Published in final edited form as:
*J Sleep Res*. 2014 April ; 23(2): 211–221. doi:10.1111/jsr.12105.

# Automatic scoring of sleep stages and cortical arousals using two electrodes on the forehead: validation in healthy adults

**Djordje Popovic**[1,2], **Michael Khoo**[2], and **Philip Westbrook**[1]

[1]Advanced Brain Monitoring Inc., Carlsbad, CA, USA

[2]Department of Biomedical Engineering, University of Southern California, Los Angeles, CA, USA

## SUMMARY

Accuracy and limitations of automatic scoring of sleep stages and electroencephalogram arousals from a single derivation ($F_{p1}$–$F_{p2}$) were studied in 29 healthy adults using a portable wireless polysomnographic recorder. All recordings were scored five times: twice by a referent scorer who viewed the standard polysomnographic montage and observed the American Academy of Sleep Medicine rules (referent scoring and blind rescoring); and once by the same scorer who viewed only the $F_{p1}$–$F_{p2}$ signal (alternative scoring), by another expert from the same institution, and by the algorithm. Automatic, alternative and independent expert scoring were compared with the referent scoring on an epoch-by-epoch basis. The algorithm's agreement with the reference (81.0%, Cohen's $\kappa = 0.75$) was comparable to the inter–rater agreement (83.3%, Cohen's $\kappa = 0.78$) or agreement between the referent scoring and manual scoring of the frontopolar derivation (80.7%, Cohen's $\kappa = 0.75$). Most misclassifications by the algorithm occurred during uneventful wake/sleep transitions, whereas cortical arousals, rapid eye movement and stable non-rapid eye movement sleep were detected accurately. The algorithm yielded accurate estimates of total sleep time, sleep efficiency, sleep latency, arousal indices and times spent in different stages. The findings affirm the utility of automatic scoring of stages and arousals from a single frontopolar derivation as a method for assessment of sleep architecture in healthy adults.



**Correspondence**. Djordje Popovic, MD, PhD, Advanced Brain, Monitoring (ABM Inc.), 2237 Faraday Avenue, Carlsbad, CA 92008, USA. Tel.: +1-760-720-0099; fax: +1-760-476-3620; popovic@usc.edu.

**DISCLOSURE STATEMENT**

DP is an employee of ABM Inc., and PW is a medical officer for ABM and Ventus Medical; both companies market devices for diagnosis or treatment of sleep disorders. MK has no conflict of interest.

**SUPPORTING INFORMATION**

Additional Supporting Information may be found in the online version of this article:

**Table S1.** Descriptors of sleep macro-structure used by FP-STAGER.

**Table S2.** Rules and thresholds in the algorithm's decision tree.

**Figure S1.** Agreement between the referent scorer (initial round, rescoring and consolidated reference) and an independent expert in the groups of sleep-deprived and well-rested volunteers. The reference is always in rows, while the independent expert scoring is in columns.

**Figure S2.** Agreement between the referent scorer (initial round, rescoring and consolidated reference) and manual scoring of the $F_{p1}$–$F_{p2}$ derivation (alternative scoring) in the groups of sleep-deprived and well-rested volunteers. The reference is always in rows, while the alternative scoring is in columns.

**Keywords**

automated sleep scoring; forehead; single channel

## INTRODUCTION

The standard somnological armamentarium lacks affordable tools for objective assessment of sleep architecture on a large scale or repeatedly over time (e.g. epidemiological studies or clinical trials). Multichannel polysomnography (PSG) is encumbering and costly for such applications; simpler devices for assessment of sleep-disordered breathing (SDB) provide only autonomic indices of sleep quality that have not been validated in non-SDB populations; actigraphs estimate duration and timing of sleep but render no information about its macro- or microstructure; finally, diaries and questionnaires are subjective. Recent years have seen the emergence of self-applicable recorders worn on the forehead that assess sleep architecture from the mixture of electroencephalograms (EEGs), electrooculograms (EOGs) and electromyograms (EMGs) recorded in a single differential derivation ($F_{p1}$–$F_{p2}$). Given their user-friendliness and low cost, devices of this type could become a valuable complement to the established somnological methods in ambulatory settings, provided the accompanying software is able to, at least, score the conventional sleep stages and cortical arousals. Unfortunately, the forehead-worn systems introduced hitherto staged sleep with modest accuracy and did not detect EEG arousals at all (Fischer *et al.*, 2004; Popovic *et al.*, 2008; Shambroom *et al.*, 2012). It is, however, unclear whether their performance was hindered by deficiencies of the devices or algorithms, or limited by insufficiency of the information contained in the $F_{p1}$–$F_{p2}$ signal. Resolving the dilemma deserves attention because the forehead is an opportune location where self-application of sensors or miniature devices is relatively easy. This study has, therefore, investigated the accuracy and limitations of automatic scoring of sleep stages and EEG arousals from the $F_{p1}$–$F_{p2}$ derivation in healthy adults under optimal recording conditions, i.e. using the standard PSG equipment. The algorithm is described in detail, and the impact of relevant factors on its accuracy is discussed.

