

RESEARCH

Open Access



Automatic sleep stage classification based on subcutaneous EEG in patients with epilepsy

Sirin W. Gangstad^{1,2}, Kaare B. Mikkelsen³, Preben Kidmose³, Yousef R. Tabar³, Sigge Weisdorf⁴, Maja H. Lauritzen⁴, Martin C. Hemmsen², Lars K. Hansen¹, Troels W. Kjaer^{4*}  and Jonas Duun-Henriksen^{2,5}

*Correspondence:

twk@regionsjaelland.dk

⁴ Center of Neurophysiology,
Department of Neurology,
Zealand University

Hospital, Vestermarksvej 11,
4000 Roskilde, Denmark

Full list of author information
is available at the end of the
article

Abstract

Background : The interplay between sleep structure and seizure probability has previously been studied using electroencephalography (EEG). Combining sleep assessment and detection of epileptic activity in ultralong-term EEG could potentially optimize seizure treatment and sleep quality of patients with epilepsy. However, the current gold standard polysomnography (PSG) limits sleep recording to a few nights. A novel subcutaneous device was developed to record ultralong-term EEG, and has been shown to measure events of clinical relevance for patients with epilepsy. We investigated whether subcutaneous EEG recordings can also be used to automatically assess the sleep architecture of epilepsy patients.

Method : Four adult inpatients with probable or definite temporal lobe epilepsy were monitored simultaneously with long-term video scalp EEG (LTV EEG) and subcutaneous EEG. In total, 11 nights with concurrent recordings were obtained. The sleep EEG in the two modalities was scored independently by a trained expert according to the American Academy of Sleep Medicine (AASM) rules. By using the sleep stage labels from the LTV EEG as ground truth, an automatic sleep stage classifier based on 30 descriptive features computed from the subcutaneous EEG was trained and tested.

Results : An average Cohen's kappa of $\kappa = 0.78 \pm 0.02$ was achieved using patient specific leave-one-night-out cross validation. When merging all sleep stages into a single class and thereby evaluating an awake–sleep classifier, we achieved a sensitivity of 94.8% and a specificity of 96.6%. Compared to manually labeled video-EEG, the model underestimated total sleep time and sleep efficiency by 8.6 and 1.8 min, respectively, and overestimated wakefulness after sleep onset by 13.6 min.

Conclusion : This proof-of-concept study shows that it is possible to automatically sleep score patients with epilepsy based on two-channel subcutaneous EEG. The results are comparable with the methods currently used in clinical practice. In contrast to comparable studies with wearable EEG devices, several nights were recorded per patient, allowing for the training of patient specific algorithms that can account for the individual brain dynamics of each patient. Clinical trial registered at ClinicalTrial.gov on 19 October 2016 (ID:NCT02946151).

Keywords: Subcutaneous EEG, Wearable EEG, Automatic sleep scoring, Sleep, Epilepsy

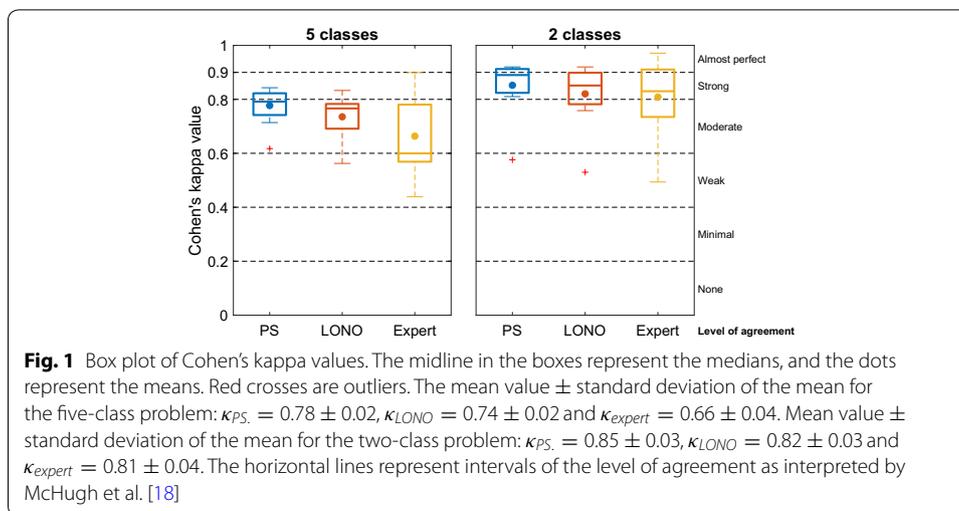


Background

The polysomnography (PSG) is the gold standard to assess sleep stages and other clinically relevant sleep parameters. However, it is resource demanding, impractical for the patient and may in itself have a negative impact on the sleep due to the obtrusive nature of the equipment. This method often limits the sleep assessment to a few days. To gain an objective measurement of sleep patterns over longer periods of time, a variety of wearable sleep trackers have emerged in the recent years. Activity-based devices monitor movements to infer information about sleep–wake patterns, and is currently the modality of choice for long-term sleep monitoring [1, 2]. A review by Sadeh et al. [3] concludes that actigraphy is reliable in individuals with normal sleep patterns. However, the authors question the validity in patients with sleep disorders, poor sleep and certain special populations such as very young children or the elderly. One of the major limitations of the actigraphy that is highlighted is the low specificity reported in several studies (the ability to recognize wakefulness, which affects estimates of, for example, wakefulness after sleep onset and sleep efficiency). In patients with epilepsy, seizures can produce movement patterns that can affect the actigraphy scoring. Sadaka et al. [4] compared actigraphy with continuous video-EEG over a 24-h period in 27 children with medically refractory epilepsy. The authors found that actigraphy reliably estimated commonly used sleep measures except number of wakings after sleep onset. They conclude that actigraphy can be used as a reliable tool for evaluating sleep patterns in children with epilepsy, but as reported in other studies, detecting wake periods after sleep onset remains a challenge.

As the American Academy of Sleep Medicine (AASM) manual mainly differentiates between the stages of sleep based on EEG features, the EEG is an essential tool in sleep monitoring. With the availability of publicly open EEG databases for benchmarking, several studies have developed algorithms for automatic sleep stage scoring of scalp EEG [5–9]. For a review of state-of-the-art feature extraction and classification techniques, see [10]. Other studies have sleep scored EEG from wearable devices such as ear plugs [11, 12], around-the-ear electrode arrays [13], head bands [14] and disposable forehead electrode arrays [15]. Most of these studies report promising results. However, some of the wearables are more suitable for ultralong-term recordings than others.

