysis, Methodology, Software, Visualization, Writing – review & editing. ES: Conceptualization,

Data curation, Investigation, Methodology, Visualization, Writing – review & editing. RS: Conceptualization, Data curation, Investigation, Methodology, Writing – review & editing. CR: Investigation, Software, Writing – review & editing. SK: Conceptualization, Methodology, Supervision, Writing – review & editing. LT: Conceptualization, Methodology, Supervision, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This study was funded by the National Institute for Health and Care Research (NIHR) Oxford Biomedical Research Centre (BRC) (NIHR203316). SD and CR are supported by the NIHR Oxford BRC (BZR05002). ES, RS, and SK are supported by the NIHR Oxford Health BRC (NIHR203316). ES is supported by the 2023 British Sleep Society Colin Sullivan Award. SK is supported by the Wellcome Trust (226784/Z/22/Z) and the NIHR EME award (NIHR131789).

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Author disclaimer

The views expressed are those of the authors and not necessarily those of the NHS, NIHR, or the Department of Health and Social Care.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2026.1693860/full#supplementary-material

References

American Academy of Sleep Medicine (2023). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, version 3*. American Academy of Sleep Medicine.

Areia, C., Young, L., Vollam, S., Ede, J., Santos, M., Tarassenko, L., et al. (2020). Wearability testing of ambulatory vital sign monitoring devices: prospective observational cohort study. *JMIR mHealth uHealth* 8:e20214. doi: 10.2196/20214

Baglioni, C., Spiegelhalder, K., Feige, B., Nissen, C., Berger, M., and Riemann, D. (2014). Sleep, depression and insomnia-a vicious circle? *Curr. Psychiatry Rev.* 10, 202–213. doi: 10.2174/1573400510666140620001507

Bakker, J. P., Ross, M., Vasko, R., Cerny, A., Fonseca, P., Jasko, J., et al. (2021). Estimating sleep stages using cardiorespiratory signals: validation of a novel algorithm across a wide range of sleep-disordered breathing severity. *J. Clin. Sleep Med.* 17, 1343–1354. doi: 10.5664/jcsm.9192

Byun, J.-H., Kim, K. T., Moon, H.-j., Motamedi, G. K., and Cho, Y. W. (2019). The first night effect during polysomnography, and patients' estimates of sleep quality. *Psychiatry Res.* 274, 27–29. doi: 10.1016/j.psychres.2019.02.011

Carrier, J., Viens, I., Poirier, G., Robillard, R., Lafortune, M., Vandewalle, G., et al. (2011). Sleep slow wave changes during the middle years of life. *Eur. J. Neurosci.* 33, 758–766. doi: 10.1111/j.1460-9568.2010.07543.x

Carter, J. F., and Tarassenko, L. (2024). wav2sleep: a unified multi-modal approach to sleep stage classification from physiological signals. *arXiv preprint arXiv:2411.04644*.

Chee, M. W., Baumert, M., Scott, H., Cellini, N., Goldstein, C., Baron, K., et al. (2025). World sleep society recommendations for the use of wearable consumer health trackers that monitor sleep. *Sleep Med.* 131:106506. doi: 10.1016/j.sleep.2025.106506

Coon, W. G., Zerr, P., Milsap, G., Sikder, N., Smith, M., Dresler, M., et al. (2025). ezscore-f: A set of freely available, validated sleep stage classifiers for forehead EEG. *bioRxiv, 2025–06*. doi: 10.1101/2025.06.02.657451

Davidson, S., Roman, C., Carter, J., Harford, M., and Tarassenko, L. (2023). "Sleep staging using wearables and deep neural networks," in *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)* (IEEE), 1–4. doi: 10.1109/BHI58575.2023.10313497

Davidson, S., Sharman, R., Kyle, S. D., and Tarassenko, L. (2025). Is it time to revisit the scoring of slow wave (n3) sleep? *Sleep* 48:zsaf063. doi: 10.1093/sleep/zsaf063

Ebrahimi, F., and Alizadeh, I. (2022). Automatic sleep staging by cardiorespiratory signals: a systematic review. *Sleep Breath.* 26, 965–981. doi: 10.1007/s11325-021-02435-8

Ferber, R., Chair, R., Millman, R., Coppola, M., Fleetham, J., Murray, C. F., et al. (1994). Asda standards of practice: portable recording in the assessment of obstructive sleep apnea. *Sleep-Lawrence* 17:378. doi: 10.1093/sleep/17.4.378

Fietze, I., Penzel, T., Partinen, M., Sauter, J., Küchler, G., Suvoro, A., et al. (2015). Actigraphy combined with EEG compared to polysomnography in sleep apnea patients. *Physiol. Meas.* 36:385. doi: 10.1088/0967-3334/36/3/385