## MATERIALS AND METHODS

### Data

The data originated from two studies that investigated the relationship between sleep and performance in healthy volunteers. One study contributed with daytime recordings of 21 subjects [nine females; age: 26 ± 6 years; body mass index (BMI): 22 ± 2 kg $m^{-2}$] who took a nap upon completion of a sleep restriction protocol; the other provided nocturnal recordings of a different 18 subjects (10 females; age: 27 ± 6 years; BMI: 23 ± 2 kg $m^{-2}$). Subjects provided written informed consent, and were compensated for their participation. The self-reported absence of sleep problems was corroborated by standard questionnaires, sleep logs and actigraphy (Actiwatch_64; Philips Respironics, Bend, OR, USA). Biomedical Research Institute of America Institutional Review Board (San Diego, CA, USA) approved both studies.

The recording conditions were identical for daytime and nocturnal sessions. Subjects slept in a dark and soundproof room monitored with an infrared camera. Scalp EEG ($C_3$–$A_2$, $C_4$–$A_1$, $F_z$–$O_z$, $C_z$–$O_z$), left and right EOG, submental EMG, the $F_{p1}$–$F_{p2}$ signal and the signal from a respiration belt were acquired using an ambulatory PSG recorder (×10, Advanced Brain Monitoring, Carlsbad, CA, USA). The forehead electrodes were placed on the frontal eminences, approximately 1 cm laterally from the $F_{p1}$ and $F_{p2}$ positions of the 10–20 system. All signals were filtered (0.1–70 Hz, 20 dB decade$^{-1}$), digitized (256 Hz), wirelessly transmitted to the monitoring station in the adjacent room, and stored on a disk.

For the purposes of this study, the available sleep recordings were divided into three groups (Table 1). The data from the training group, which included 10 randomly selected daytime recordings, served to derive parameters of the algorithm for automatic scoring of sleep stages and cortical arousals from the $F_{p1}$–$F_{p2}$ signal. The algorithm was subsequently validated on the remaining 11 daytime and all 18 nocturnal recordings that, respectively, constituted the ‘sleep-deprived’ and ‘well-rested’ validation group. The groups were similar with regard to the demographic and sleep variables, and the sleep-deprived and well-rested groups were therefore often combined in the analyses to increase the statistical power. However, the validation groups were analysed separately when the between-group differences in sleep architecture could help elucidate limitations of either the algorithm or $F_{p1}$–$F_{p2}$ signal morphology.

## Algorithm

FP-STAGER is a modification of our algorithm that scores the referential EOG leads (Levendowski *et al.*, 2012). It is implemented in MATLAB 7.7 (MathWorks, Natick, MA, USA) and involves four major steps (Fig. 1): spectral decomposition of the input signal; computation of descriptors of sleep macro- and microstructure; artefact detection; and classification of 30-s epochs into one of the five stages [wake, rapid eye movement (REM), non-(N)REM1, NREM2 or NREM3]. The input signal is decomposed into delta, theta, alpha, sigma, beta and EMG bands using digital filters. Two signals are derived in the delta band, one from the raw signal and one after removal of ocular artefacts with a median filter. The other bands are extracted directly from the raw signal (eye movements had little impact on the signal power >4 Hz). Descendant signals in each band are integrated and fed to the feature extraction block.

Six descriptors of sleep macro-structure [sigma-beta index (SBI), delta-beta index (DBI), eye movement index (EMI), beta-EMG index (BEI), average EMG activity ($\overline{\mathrm{EMG}}$); average beta activity ($\bar{\beta}$)] are derived for each 30-s epoch (for details, see Table S1 in the online supplement); their selection was guided by the literature (Uchida *et al.*, 1994) and attempts to mitigate between-subject variability of the envelopes in each band. Three descriptors of microstructure are also determined: number of spindles, number of arousals and total length of all arousals in the epoch. Spindles and arousals are detected by contrasting short-term fluctuations to long-term trends in the signal, as in De Carli *et al.* (1999). Spindles are identified as 0.5- to 2-s segments of the signal during which the sigma envelope is larger than the theta, alpha and beta envelopes, and its instantaneous value exceeds by a factor of two the median value of the sigma envelope calculated over the preceding 30 s. Cortical

arousals during NREM sleep are detected as 3- to 15-s segments during which either alpha or beta envelope exceeds by a factor of two the respective median values calculated over the preceding 90 s; in REM sleep, the EMG envelope within the segment must also exceed the median EMG calculated over 90 s.

Artefact detection proceeds in parallel with the previous two steps and is based on the algorithm devised earlier by our group to detect motion, cardio-respiratory and EMG artefacts in daytime EEG recordings (Berka *et al.*, 2004). The presence of artefacts is assessed in successive non-overlapping 1-s segments of EEG data by evaluating four variables in each segment: the peak-to-peak amplitude and largest slope of the band-pass filtered (0.5–7 Hz) $F_{p1}$–$F_{p2}$ signal, and average beta and EMG power levels of the original signal. The segments for which the four variables exceed pre-defined (fixed) thresholds are considered to be contaminated with artefacts and are consequently excluded from the calculations of the six aforementioned descriptors of sleep macrostructure. Furthermore, the proportion of EEG signal contaminated with artefacts [the artefact index (ARI)] is determined for each 30-s epoch.