An emerging EEG modality is subcutaneous EEG. In contrast to currently available wearable EEG solutions, the electrodes are situated in a protected position underneath the skin and can provide continuous measurements with consistent location and impedance for months. Subcutaneous EEG solutions are therefore well suited for ultralong-term monitoring, meaning continuous recordings for > 2 weeks. If wearing a recording device in everyday life is well tolerated by the user, it is believed that long-term monitoring could provide a great help in treatment optimization and alarming of caregivers of patients with epilepsy [16]. It has already been shown that recordings from subcutaneous channels were comparable to those of scalp channels at similar locations [2, 17], and that subcutaneous EEG could be used to detect clinically relevant events in epilepsy patients [2]. The current study shows that subcutaneous recordings can be used to sleep score the same patient population to produce clinically relevant sleep measures. Subcutaneous EEG and LTV EEG from four inpatients were independently sleep scored by a trained expert. By using the sleep stage labels from the LTV EEG as ground truth, an



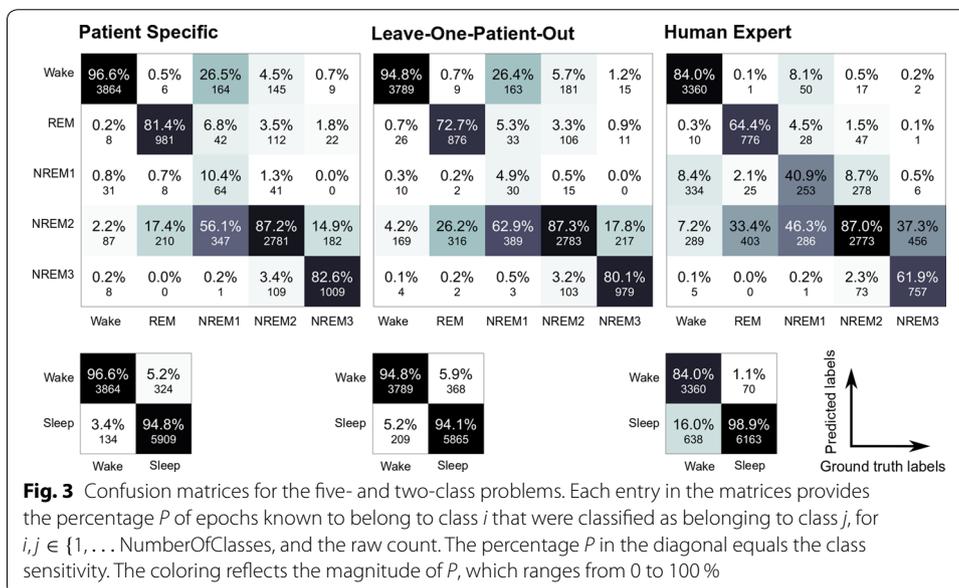
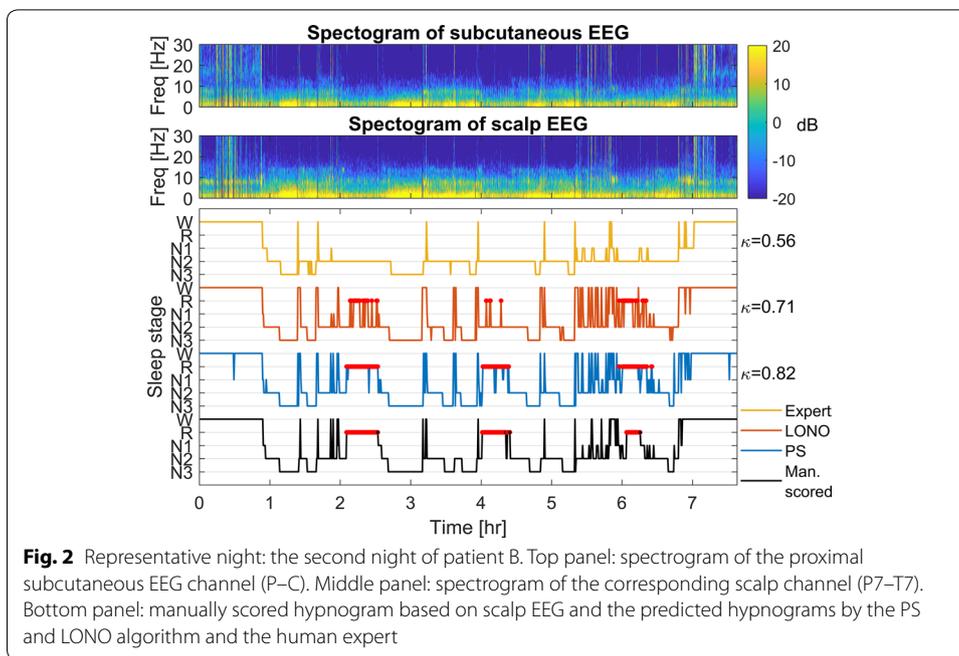
automatic sleep stage classifier based on the subcutaneous EEG was trained and tested. The algorithm was cross-validated (CV) using two strategies: a patient-specific (PS) approach and a leave-one-night-out (LONO) approach. In addition to evaluating the algorithms and the human expert on the five-class sleep staging problem, the hypnograms were converted to sleep–wake traces to create a simpler two-class classification task. Furthermore, some common sleep measures computed from the ground truth hypnograms and the predicted hypnograms were compared.

Results

Sleep stage classification

The Cohen's kappa values for the algorithms and the human expert when scoring five and two classes are plotted in Fig. 1 and tabulated in Appendix C. The CV strategy producing the best average Cohen's kappa value across nights was the PS approach. The mean kappa value \pm standard deviation of the mean is $\kappa_{PS} = 0.78 \pm 0.02$, while the mean kappa value for the LONO approach is $\kappa_{LONO} = 0.74 \pm 0.02$. For comparison, the mean kappa value for the human expert evaluating the subcutaneous EEG is $\kappa_{expert} = 0.66 \pm 0.04$. An exact paired permutation test revealed that the mean kappa values for both algorithms were significantly higher than for the human expert ($p_{PS} = 0.0016$, $p_{LONO} = 0.015$). For an illustration of a representative night showing the manually labeled and predicted hypnograms, see Fig. 2. For the two-class problem, the PS models had a mean kappa value of $\kappa_{PS} = 0.85 \pm 0.03$, the LONO-approach had a kappa of $\kappa_{LONO} = 0.82 \pm 0.03$, and the human expert had a kappa of $\kappa_{expert} = 0.81 \pm 0.04$.