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. doi: 10.1109/CVPR.2016.90

Himali, J. J., Baril, A.-A., Cavuoto, M. G., Yiallourou, S., Wiedner, C. D., Himali, D., et al. (2023). Association between slow-wave sleep loss and incident dementia. *JAMA Neurol*. 80, 1326–1333. doi: 10.1001/jamaneurol.2023.3889

Kotzen, K., Charlton, P. H., Salabi, S., Amar, L., Landesberg, A., and Behar, J. A. (2022). Sleepppg-net: a deep learning algorithm for robust sleep staging from continuous photoplethysmography. *IEEE J. Biomed. Health Inform*. 27, 924–932. doi: 10.1109/JBHI.2022.3225363

Kyle, S. D., Siriwardena, A. N., Espie, C. A., Yang, Y., Petrou, S., Ogburn, E., et al. (2023). Clinical and cost-effectiveness of nurse-delivered sleep restriction therapy for insomnia in primary care (habit): a pragmatic, superiority, open-label, randomised controlled trial. *Lancet* 402, 975–987. doi: 10.1016/S0140-6736(23)00683-9

Lins, O. G., Picton, T. W., Berg, P., and Scherg, M. (1993). Ocular artifacts in EEG and event-related potentials i: Scalp topography. *Brain Topogr*. 6, 51–63. doi: 10.1007/BF01234127

Luyster, F. S., Strollo Jr, P. J., Zee, P. C., and Walsh, J. K. (2012). Sleep: a health imperative. *Sleep* 35, 727–734. doi: 10.5665/sleep.1846

Ong, J. L., Golkashani, H. A., Ghorbani, S., Wong, K. F., Chee, N. I., Willoughby, A. R., et al. (2023). A data driven approach for choosing a wearable sleep tracker. *medRxiv, 2023–10*. doi: 10.1101/2023.10.12.23296981

Santos, M. D., Roman, C., Pimentel, M. A., Vollam, S., Areia, C., Young, L., et al. (2021). A real-time wearable system for monitoring vital signs of covid-19 patients in a hospital setting. *Front. Digital Health* 3:630273. doi: 10.3389/fdgth.2021.630273

Schipper, F., van Sloun, R. J., Grassi, A., Derkx, R., Overeem, S., and Fonseca, P. (2021). Estimation of respiratory rate and effort from a chest-worn accelerometer using constrained and recursive principal component analysis. *Physiol. Meas.* 42:045004. doi: 10.1088/1361-6579/abf01f

Silber, M. H., Ancoli-Israel, S., Bonnet, M. H., Chokroverty, S., Grigg-Damberger, M. M., Hirshkowitz, M., et al. (2007). The visual scoring of sleep in adults. *J. Clin. Sleep Med*. 3, 121–131. doi: 10.5664/jcsm.26814

Stanyer, E. C., Sharman, R., Davidson, S., Ribeiro Pereira, S. I., Espie, C., Kyle, S., et al. (2024). Examining the mechanisms of sleep RESTriction therapy fOR insomnia in people with dEpressive symptoms (The RESTORE Study). *OSF*. doi: 10.17605/OSF.IO/3MEUZ

Steiger, A., and Pawlowski, M. (2019). Depression and sleep. *Int. J. Mol. Sci*. 20:607. doi: 10.3390/ijms20030607

Tasali, E., Leproult, R., Ehrmann, D. A., and Van Cauter, E. (2008). Slow-wave sleep and the risk of type 2 diabetes in humans. *Proc. Nat. Acad. Sci*. 105, 1044–1049. doi: 10.1073/pnas.0706446105

Tse, K. Y. K., Maurer, L. F., Espie, C. A., and Kyle, S. D. (2024). The effect of single-component sleep restriction therapy on depressive symptoms: a systematic review and meta-analysis. *J. Sleep Res*. 33:e14180. doi: 10.1111/jsr.14180

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*, 30.

Wang, K., Yang, J., Shetty, A., and Dunn, J. (2025). DREAMT: dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology (version 2.1.0). *PhysioNet*. doi: 10.13026/7r9r-7r24

Wang, W. K., Chen, B., Yang, J., Jeong, H., Hershkovich, L., Islam, S. M. M., et al. (2025). Watchsleepnet: a novel model and pretraining approach for advancing sleep staging with smartwatches. *Proc. Mach. Learn. Res*. 287, 1–20. doi: 10.1088/3049-4761/addac6

Zavecz, Z., Shah, V. D., Murillo, O. G., Vallat, R., Mander, B. A., Winer, J. R., et al. (2023). Nrem sleep as a novel protective cognitive reserve factor in the face of Alzheimer's disease pathology. *BMC Med*. 21, 1–12. doi: 10.1186/s12916-023-02811-z

Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., et al. (2018). The national sleep research resource: towards a sleep data commons. *J. Am. Med. Inform. Assoc*. 25, 1351–1358. doi: 10.1093/jamia/ocy064