The ARI and the descriptors of sleep macro- and microstructure are fed to a hierarchical decision tree with eight nodes. Epochs where the EEG is contaminated with artefacts more than 50% of the time are classified as artefact (A) at node R0. At the next level, node R1 classifies epochs that are not dominated by artefacts into NREM cluster (NREM2; NREM3; some NREM1) or beta-dominated cluster (wake; REM; most of NREM1). The NREM cluster is further separated into light (NREM1/2) and deep sleep (NREM3), whereas the beta-dominated cluster is divided into REM/NREM1 and wake/NREM1 sub-clusters (nodes R2, R3). REM sleep is identified in two steps that resemble the American Academy of Sleep Medicine (AASM) rules for initiation and continuation of REM scoring: seed epochs are first identified with high precision using one set of thresholds, followed by examination of the 3-min segments around each seed against another set of thresholds. At the next level, nodes R4 and R5 separate NREM1 epochs with arousals from the NREM1/2 and REM/NREM1 clusters, and node R6 identifies wake and arousal-free NREM1 epochs. The epochs unclassified at nodes R1–R6 are assigned a stage using a simple ‘score-through’ rule (node R7). The rules and thresholds in the decision tree were derived by a step-wise (node-by-node) maximization of the epoch-by-epoch agreement between the algorithm and manual scoring reference in the training group (Table 1). The rules are summarized in Table S2 in the online supplement.

## Analyses

The study aimed not only to validate the algorithm’s output against a manual reference, but also to assess the impact of morphology of the $F_{p1}$–$F_{p2}$ signal on the algorithm’s accuracy and distinguish it from the effects of the algorithm’s deficiencies or ambiguity of the scoring rules of the AASM (Iber *et al.*, 2007). The traditional analysis of agreement between the algorithm and a human scorer was, therefore, complemented by comparisons of the algorithm’s accuracy to the agreement between two expert scorers and performance of an expert scorer challenged with the task of determining sleep stages from the single frontopolar derivation. Accordingly, the recordings from the two validation groups were

independently scored five times: (1) by an in-house specialist who observed the AASM guidelines [initial referent scoring (IREF)]; (2) by the same expert who, 6 months later, blindly rescored the recordings while viewing only the $F_{p1}$–$F_{p2}$ signal [alternative scoring (ALT)]; (3) by the same expert who, more than 1 year later, blindly rescored all the recordings in accord with the AASM guidelines [repeated referent scoring (RREF)]; (4) by another in-house expert who observed the AASM rules [independent expert scoring (EXP)]; and (5) by the FP-STAGER algorithm [automatic scoring (AUT)]. In all scoring rounds the scorers were instructed to mark off the epochs contaminated with excessive noise (>50% of an epoch) as 'artefact' ($n$ = 136, or 0.7% of all epochs).

The algorithm's performance was first assessed in each validation group by quantifying its epoch-by-epoch agreement with the referent scoring (both IREF and RREF) on all available epochs, including those labelled as artefacts ($n$ = 19 445 from both validation groups). The epoch-by-epoch comparison was also performed against the consolidated reference (CREF), i.e. a subset of 17 438 epochs (89.7% of all epochs) that had been assigned the same stage by the referent scorer during both rounds (IREF and RREF). Both sets of the referent hypnograms were additionally compared with the independent expert scoring (IREF–EXP, RREF–EXP) and the manual scoring of the $F_{p1}$–$F_{p2}$ signal (IREF–ALT, RREF–ALT). Overall percentage agreement, all-stage Cohen's kappa, stage-specific sensitivity (SEN) and positive predictive value (PPV) were calculated from the resultant contingency tables.

The algorithm's estimates of total sleep time (TST), sleep latency (SL), sleep efficiency (SE), wake time after sleep onset (WASO), time spent in stages REM and NREM3 (REMT, SWST), latency to REM and NREM3 (REML, SWSL), and arousal indices (AIs) were then calculated for all 29 recordings and compared with the corresponding variables derived from the referent (IREF and RREF) hypnograms (all epochs included, $n$ = 19 445). Arousal indices were calculated by dividing the total number of arousals identified by each method with the TST determined by the same scoring method. Statistical significance of differences among the five sets of estimates was tested with one-way repeated-measures ANOVA for normally distributed variables, and Kruskal–Wallis tests for the variables with a bimodal distribution (TST, WASO, REMT and SWST). Bland–Altman plots and intra-class correlation coefficients were used to additionally analyse differences between the algorithm and referent scoring for variables whose automatic estimates might be influenced by the (relatively frequent) confusion among stages wake, NREM1 and REM by the algorithm (variables TST, SE, SL, WASO, REMT and REML). The Bland–Altman plots were made only against the initial round of referent scoring (IREF) because the epoch-by-epoch comparisons revealed insignificant differences between the IREF–AUT and RREF–AUT agreement.

Arousals identified by the referent scorer were compared with events marked during the independent, alternative and automatic scoring, and sensitivity and PPV were calculated for the three methods. All segments marked as arousals were counted, including those occurring in epochs scored as 'artefact' by the scorers or the algorithm (38 such instances, or <3% of all segments marked as arousals). The algorithm was also evaluated against the 'consensus' arousals (those marked by both referent and independent scorers). Any overlap between arousal strips was counted as a true positive, whereas segments marked only by the referent

or comparative method were counted as true negatives and false positives, respectively. The Bland–Altman plot was created to compare the duration of referent arousals and their algorithm-generated counterparts.