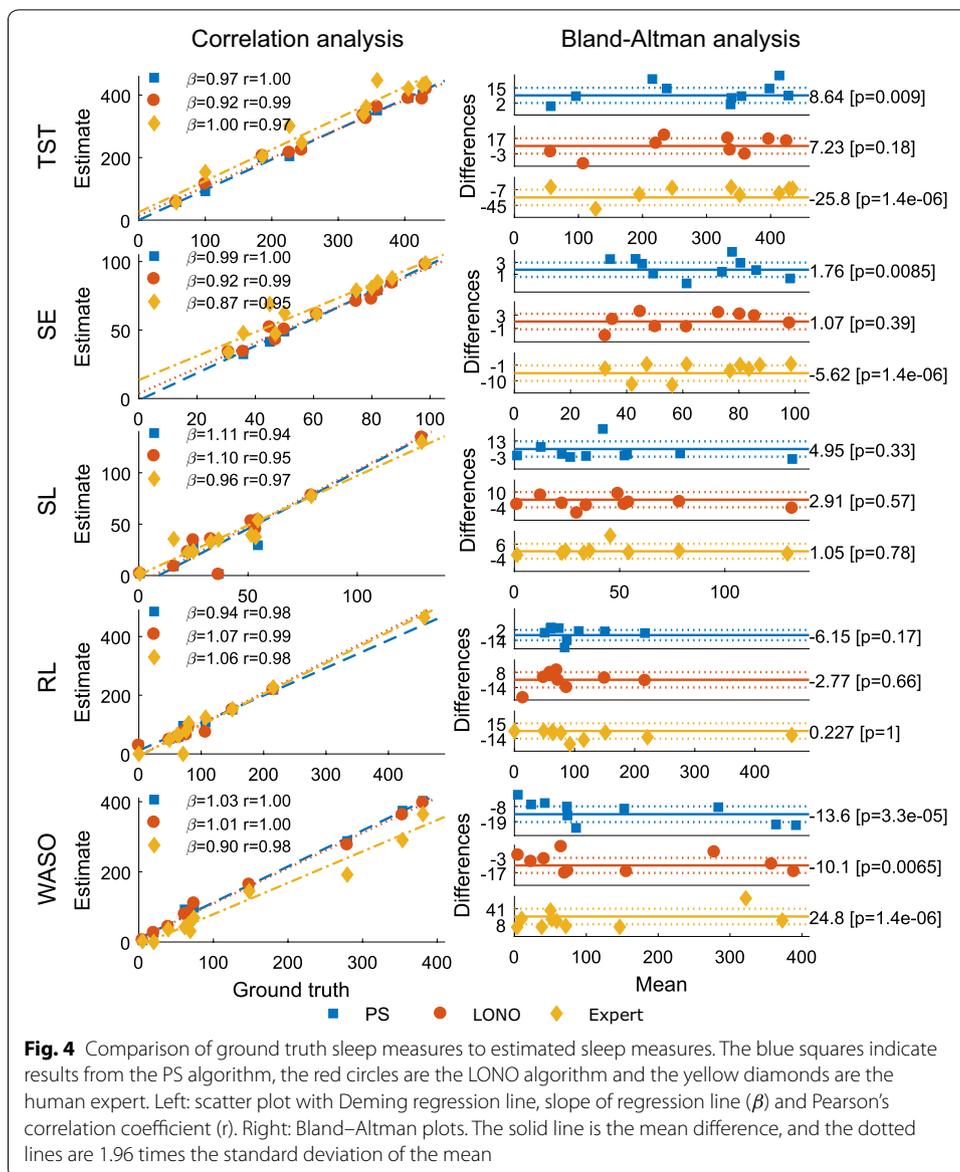
The confusion matrices can be seen in Fig. 3. The PS algorithm classified 96.6% of the wake epochs correctly. The second best class sensitivity was seen for N2, where the individual approach classified 87.2% correctly. Then followed REM sleep with a class sensitivity of 81.4%, N3 with 82.6%, and lastly the N1 class with a poor class sensitivity of 10.4%. The order of the classes according to their class sensitivity was the same for the LONO approach and the human expert. However, the human expert had a substantially better performance on the N1 class, with a class sensitivity of 40.9%. On the simpler sleep–wake classification task, the performances were higher. The PS approach had a



specificity of 96.6% and a sensitivity of 94.8%. The LONO approach performed similarly, with a specificity of 94.8% and a sensitivity of 94.1%. The human expert had a specificity of 84.0% and a sensitivity of 98.9%.

Sleep measures

A comparison of the estimated measures and the ground truth are shown Fig. 4. In general, there is a high agreement between the estimated and ground truth sleep measures as measured by the Deming slope and correlation coefficient, except for the estimation



of REM latency by the PS algorithm. Here, the slope β of the Deming regression line is 0.23 and the correlation coefficient $r = 0.47$ due to a single outlier (patient D, night 3) where the ground truth latency is 457.5 min, and the estimated value is 100.5 min. As can be seen in Fig. 6, this night contains two sleep periods with a long wake period in between. According to the manually labeled hypnograms, the first REM epoch occurs after 457.5 min in the second sleep period. However, the PS algorithm predicts a single REM epoch after 100.5 min in the first sleep period.

The Bland–Altman analysis shows that there are three sleep measures where the difference in estimated and ground truth values are significant. The first is the TST, where the PS algorithm underestimates the measure by 8.64 min ($p = 0.009$) and the human expert overestimates it by 25.8 min ($p < 1e-4$). The second measure is the SE, where the PS algorithm underestimates it by 1.76 min ($p = 0.0085$) and the expert overestimates it

by 5.62 min ($p < 1e-4$). The third is the WASO, where the PS algorithm overestimates it by 13.6 min ($p < 1e-4$), LONO overestimates it by 10.1 min ($p = 0.0065$) and the expert underestimates it by 24.8 min ($p < 1e-4$).

Discussion

We have automatically and manually sleep scored subcutaneous EEG from four patients with epilepsy, and achieved good results as compared to manually labeled LTV EEG. When classifying five stages, our best performing algorithm (the PS approach) achieved a mean Cohen's kappa value of 0.78 across patients, which according to McHugh et al. [18] represents a *moderate agreement*. When classifying sleep vs. wake, we achieved a mean kappa value of 0.85, which is in *strong agreement*.

The PS models outperformed the LONO models on almost all nights, even though the LONO models have more training data available in each CV fold. However, the majority of data in each training fold in the LONO approach originates from other patients than the one whose night is in the test fold. The difference in performance highlights the strength of long-term recording devices: the ability to provide large amounts of data from a specific person that can give rise to highly personalized algorithms rather than one-size-fits-all solutions.

The estimated sleep measures TST, SE, SL, RL and WASO were in good agreement with the ground truth values of the sleep measures computed from the manually labeled LTV EEG, as measured by the slope of the Deming regression line and correlation coefficient. Bland–Altman analysis revealed that there was a significant difference between the estimated values and the ground truth values for the TST, SE and WASO parameters across nights in the data set. However, the differences were relatively small for the PS models.

The algorithms outperformed the human expert on all performance measures except the class sensitivity of N1. However, this comparison is hardly fair, as the human expert had no training on subcutaneous EEG before scoring. There could possibly have been some learning for the human scorer during the course of analysis, but this was not tested for.

The result should be interpreted in the light of the weaknesses and strengths of this study. Firstly, it is based on a small data set with only four patients. However, multiple nights were recorded for each patient, allowing for training of patient-specific models with cross-validation schemes that is not based on random splitting epochs into train and test folds. By training on all nights except one and testing on the last night, the strong temporal inter-dependence between epochs from the same night is respected.