Finally, in order to assess the degree to which morphology of the $F_{p1}$–$F_{p2}$ signal limited the performance of the algorithm, the accuracy of automatic scoring was contrasted to the agreement between the experts and the performance of the referent scorer when scoring sleep stages from the single frontopolar derivation. In order to simplify the interpretation and eliminate the effects of intra–rater (score–rescore) differences on the agreement measures, the comparison was performed only on the CREF, i.e. a subset of epochs that had been assigned the same stage during both initial (IREF) and repeated scoring (RREF). The CREF included 1527 epochs of wake, 1301 epochs of stage NREM1, 6752 epochs of stage NREM2, 4182 epochs of stage NREM3 and 3541 epochs of REM sleep ($n$ = 17 438 or 89.7% of all epochs), and was therefore representative of all sleep stages and their typical proportions in healthy sleepers. (The CREF also included 135 epochs labelled as ‘artefact’, but they were not of interest for this analysis due to their low count and generally very high agreement among the scorers and algorithm when scoring artefacts.) The inter–rater agreement on such a subset (CREF–EXP) was considered to reflect the ambiguity of the EEG patterns in our dataset and, thus, to represent an empirical estimate of the upper limit of performance of any algorithm. The between-montage agreement (CREF–ALT) on the other hand served as a heuristic measure of sufficiency of the information in the frontopolar derivation, as compared with the information available in the PSG montage. Box plots were created of the distributions of SEN and PPVs for each of the three methods (EXP, ALT and AUT), and differences were tested using the Kruskal–Wallis test. A significant negative difference between the expert and alternative scoring with respect to a stage-specific metric was interpreted as an indication that the morphology of the $F_{p1}$–$F_{p2}$ signal was insufficient for a reliable identification of that stage. Likewise, a negative difference between the alternative and automatic scoring suggested a limitation of the algorithm. As the latter was the case for stages wake and NREM1 (see Results), we additionally investigated whether the algorithm’s performance was hampered by between-subject variability of features used by FP-STAGER to identify these two stages. To test this hypothesis, the group-based thresholds at node R6 of the algorithm were ‘individualized’, i.e. replaced by values optimized separately for each recording by finding the combination that optimizes the wake/sleep agreement between the algorithm and referent scoring for that particular recording. (This procedure was used only to an analytic tool, and it is ‘not’ part of the algorithm’s training or its regular use.)

In all analyses, sensitivity denoted the proportion of all elements of a referent class that were correctly identified by the algorithm (SEN = (True Positives)/ (True Positives + False Negatives)). The PPV expressed the proportion of all elements assigned to a class by the algorithm that indeed belong to that class (PPV = (True Positives)/(True Positives + False Positives)). Agreement was quantified as a percentage and with Cohen’s kappa statistics, which measures the agreement beyond chance (Cohen, 1960). Kappa values between 0 and 0.20, 0.21 and 0.40, 0.41 and 0.60, 0.61 and 0.80, and 0.81 and 1, respectively, indicated poor, modest, moderate, substantial and excellent agreement (Landis and Koch, 1977).

# RESULTS

## Sleep stages

The pooled epoch-by-epoch agreement between the algorithm (AUT) and referent scoring (IREF, RREF and CREF) is shown in Fig. 2. The agreement was substantial for all pair-wise comparisons in both validation groups, and was (expectedly) higher in the case of CREF. There was no significant difference between the IREF–AUT and RREF–AUT agreement patterns in either the sleep-deprived or well-rested group: generally, REM and solid NREM sleep were identified with high sensitivity and PPV, whereas the detection of wake was slightly less accurate, and that of stage NREM1 considerably less accurate. Wake epochs were most frequently confused with NREM1, and NREM1 epochs were frequently misclassified as REM and NREM2 by the algorithm. The agreement between the algorithm and referent scorer was comparable to the agreement between the referent and alternative scoring, and only slightly lower than the inter–rater agreement for the same data sets and same choice of reference (Table 2; the respective contingency tables are shown in Figs S1 and S2 in the online supplement).

## All-night sleep structure

There was no difference among the referent, independent, alternative and automatic scoring with respect to the estimates of all but one all-night measures (Table 3). The algorithm underestimated latency to NREM3, but the magnitude of this statistically significant difference (maximum: 8 min) nonetheless seemed acceptable for most practical applications. Bland–Altman plots (Fig. 3) showed negligible biases for all variables and clinically acceptable variances for TST, SE, SL and REMT. Standard deviations of the WASO and REML estimates were, however, large in comparison to the range of these parameters, which could limit their utility in practice.

## Cortical arousals

The algorithm detected 1317 of the 1826 arousals marked by the referent scorer, and made 648 false detections ($SEN_{AUT}$ = 72.2%, $PPV_{AUT}$ = 67.0%). The alternative scoring identified 1379 true and 668 false events ($SEN_{ALT}$ = 75.5%, $PPV_{ALT}$ = 67.3%). The independent scorer marked a total of 1860 events ($SEN_{EXP}$ = 68.9%, $PPV_{EXP}$ = 67.6%). The algorithm detected 1031 out of the 1258 consensus arousals (SEN = 82%). The duration of algorithm-marked arousals was in a good agreement with the duration of their counterparts identified by the referent scorer (Fig. 4).

## Sufficiency of information contained in the $F_{p1}–F_{p2}$ derivation

The distributions of SENs and PPVs across the 29 recordings are shown in Fig. 5 for the CREF–EXP, CREF–ALT and CREF–AUT comparisons. The automatic identification of wakefulness was inferior in comparison to the other two methods

($SEN^{Wake}_{CREF-EXP}=86.0\pm9.2\%, SEN^{Wake}_{CREF-ALT}=83.4\pm10.6\%, SEN^{Wake}_{CREF-AUT}=81.7\pm10.1\%$, $P < 0.05$;

$PPV^{Wake}_{CREF-EXP}=88.5\pm9.2\%, PPV^{Wake}_{CREF-ALT}=79.1\pm10.6\%, PPV^{Wake}_{CREF-AUT}=76.3\pm10.1\%$, $P < 0.05$). The large difference in PPV (and a smaller but notable difference in sensitivity)

between the independent scoring of the standard montage and manual scoring of the $F_{p1}$–$F_{p2}$ signal suggests that frontopolar cues (e.g. slow eye movements, theta bursts) were not as reliable indicators of wake/sleep transitions as was the attenuation of occipital alpha in the PSG montage. The comparatively smaller differences between the manual and automatic wake/sleep scoring from the $F_{p1}$–$F_{p2}$ signal were related to between-subject variability of the features used by FP-STAGER to identify wakefulness: when the group-based thresholds at node R6 were replaced with values optimized separately for each subject, the automatic detection of wakefulness became as sensitive and precise as the alternative scoring ($\mathrm{SEN}^{\mathrm{Wake}}_{\mathrm{CREF-AUT(ind)}}=83.1\pm11.2\%, \mathrm{PPV}^{\mathrm{Wake}}_{\mathrm{CREF-AUT(ind)}}=79.7\pm9.9\%$). The detection of stage NREM1 was challenging for all three scoring methods, but the algorithm was the least accurate among them

($\mathrm{SEN}^{\mathrm{NREM1}}_{\mathrm{CREF-EXP}}=67.9\pm10.2\%, \mathrm{SEN}^{\mathrm{NREM1}}_{\mathrm{CREF-ALT}}=62.8\pm9.1\%, \mathrm{SEN}^{\mathrm{NREM1}}_{\mathrm{CREF-AUT}}=56.4\pm12.7\%$, $P < 0.05$;

$\mathrm{PPV}^{\mathrm{NREM1}}_{\mathrm{CREF-EXP}}=55.1\pm8.9\%, \mathrm{PPV}^{\mathrm{NREM1}}_{\mathrm{CREF-ALT}}=52.0\pm9.4\%, \mathrm{PPV}^{\mathrm{NREM1}}_{\mathrm{CREF-AUT}}=46.2\pm12.1\%$, $P < 0.05$). The large difference between the stage-specific REF–EXP and REF–AUT sensitivity and PPV pointed, again, to the inconsistency of features in the $F_{p1}$–$F_{p2}$ signal as a key problem during uneventful transitions from wake through NREM1 to stable NREM sleep (stage NREM2). The differences in NREM1 sensitivity and PPV between the alternative and automatic scoring were also, to a large degree, annulled by the individualization of the thresholds at node R6 ($\mathrm{SEN}^{\mathrm{NREM1}}_{\mathrm{CREF-AUT(ind)}}=60.9\pm8.8\%, \mathrm{PPV}^{\mathrm{NREM1}}_{\mathrm{CREF-AUT(ind)}}=51.3\pm10.3\%$). Stages NREM2 and NREM3 were identified accurately (average SEN and PPV in excess of 85%), and the respective stage-specific measures did not significantly differ among the three methods. Automatic detection of REM sleep was also accurate, but its visual identification from the forehead was significantly hampered by difficulties with assessment of the muscle tone from the $F_{p1}$–$F_{p2}$ signal

($\mathrm{SEN}^{\mathrm{REM}}_{\mathrm{CREF-EXP}}=92.1\pm12.1\%, \mathrm{SEN}^{\mathrm{REM}}_{\mathrm{CREF-ALT}}=85.0\pm11.3\%, \mathrm{SEN}^{\mathrm{REM}}_{\mathrm{CREF-AUT}}=90.9\pm10.4\%$, $P < 0.05$;

$\mathrm{PPV}^{\mathrm{REM}}_{\mathrm{CREF-EXP}}=92.3\pm11.5\%, \mathrm{PPV}^{\mathrm{REM}}_{\mathrm{CREF-ALT}}=85.7\pm9.2\%, \mathrm{PPV}^{\mathrm{REM}}_{\mathrm{CREF-AUT}}=90.4\pm8.8\%$, $P < 0.05$).

## DISCUSSION

The agreement between the manual reference and automatic staging of the single frontopolar derivation was similar to the inter–rater and between-montage agreement for the same data, and lied within the range of inter–rater agreements reported for experts from the same laboratory and samples composed of healthy subjects (Anderer *et al.*, 2005; Berthomier *et al.*, 2007; Schaltenbrand *et al.*, 1996; Shambroom *et al.*, 2012). The accuracy of the automatic detection of arousals was comparable to the inter–rater and intra–rater concordances reported in the literature (De Carli *et al.*, 1999; Whitney *et al.*, 1998). Furthermore, the algorithm's performance was consistent and satisfactory across the individual recordings, and its estimates of the common descriptors of all-night sleep architecture were in close agreement with the referent values. The results, therefore, affirm

automatic scoring of sleep stages and EEG arousals from the $F_{p1}$–$F_{p2}$ derivation as an accurate method for assessment of sleep architecture in healthy adults.