Secondly, the ground truth is manually scored hypnograms based on the full scalp EEG from LTV EEG, and not a PSG as according to the AASM guidelines. Although long-term PSG recordings would have been preferable, it was considered impractical and an unacceptable additional burden for the patients.

Thirdly, the placement of the subcutaneous electrodes was chosen to maximize the probability of recording temporal lobe seizures. This placement might be sub-optimal for sleep scoring, as some common sleep phenomena are best seen fronto-centrally [2]. As the implant can be placed at a variety of positions as long as the disk-shaped housing is located behind the ear, one might expect better results with a more optimal placement.

Finally, the 11 ground truth-hypnograms reflect how the patients slept poorly in the EMU. Furthermore, it is well established that epilepsy can induce sleep disturbances [19, 20]. The present classification performance might have improved if the data set consisted of 11 nights from four healthy subjects that were good sleepers.

As already mentioned, several relevant studies have sleep scored wearable EEG, and it is natural to compare these to the present study. Studies conducted on ear-EEG are perhaps the most relevant, as this modality also has the potential to provide unobtrusive, ultralong-term measurements. However, the hardware is still under development and there are no commercially available solutions yet. Nakamura et al. [13] conducted a small study on four healthy male subjects, where they recorded 45 min daytime naps after a sleep-deprived night using ear-EEG. They achieved a Cohen's kappa of 0.65 when classifying the four stages W, N1, N2 and N3, and a kappa of 0.8 when classifying sleep vs. wake. Mikkelsen et al. [11] conducted a larger study using ear-EEG with nine healthy subjects. One night was recorded per subject. They achieved an average Cohen's kappa value of 0.65 on five classes across subjects when training subject-specific models. By merging all sleep labels into a single sleep class, they achieved a sensitivity of 81% and a specificity of 97%. The present study shows better results, probably due to the fact that multiple nights were recorded per subject. In 2019, Mikkelsen et al. [12] did a larger study on 15 healthy subjects, this time using around-the-ear, flex-printed electrode arrays and Actiwatches (MW8, CamNtech, UK). Although the electrode arrays are less suitable for ultralong-term recordings, the experimental setup and aim of the study were similar to the present study. One night was recorded for each subject, and they used a leave-one-subject-out CV strategy. They achieved a mean Cohen's kappa of 0.54 for the five-class problem and a mean Cohen's kappa of 0.69 for the two-class problem. They found that the EEG-based device outperformed the Actiwatch in sleep detection. Mikkelsen et al. also estimated the five sleep measures TST, SE, SL, RL and WASO, and found that the estimates were in agreement with the true values except for RL and WASO. The significant underestimation of brief wake periods reported across several studies calls for future research on how to accurately estimate this parameter using wearable sleep monitors.

For sleep-wake detection, multiple studies have compared actigraphy to PSG. Most studies find that the sleep measures produced by the actigraphs are well correlated with the measures reported by PSG, but the modality often suffers from poor specificity. Kosmadopoulos et al. [21] assessed the validity of an Actiwatch-64 (Mini-Mitter Philips Respironics, Bend, OR) against PSG in 22 healthy subjects. By adjusting the activity threshold in the embedded sleep detection algorithm, they achieved sets of sensitivity and specificity ranging from 87.6 to 61.5 % (very low activity threshold) to 97.8% and 26.9% (high activity threshold). The Cohen's kappa values ranged from 0.30 to 0.37. Slater et al. [22] assessed a GTX3+ Actigraph, and found a sensitivity, specificity and accuracy of 90, 46 and 84%, respectively. An extensive literature review on the subject was out of scope for this paper, but these results are in line with the review of Sadeh et al. [3]. Compared to actigraphy, EEG-based devices for ultralong-term sleep monitoring have two advantages: the ability to distinguish between sleep stages and a much higher specificity.

When developing and evaluating an alternative method, the question of "how good is good enough" naturally arises. Surely, the goal of a medical device must be to obtain clinical relevance. Werner et al. [23] compared the sleep patterns of children as reported by actigraphy and sleep diaries. They compared several sleep measures, including TST and WASO. Based on the author's clinical experience, the difference between the estimates based on actigraphy and sleep diary were considered in clinically acceptable agreement if it was less than 30 min. In the current study, the PS estimates of both TST and WASO as compared to the true values were within the 30 min limit for all nights. For the LONO approach, the estimates were all within the limit except for a single night.

When comparing Cohen's kappa values, a reasonable benchmark for clinical relevance could be the inter-rater reliability (IRR) one might expect between different trained scorers. Danker-Hopfe et al. [24] found that the IRR as measured by Cohen's kappa between scorers from eight European sleep laboratories was 0.6816. When the AASM guideline was introduced, the IRR increased to a kappa value of 0.76. As our best-performing algorithm achieved similar kappa values, we argue that the proposed method could deliver performances on par with current clinical practice.

As the data set only contained nights spent in the EMU, it remains an open question whether the algorithm can generalize to nights recorded in the patients' own homes. Monitoring the sleep quality in patients with epilepsy over ultralong time periods could potentially have clinical value, as there is a complex interplay between sleep quality and epileptic activity. Studies have shown an improvement in seizure control when sleep disturbances were treated [20, 25]. Future research is needed to illuminate the clinical utility of ultralong-term EEG monitoring of sleep patterns in epilepsy patients recorded "in the wild".

Conclusion

By recording several nights per patient, we were able to train patient-specific models and achieved a mean Cohen's kappa value of 0.78 across recordings. This is higher than the inter-rater agreement one would expect between two human raters from different sleep laboratories, as reported by Danker-Hopf et al. When detecting sleep vs. wake, we achieved a sensitivity of 94.8% and a specificity of 96.6%, which is an improvement over the widely used actigraphy.

Of the five sleep measures TST, SE, SL, RL and WASO, we found significant differences in TST, SE and WASO. The differences were small and within reported clinically acceptable limits.

In conclusion, we are the first to show that sleep monitoring patients with epilepsy using subcutaneous EEG and automatic scoring algorithms is possible and can produce results of clinical relevance. Ultralong-term EEG combines the strengths of the PSG and actigraphy, providing both accurate sleep stage scoring and long-term measurements. With the possibility of recording ultralong measurements, there is a potential to develop strong patient-specific sleep scoring algorithms that could illuminate sleep pattern over weeks and months.

Methods

Data collection

Four adult epilepsy patients (one male, three female) with a temporally implanted two-channel EEG system were admitted for full channel workup in the Epilepsy Monitoring Unit (EMU) at Zealand University Hospital [2]. The four adults are a sub-population of a clinical study comprising nine adults with the implanted EEG system, but only four of the trial participants were admitted to the EMU [2]. During the patient's EMU stay, LTV EEG and subcutaneous EEG were recorded simultaneously.