The accuracy of both visual and automatic scoring of the $F_{p1}$–$F_{p2}$ derivation was primarily determined by the morphology of the signal, and varied significantly by stage. Stages NREM2 and NREM3, and NREM1 epochs associated with arousals were identified accurately because the detection rested on waveforms that, for the most part, coincided in the frontopolar and scalp EEG derivations during the respective stages (e.g. EEG arousals, spindles, K-complexes or delta waves). REM sleep detection was challenging for a human eye because the frontopolar EEG is similar in wakefulness, REM and light NREM sleep, while changes in the generally meager EMG activity of the frontalis muscle were difficult to distinguish from variations in the background beta EEG. The algorithm, nonetheless, accurately delineated REM from wake or NREM1 epochs by contrasting the average beta power to the power in the sigma and EMG bands (an approach similar to, e.g. Berthomier *et al.*, 2007; Shambroom *et al.*, 2012). Identification of wake and NREM1 epochs was, on the other hand, challenging for both automatic and visual scoring of the $F_{p1}$–$F_{p2}$ signal because of its features (or lack thereof): the absence of alpha rhythm, inconsistent relationship between the attenuation of the posterior alpha and frontopolar waveforms suggestive of sleep onset, and early occurrence of slower spindles (12–14 Hz) that had no counterparts in the scalp EEG (for details on different classes of spindles see, e.g. Huupponen *et al.*, 2008). Differences between the automatic and visual scoring of the $F_{p1}$–$F_{p2}$ signal resulted mostly from the algorithm's inability to handle between-subject variability of certain features, and were mitigable through individualization of pertinent thresholds in the algorithm's decision tree. Beyond its use as an analytic tool, the individualization is, however, of limited practical value, as it requires at least one baseline PSG study per subject, which may not be feasible or economical in all applications. Future work, thus, needs to concentrate on clarifying the relationship between the standard (parieto-occipital) and frontopolar hallmarks of wake, NREM1 and early NREM2 stage on a representative sample of subjects.

Table 4 compares FP-STAGER with the performance of other algorithms for staging sleep from simplified recording configurations, such as the ASEEGA system (Berthomier *et al.*, 2007), algorithms that operate on EOG channels (Levendowski *et al.*, 2012; Virkkala *et al.*, 2007, 2008), and the forehead-worn ZEO and ARES systems (Popovic *et al.*, 2008; Shambroom *et al.*, 2012). For a fair comparison to algorithms that were trained and validated against different scorers (e.g. ZEO) or a consensus reference created by eliminating the epochs where the scorers disagreed (e.g. ASSEGA), FP-STAGER was also compared with the independent expert and the subset of epochs where the referent and independent scorer agreed. FP-STAGER performed comparably to algorithms validated with conventional PSG equipment, and clearly outperformed the forehead-worn ZEO and ARES systems or the Alive Heart Monitor that recorded from a single differential EOG lead (Virkkala *et al.*, 2008). The performance of the forehead-worn systems was, therefore, likely hindered by properties of the devices, such as the thermal noise of utilized dry sensors and/or narrow bandwidth of the amplifiers (e.g. 2–47 Hz for the ZEO). FP-STAGER performed similarly to the algorithms for staging of the full montage that have been validated by an epoch-by-epoch comparison to the reference (Anderer *et al.*, 2005; Pitman *et al.*, 2004; Schaltenbrand *et al.*,

1996; Svetnik *et al.*, 2007). Differentiation of wakefulness from light sleep was problematic for some of these algorithms too, even though they operated on the scalp derivations, where alpha rhythm is easily captured (Pittman *et al.*, 2004; Svetnik *et al.*, 2007).

Wrist actigraphs have been used for objective, unobtrusive and inexpensive estimation of wake/sleep patterns in ambulatory settings for over 50 years; thus, a considerable audience might be interested in comparative advantages and disadvantages of the automatic scoring of frontopolar EEG. Although the two methods were not directly compared in this study, the sensitivity and PPV of the automatic wake/sleep scoring were higher than the values typically reported for actigraphy in healthy individuals (Ancoli-Israel *et al.*, 2003; Martin and Hakim, 2011). Considering its different sensitivity to solid and transient wakefulness, automatic scoring of the $F_{p1}$–$F_{p2}$ derivation is likely to be more accurate than actigraphy in clinical populations with lots of wake periods during which subjects lie still (e.g. insomnia); some support for this assertion is provided by a study of 26 volunteers that found actigraphy inferior to the forehead-worn ZEO system in subjects with low sleep efficiency (Shambroom *et al.*, 2012). However, the main advantage of the automatic scoring of frontopolar EEG over actigraphy is its ability to accurately identify restorative sleep (NREM2, NREM3 and REM) and cortical arousals (in addition to the wake/sleep classification).

The algorithm described herein has four major limitations. First, it was developed and validated on a relatively small population of young, healthy subjects whose recordings mostly consisted of long periods of uninterrupted sleep and contained relatively few arousals. Further studies on elderly subjects and various clinical populations are, therefore, warranted before the algorithm is eventually introduced into clinical practice. Second, the algorithm was validated on the data acquired in a sleep laboratory and containing relatively few artefacts (0.7% of all epochs). Although the artefacts were detected with high sensitivity and precision in this sample, the true ability of the algorithm to deal with artefacts will be known only after a study performed outside of a dedicated facility, where artefacts can occupy as much as 10–15% of the recording. Third, the algorithm's thresholds as reported herein were optimized for the utilized sensors and hardware, and may need to be adjusted prior to eventual use of the algorithm with sensors or recorders that have different noise spectra, amplification gains or channel bandwidths. Finally, the reported thresholds reflect the scoring habits of our sleep expert(s), and might be a suboptimal combination if used by researchers with significantly different scoring styles. It is, therefore, advisable that the thresholds be adjusted against a representative set of recordings scored by a local referent scorer prior to any large-scale deployment of the algorithm by other groups.

In summary, the study affirmed the utility of automatic scoring of sleep stages from a single frontopolar derivation, and demonstrated that its accuracy is primarily limited by physiological differences between the frontopolar and scalp EEG. The method could be used for assessment of sleep quality of healthy individuals in domicile or operational settings (e.g. shift workers, deployed military servicemen). Upon a successful validation in relevant clinical populations, the algorithm or its improved variants could also be used in in-home diagnostic sleep studies, clinical trials or epidemiological studies.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This study was supported by the US Defense Advanced Research Projects Agency (#W31P4Q-08-C-0123). The views, opinions and findings in this article are those of the authors, and should not be interpreted as representing the official views or policies of the Defense Advanced Research Project Agency or the Department of Defense.

## REFERENCES

Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. Sleep. 2003; 26:342–392. [PubMed: 12749557]

Anderer P, Gruber G, Prapatics S, et al. An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24 × 7 utilizing the SIESTA database. Neuropsychobiology. 2005; 51:115–133. [PubMed: 15838184]

Berka C, Levendowski D, Cvetinovic M, et al. Real-time analysis of EEG indices of alertness, cognition and memory acquired with a wireless EEG headset. Int. J Hum. Comput. Interact. 2004; 17:151–170.

Berthomier C, Drouot X, Herman-Stoica M, et al. Automatic analysis of single channel sleep EEG: validation in healthy individuals. Sleep. 2007; 30:1587–1595. [PubMed: 18041491]

Cohen J. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 1960; 20:37–46.

De Carli F, Nobili L, Gelcich P, Ferrillo F. A method for the automatic detection of arousals during sleep. Sleep. 1999; 22:561–572. [PubMed: 10450591]

Fischer Y, Junge-Hulsing B, Rettinger G, Panis A. The use of an ambulatory, automatic sleep recording device in the evaluation of primary snoring and sleep apnea. Clin. Otolaryngol. 2004; 29:18–23. [PubMed: 14961847]

Huupponen E, Kulkas A, Tenhunen M, Saastamoinen A, Hasan J, Himanen SL. Diffuse sleep spindles show similar frequency in central and frontopolar positions. J. Neurosci. Methods. 2008; 172:54–59. [PubMed: 18482770]

Iber, C.; Anconi-Israel, S.; Chesson, AL.; Quan, SF. The AASM Manual for the Scoring of Sleep and Associated Events. Westchester, IL: American Academy of Sleep Medicine; 2007. for the American Academy of Sleep Medicine.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–174. [PubMed: 843571]

Levendowski DJ, Popovic D, Berka C, Westbrook PR. Retrospective cross-validation of automated sleep staging using electroocular recording in patients with and without sleep disordered breathing. Int. Arch. Med. 2012; 5:21. [PubMed: 22726270]

Martin JL, Hakim AD. Wrist actigraphy. Chest. 2011; 139:1514–1527. [PubMed: 21652563]

Pittman S-D, MacDonald M, Fogel R-B, et al. Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing. Sleep. 2004; 27:1394–1403. [PubMed: 15586793]

Popovic D, Levendowski D, Ayappa I, et al. Accuracy of automated sleep staging using signals from a single forehead site. Sleep. 2008:31. (Abstract supplement): A332.

Schaltenbrand N, Lengelle R, Toussaint M, et al. Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients. Sleep. 1996; 19:26–35. [PubMed: 8650459]

Shambroom JR, Fabregas SE, Johnstone J. Validation of an automated wireless system to monitor sleep in healthy adults. J. Sleep Res. 2012; 21:221–230. [PubMed: 21859438]

Svetnik V, Ma J, Soper KA, et al. Evaluation of automated and semi-automated scoring of polysomnographic recordings from a clinical trial using zolpidem in the treatment of insomnia. Sleep. 2007; 30:1562–1574. [PubMed: 18041489]

Uchida S, Maloney T, Feinberg I. Sigma (12–16 Hz) and beta (20–28 Hz) EEG discriminate NREM and REM sleep. Brain Res. 1994; 659:243–248. [PubMed: 7820669]

Virkkala J, Hasan J, Varri A, Himanen S-L, Muller K. Automatic sleep stage classification using two channel electro-oculography. J. Neurosci. Methods. 2007; 166:109–115. [PubMed: 17681382]

Virkkala J, Velin R, Himanen S-L, Varri A, Muller K, Hasan J. Automatic sleep stage classification using two facial electrodes. Conf. Proc. IEEE Eng. Med. Biol. Soc. 2008; 30:1632–1646.