The subcutaneous system consists of an implant and an externally worn device. The implant has three electrodes, where the center electrode acts as a reference to create two bipolar channels. It is implanted under the skin behind the ear under local anesthesia, and were placed such that the electrodes span the temporal lobe. The electrodes are named Distal (*D*), Center (*C*) and Proximal (*P*), where *P* is the closest to the ear. The external device contains a rechargeable battery and a memory chip. The external device both powers the implant and receives data through an inductive link across the skin, and has a sampling frequency of 207 Hz. The device is produced by UNEEG medical A/S (Lyngby, Denmark), and in this study, a beta version of the 24/7 EEG SubQ device was used. The commercially available device is marketed as a tool for treatment optimization by providing an objective estimation of the seizure burden. An illustration is provided in Fig. 5. The patients were given two external devices, one to wear during daytime (awake) and one for nighttime (sleep). The external devices were recharged when not in use. The time stamp for the start of the "nighttime device"-recording is considered as "lights off". The LTV EEG was recorded with a NicoletOne wireless 64-channel head box (CareFusion 209) with a sampling rate of 1024 Hz. The 25 scalp EEG electrodes were placed according to the international 10-20 system with additional low row. The exact electrode placement can be seen in Appendix A.

The recordings from the two EEG modalities were sleep stage scored by a trained expert according to the AASM guidelines. The expert scored the recordings manually (and not computer assisted), as this is customary in Denmark. Only recordings from the period where the patient was wearing the subcutaneous "night device" were considered,



Fig. 5 Illustration of the subcutaneous recordings system. Left: illustration of the implant and the beta-version of the external device used to collect data in the present study. The placement of the Proximal (*P*), Center (*C*) and Distal (*D*) electrodes are indicated by the letters. The length of the implant is approximately 11 cm. Right: illustration of the commercially available device. The device is worn under the shirt and secured in place by a magnet (gray circle)

therefore any daytime naps are excluded. Both subcutaneous and scalp recordings were scored using Nicolet One version 5.95. For each patient, the subcutaneous recordings were scored before the scalp recordings, to get as unbiased subcutaneous scorings as possible.

Data set

In total, 11 nights with concurrent scalp and subcutaneous EEG were recorded. Patient B had two nights, and the rest had three nights each. Patient A had two nights and patient C had one night where the external device was removed during the night. The part of the nights where both EEG modalities were recorded were included in the data set, as they were considered usable despite their short length. For an overview of age, gender, seizure onset zone, anti-epileptic drug intake, relevant MRI findings and total duration of EEG recordings for each EEG modality during the course of the EMU stay, the reader is referred to [2].

The hypnograms scored based on the LTV EEG, which are considered the ground truth, are visualized in Fig. 6. The hypnograms show that the nights are rather diverse. Patient D had trouble sleeping and patient C barely had any deep sleep (N3). The EEG technician at the EMU clinic noted that the patient most likely has an undiagnosed sleep apnea. Patient B had a seizure few hours before "lights off" on both evenings, and patient D had >80 seizures during the EMU stay, the majority occurring in the evening before night 3. Observing poor sleep quality in the data set is therefore expected, as the extensive equipment setup, having epilepsy and the occurrence of seizures, can negatively impact sleep [11, 19].

Data pre-processing and feature extraction

The EEG recordings and the hypnograms were imported into MATLAB version 2017a (MathWorks), in which all subsequent analysis was made. To synchronize the scalp and subcutaneous recordings, derivations with electrode placements similar to the subcutaneous placement were extracted from the scalp recordings. Both EEG recordings were band-pass filtered between 0.5 and 100 Hz, and notch filtered around 50 Hz. The scalp channels were downsampled to 207 Hz so that the EEG in the two recordings could be synchronized using the cross correlation. A crude artifact rejection was performed by removing parts of the EEG with an amplitude larger than 300 μ V, to avoid having high amplitude artifacts dominate the cross correlation. The artifacts were removed from all subsequent EEG analysis. An overview of the amount of rejected artifact for each night can be seen in Appendix B. The synchronization was performed on a 30-s epoch-by-epoch basis, such that each scored epoch in the scalp EEG had a corresponding synchronized subcutaneous EEG epoch.

Thirty features were computed for each subcutaneous channel and epoch, and they are listed in Table 1. The features are based on the power distribution in the traditional frequency bands computed using MATLAB's continuous wavelet transform (cwt). These features were chosen, as the AASM scoring rules to a large extent are based on activity in these frequency ranges. When computing the wavelet transform, the analytic Morse (3,60) wavelet was used, where 3 is the symmetry parameter and 60 is the time-bandwidth product. The traditional frequency bands were defined as delta: 0.5–4 Hz, theta:

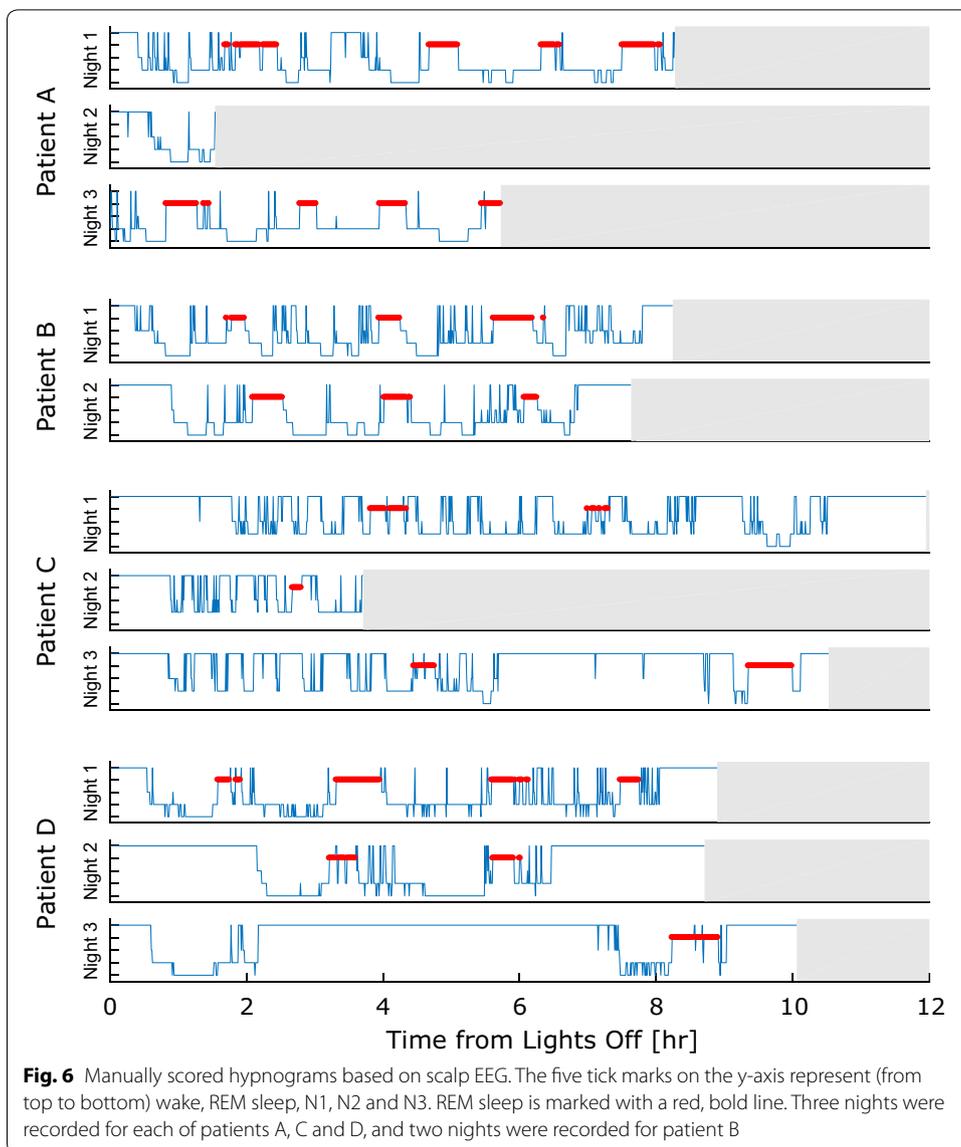


Table 1 Description of the 30 features that were computed for each EEG channel. The five frequency bands are the delta, theta, alpha, lower beta and upper beta

| Feature number | Feature description |
|----------------|---|
| 1–5 | Mean power in the five frequency bands |
| 6–10 | Variance of the power distribution in the five frequency bands |
| 11–15 | Skewness of the power distribution in the five frequency bands |
| 16–20 | Kurtosis of the power distribution in the five frequency bands |
| 21–25 | Shannon entropy of the power distribution in the five frequency bands |
| 26–30 | Duration of the activation of the power in the five frequency bands |

4–8 Hz, alpha: 8–13, lower beta: 13–22 Hz and upper beta: 22–32 Hz. Note that the beta band has been divided into two separate bands, as it the traditional definition of this band is quite broad.

The five last features listed in the table are the duration of the activation of the power in the frequency bands. Here, "activation" is defined as the mean power in a given frequency band exceeding a threshold of 1.5 times the median of the power in the range 0.5–100 Hz. As an example, the duration of the activation of the delta band for a given epoch is the amount of time that the mean power in the delta band exceeds the threshold during the 30 s epoch.

As the manual scoring of an epoch may depend on the label of the surrounding epochs, the epoched features were concatenated with the features from the preceding epoch and two subsequent epochs. Each epoch of EEG was thus represented by the features associated with four epochs.

Classification

By using the hypnogram from the scalp EEG as ground truth, an automatic sleep stage algorithm was trained on the features extracted from the synchronized two-channel subcutaneous EEG. The sleep staging algorithm used in this study is a random forest similar to [11, 12] that has scored around-the-ear-EEG with success. The forest consists of 100 decision trees fitted using the `fitensemble` function with the Bag method. The trees were trained by resampling the training data set with replacement. Every tree in the ensemble randomly selected features for node splitting. The tree nodes were split based on their impurity (Gini's Diversity Index) using the standard CART algorithm. The splitting of a node continued until either the node was pure, there were fewer than ten observations in the node, or the algorithm had split $N_{TrainingSetSize} - 1$ nodes. The model was cross-validated using two different schemes: patient specific (PS) and leave-one-night-out (LONO). In the PS scheme, one model was trained for each patient, in a leave-one-night-out manner. In the LONO approach, an 11-fold CV was achieved by using all possible combination of ten nights as training set, and using the remaining night as test set. Common for both CV strategies is that data samples from the same night are always in the same fold. This is in contrast to the often-used CV technique where epochs are assigned to training and test folds randomly, which ignores the strong dependence between epochs recorded close in time [5, 26].

Furthermore, we computed five sleep measures recommended by the AASM manual version 2.4 [27] from the hypnograms. The first measure is the total sleep time (TST), which is the total time spent asleep, and is the sum of the time spent in any of the sleep stages. Inspired by ANSI/CTA 2052.1 Standard *Definitions and Characteristics for Wearable Sleep Monitors* [28], we computed the time attempting to sleep (TATS) instead of the total recording time, which is used in the AASM manual. The TATS is a more suitable measure for continuous long-term recordings, and is defined as the time when the patient is in bed and starts attempting to sleep, until the patient is no longer attempting to sleep [28, 29]. TATS is in this study indicated by the mounting and dismounting of the nighttime subcutaneous EEG device. Sleep efficiency percentage (SE) is defined as $TST/TATS \times 100$. Sleep latency (SL) is the time from when the patient begins attempting to sleep until the first sleep epoch of any stage occurs. Similarly, REM latency (RL) is defined as the time from first attempt to sleep to the first epoch of REM sleep. Wake after sleep onset (WASO) is defined as $TATS - SL - TST$. All sleep measures that are not given in percent are given in minutes. A set of "ground truth sleep measures" was

computed from the manually labeled hypnograms based on the scalp EEG and a set of "estimated measures" was computed from the hypnograms predicted by the algorithms and the human expert based on the subcutaneous EEG.

In addition to computing hypnograms, we also derived sleep–wake traces, which is currently the standard in long-term sleep monitoring. The traces were derived by merging the sleep stages in the ground truth and predicted hypnograms post-analysis into a single sleep class. The algorithm was not retrained to do this, and the human expert did not re-score the same data.

Evaluation

Sleep stage classification

To assess agreement between the predicted and ground truth hypnograms and sleep–wake traces, we computed Cohen's kappa and confusion matrices. Cohen's kappa coefficient (κ) is a statistic which measures interrater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as (κ) accounts for the possibility of the agreement occurring by chance. A kappa value was computed for each the 11 predicted hypnograms and sleep–wake traces for each method. A single confusion matrix was computed for each method. This corresponds to computing a population average weighted by the number of patient epochs. The confusion matrices report both the count and the percentage of epochs known to belong to class i that was classified as belonging to class j for $i, j \in \{1, \dots, \text{NumberOfClasses}\}$. The percentage in the diagonal can be thought of as a class sensitivity. For the binary sleep detection problem, the first entry in the diagonal is the class sensitivity of the wake class. In the sleep literature however, it is customary to consider sleep epochs as positive data samples and wake epochs as negative samples. The first entry in the diagonal therefore equals the percentage of known wake epochs that were classified as such, $TN/(TN+FP)$, which is often referred to as the specificity. Similarly, the second entry in the diagonal is the percentage of sleep epochs that were classified as sleep and is often referred to as the sensitivity.

Sleep measures

The ground truth and estimated values for the sleep measures were compared by means of scatter plots and Bland–Altman analysis. For the scatter plots, a straight line was fitted using Deming regression. Deming regression is suitable for data sets where both the X and Y variables are subject to measurement errors. If the sleep measures computed from the manually labeled hypnograms and the predicted hypnograms were in agreement, the slope of the Deming regression line should be close to 1. The Pearson correlation coefficient was also computed between the ground truth and estimated values.

Furthermore, a comparison between the ground truth and the estimated values were made using Bland–Altman analysis [30]. For each night and each sleep measure, the mean of the value of the ground truth sleep measure and the estimated sleep measure was computed, as well as the difference between the two values. The differences were plotted against their mean value, along with the mean difference value and its 95% confidence interval. A permutation test was performed on the differences to test whether the mean difference was different from zero [31]. A permutation test builds a reference

distribution by resampling the observed data as opposed to assuming a reference distribution, as is done in a *t* test. As the number of samples here was low (2×11), it was feasible to run an exact test by considering all possible permutations of the samples when building the reference distribution. The significance level was set to 5%. Significant differences of positive sign implied that the proposed method was underestimating the value of the sleep measure as compared to the ground truth. Conversely, a negative mean difference implied that the proposed method overestimated the sleep measure.

Abbreviations

AASM: American Academy of Sleep Medicine; CART: Classification And Regression Tree; CV: cross-validation; EEG: electroencephalography; EMU: epilepsy monitoring unit; IRR: inter-rater reliability; LONO: leave-one-night-out; LTV EEG: long-term video-EEG; MRI: magnetic resonance imaging; N1: non-rapid eye movement stage 1; N2: non-rapid eye movement stage 2; N3: non-rapid eye movement stage 3; PS: patient specific; PSG: polysomnography; REM: rapid eye movement; RL: rapid eye movement-latency; SE: sleep efficiency; SL: sleep latency; TATS: time attempting to sleep; TST: total sleep time; WASO: wakefulness after sleep onset.

Acknowledgements

The authors would like to thank Søren Kilsgaard and Nelly Richard for the illustration of the subcutaneous EEG recording system.

Authors' contributions

SWG wrote the manuscript, performed pre-processing and data analysis. KM and YRT contributed with sleep scoring algorithms and provided critical feedback. SW collected the data, and MHL manually sleep scored it. PK, TWK, LKH and JDH contributed substantially to the conception of the study, and together with MCH supervised the findings of this work. All authors read and approved the final manuscript.

Funding information

SWG is partially funded by Innovation Fund Denmark, project number 5189-00071A.

Availability of data and materials

The datasets analyzed during the current study are not publicly available due to the interests of UNEEG™ medical A/S. The code used to sleep score can be made available upon request.

Ethics approval and consent to participate

The study was conducted according to the Declaration of Helsinki and was approved by the regional committee of science ethics for Region Zealand (SJ-551). All patients provided written informed consent before enrollment in the study.

Consent for publication

Persons depicted in illustrations gave consent for publication.

Competing interests

SWG, JDH and MCH are employed by UNEEG™ medical A/S. MHL and SW is partially funded by UNEEG™ medical A/S. TWK consults for UNEEG™ medical A/S. The remaining authors declare no competing interests.

Author details

¹ Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Bygning 324, 2800 Kgs. Lyngby, Denmark. ² UNEEG medical A/S, Nymoellevvej 6, 3540 Lyngby, Denmark. ³ Department of Engineering, Aarhus University, Finlandsgade 22, 8200 Aarhus N, Denmark. ⁴ Center of Neurophysiology, Department of Neurology, Zealand University Hospital, Vestermarksvej 11, 4000 Roskilde, Denmark. ⁵ Department of Basic and Clinical Neuroscience, King's College London, 5 Cutcombe Road, SE5 9RX London, UK.

Appendix A

The scalp electrodes available for manual sleep scoring of the long-term video EEG are: Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8, Fz, Cz, Pz, F9, F10, T9, T10, P9, and P10.

Appendix B

A crude artifact rejection was performed on both EEG modalities before synchronization. After synchronization, the parts that were excluded from one of the modalities, were also excluded from the other, in order to make sure that both modalities had the same amount of data.

| Patient | Night | Rejected data (min) |
|---------|-------|---------------------|
| A | 1 | 104 |
| | 2 | 41.9 |
| | 3 | 45.3 |
| B | 1 | 3.28 |
| | 2 | 13.7 |
| C | 1 | 40.6 |
| | 2 | 10.3 |
| | 3 | 11.5 |
| D | 1 | 11.6 |
| | 2 | 21.1 |
| | 3 | 40.3 |

Appendix C

List of Cohen's kappa values for the sleep stage classification and the sleep-wake classification. See Tables 2, 3.

Table 2 Cohen's kappa values for the 5-class problem

| Patient | Night | PS κ | LONO κ | Expert κ |
|------------------|-------|--------------------|--------------------|--------------------|
| A | 1 | 0.81 | 0.79 | 0.80 |
| | 2 | 0.84 | 0.83 | 0.80 |
| | 3 | 0.84 | 0.83 | 0.90 |
| B | 1 | 0.62 | 0.56 | 0.56 |
| | 2 | 0.82 | 0.71 | 0.56 |
| C | 1 | 0.71 | 0.69 | 0.59 |
| | 2 | 0.82 | 0.67 | 0.43 |
| | 3 | 0.78 | 0.77 | 0.59 |
| D | 1 | 0.73 | 0.70 | 0.60 |
| | 2 | 0.79 | 0.77 | 0.73 |
| | 3 | 0.79 | 0.77 | 0.74 |
| Mean (\pm SD) | | 0.78 (\pm 0.02) | 0.74 (\pm 0.02) | 0.66 (\pm 0.04) |

Table 3 Cohen's kappa values for the 2-class problem

| | Night | PS κ | LONO κ | Expert κ |
|------------------|-------|--------------------|--------------------|--------------------|
| A | 1 | 0.81 | 0.79 | 0.83 |
| | 2 | 0.92 | 0.92 | 0.94 |
| | 3 | 0.58 | 0.53 | 0.70 |
| B | 1 | 0.84 | 0.83 | 0.81 |
| | 2 | 0.88 | 0.85 | 0.87 |
| C | 1 | 0.90 | 0.85 | 0.74 |
| | 2 | 0.89 | 0.76 | 0.49 |
| | 3 | 0.91 | 0.92 | 0.73 |
| D | 1 | 0.82 | 0.78 | 0.88 |
| | 2 | 0.91 | 0.90 | 0.97 |
| | 3 | 0.92 | 0.90 | 0.92 |
| Mean (\pm SD) | | 0.85 (\pm 0.03) | 0.82 (\pm 0.03) | 0.81 (\pm 0.04) |

Received: 13 June 2019 Accepted: 16 October 2019

Published online: 30 October 2019

References

- Smith MT, McCrae CS, Cheung J, Martin JL, Harrod CG, Heald JL, Carden KA. Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: an american academy of sleep medicine systematic review, meta-analysis, and grade assessment. *J Clin Sleep Med*. 2018;14(7):1209–30. <https://doi.org/10.5664/jcsm.7228>.
- Weisdorf S, Gangstad SW, Duun-Henriksen J, Mosholt KSS, Kjær TW. High similarity between eeg from subcutaneous and proximate scalp electrodes in patients with temporal lobe epilepsy. *J Neurophysiol*. 2018;120(3):1451–60. <https://doi.org/10.1152/jn.00320.2018>.
- Sadeh A. The role and validity of actigraphy in sleep medicine: an update. *Sleep Med Rev*. 2011;15(4):259–67. <https://doi.org/10.1016/j.smrv.2010.10.001>.
- Sadaka Y, Sadeh A, Bradbury L, Massicotte C, Zak M, Go C, Shorer Z, Weiss SK. Validation of actigraphy with continuous video-electroencephalography in children with epilepsy. *Sleep Med*. 2014;15(9):1075–81. <https://doi.org/10.1016/j.sleep.2014.04.021>.
- Vilamala A, Madsen KH, Hansen LK. Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring. 2017;8.
- Tsinalis O, Matthews PM, Guo Y. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Ann Biomed Eng*. 2016;44(5):1587–97. <https://doi.org/10.1007/s10439-015-1444-y>.
- Tsinalis O, Matthews PM, Guo Y, Zafeiriou S. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. 2016;12.
- Mousavi S, Afghah F, Acharya UR. Sleeppegnet: automated sleep stage scoring with sequence to sequence deep learning approach. *PLoS ONE*. 2019;14(5):0216456. <https://doi.org/10.1371/journal.pone.0216456>.
- O'Reilly C, Gosselin N, Carrier J, Nielsen T. Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. *J Sleep Res*. 2014;23(6):628–35. <https://doi.org/10.1111/jsr.12169>, <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.12169>.
- Aboalayon KAI, Faezipour M, Almuhammadi WS, Moslehpour S. Sleep stage classification using EEG signal analysis: a comprehensive survey and new investigation. *Entropy*. 2016;18(9):272. <https://doi.org/10.3390/e18090272>.
- Mikkelsen KB, Villadsen DB, Otto M, Kidmose P. Automatic sleep staging using ear-EEG. *Biomed Eng Online*. 2017;16(1):111. <https://doi.org/10.1186/s12938-017-0400-5>.
- Mikkelsen KB, Ebajemito JK, Bonmati-Carrion MA, Santhi N, Revell VL, Atzori G, della Monica C, Debener S, Dijk DJ, Sterr A, de Vos M. Machine-learning-derived sleep-wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy. *J Sleep Res*. 2019;28(2):12786. <https://doi.org/10.1111/jsr.12786>.
- Nakamura T, Goverdovsky V, Morrell MJ, Mandic DP. Automatic sleep monitoring using ear-EEG. *IEEE J Transl Eng Health Med*. 2017;5:7959059. <https://doi.org/10.1109/JTEHM.2017.2702558>.
- Griessenberger H, Heib DPJ, Kunz AB, Hoedlmoser K, Schabus M. Assessment of a wireless headband for automatic sleep scoring. *Sleep Breath*. 2013;17(2):747–52. <https://doi.org/10.1007/s11325-012-0757-4>.
- Mylymaa S, Muraja-Murro A, Westeren-Punnonen S, Hukkanen T, Lappalainen R, Mervaala E, Töyräs J, Sipilä K, Mylymaa K. Assessment of the suitability of using a forehead EEG electrode set and chin emg electrodes for sleep staging in polysomnography. *J Sleep Res*. 2016;25(6):636–45. <https://doi.org/10.1111/jsr.12425>.

16. Baumgartner C, Koren JP. Seizure detection using scalp-EEG. *Epilepsia*. 2018;59(Suppl. 1):14–22. <https://doi.org/10.1111/epi.14052>.
17. Duun-Henriksen J, Kjaer TW, Looney D, Atkins MD, Sørensen JA, Rose M, Mandic DP, Madsen RE, Juhl CB. Eeg signal quality of a subcutaneous recording system compared to standard surface electrodes. *J Sensors*. 2015;2015:341208. <https://doi.org/10.1155/2015/341208>.
18. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med*. 2012;. <https://doi.org/10.11613/BM.2012.031>.
19. Bazil CW. Epilepsy and sleep disturbance. *Epilepsy Behav*. 2003;4(2):39–453945. <https://doi.org/10.1016/j.yebeh.2003.07.005>.
20. Lanigar S, Bandyopadhyay S. Sleep and epilepsy: a complex interplay. *Mo Med*. 2017;114(6):453–457453457.
21. Kosmadopoulos A, Sargent C, Darwent D, Zhou X, Roach GD. Alternatives to polysomnography (PSG): a validation of wrist actigraphy and a partial-PSG system. *Behav Res Methods*. 2014;46(4):1032–41. <https://doi.org/10.3758/s13428-013-0438-7>.
22. Slater JA, Botsis T, Walsh J, King S, Straker LM, Eastwood PR. Assessing sleep using hip and wrist actigraphy. *Sleep Biol Rhythm*. 2015;13(2):172–80. <https://doi.org/10.1111/sbr.12103>.
23. Werner H, Molinari L, Guyer C, Jenni OG. Agreement rates between actigraphy, diary, and questionnaire for children's sleep patterns. *Archiv Pediatr Adolesc Med*. 2008;162(4):350–8. <https://doi.org/10.1001/archpedi.162.4.350>.
24. Danker-Hopfe H, Kunz D, Gruber G, Klösch G, Lorenzo JL, Himanen SL, Kemp B, Penzel T, Röschke J, Dorn H, Schlögl A, Trenker E, Dorffner G. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *J Sleep Res*. 2004;13(1):63–9. <https://doi.org/10.1046/j.1365-2869.2003.00375.x>.
25. Pornsriniyom D, Kim HW, Bena J, Andrews ND, Moul D, Foldvary-Schaefer N. Effect of positive airway pressure therapy on seizure control in patients with epilepsy and obstructive sleep apnea. *Epilepsy Behav*. 2014;37:270–5. <https://doi.org/10.1016/j.yebeh.014.07.005>.
26. Munk AM, Olesen KV, Gangstad SW, Hansen LK. Semi-supervised sleep-stage scoring based on single channel EEG. In: Proceedings of 2018 IEEE international conference on acoustics, speech and signal processing. 2018;2018:2551–2555. <https://doi.org/10.1109/ICASSP.2018.8461982>.
27. Berry R, Brooks R, Gamaldo C et al. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. Version 2.4 ed. 2017.
28. Consumer Technology Association, National Sleep Foundation. Definitions and characteristics for wearable sleep monitors, Ansi/cta-2052.1 edn. Consumer Technology Association, National Sleep Foundation; 2016.
29. Reed DL, Sacco WP. Measuring sleep efficiency: what should the denominator be? *J Clin Sleep Med*. 2016;12(2):263–6. <https://doi.org/10.5664/jcsm.5498>.
30. Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–10.
31. Ernst MD. Permutation methods: a basis for exact inference. *Stat Sci*. 2004;19(4):676–85. <https://doi.org/10.1214/088342304000000396>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