Whitney CW, Gottlieb DJ, Redline S, et al. Reliability of scoring respiratory disturbance indices and sleep staging. Sleep. 1998; 21:749–757. [PubMed: 11286351]

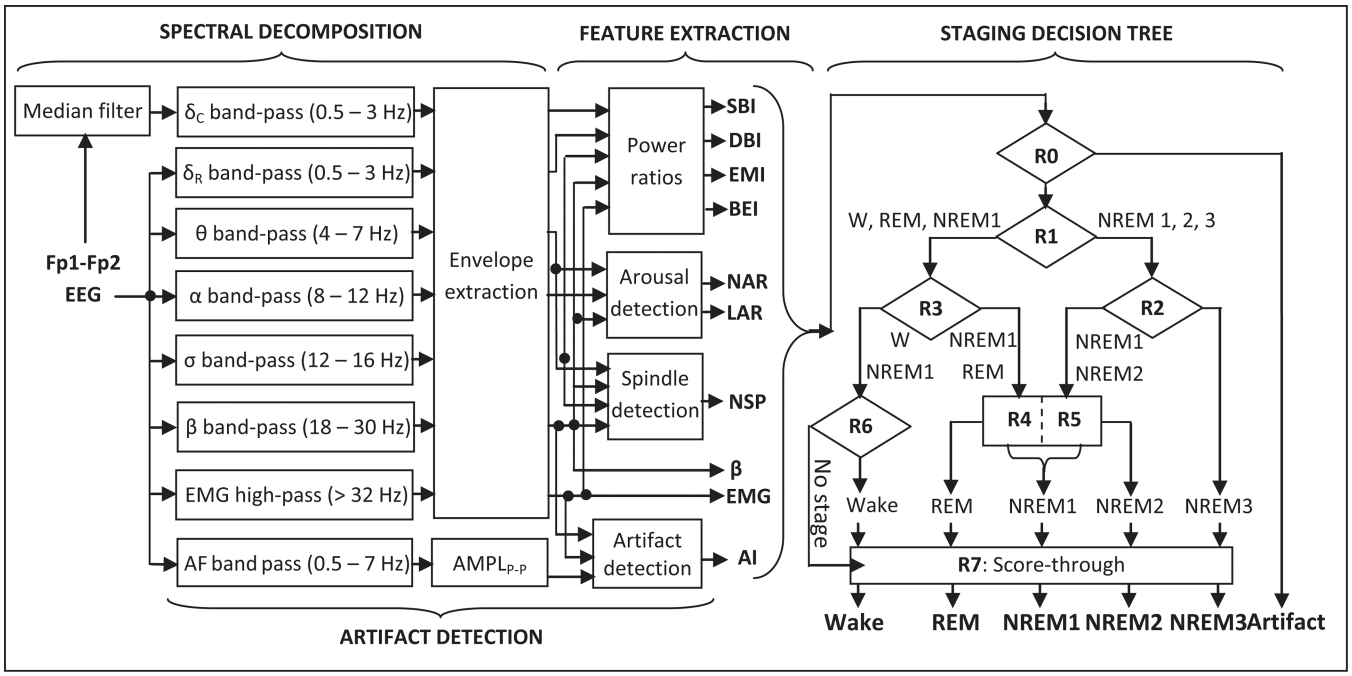


**Figure 1.**
Block diagram of the FP-STAGER algorithm.

## SLEEP-DEPRIVED GROUP

### INITIAL REFERENT SCORING vs. ALGORITHM

| IREF \ AUT | W | N1 | N2 | N3 | R | A | Sum |
|---|---|---|---|---|---|---|---|
| Wake | **273** | 29 | 19 | 2 | 20 | 4 | 347 |
| NREM1 | 55 | **183** | 48 | 25 | 43 | 3 | 357 |
| NREM2 | 34 | 63 | **891** | 97 | 18 | 3 | 1106 |
| NREM3 | 4 | 21 | 94 | **1021** | 0 | 0 | 1140 |
| REM | 0 | 19 | 36 | 1 | **491** | 0 | 547 |
| Artifact | 3 | 0 | 0 | 0 | 0 | **11** | 14 |
| Sum | 369 | 315 | 1088 | 1146 | 572 | 21 | **3511** |
| SEN(%) | **78.7** | **51.3** | **80.6** | **89.6** | **89.8** | **78.6** | |
| PPV(%) | **74.0** | **58.1** | **81.9** | **89.1** | **85.8** | **52.4** | |
| Agreement (%) | **81.7** | | Cohen s Kappa | | | | **0.76** |

### REPEATED REFERENT SCORING vs. ALGORITHM

| RREF \ AUT | W | N1 | N2 | N3 | R | A | Sum |
|---|---|---|---|---|---|---|---|
| Wake | **267** | 26 | 18 | 1 | 19 | 4 | 335 |
| NREM1 | 53 | **185** | 43 | 23 | 38 | 4 | 346 |
| NREM2 | 35 | 65 | **893** | 87 | 31 | 2 | 1113 |
| NREM3 | 5 | 17 | 92 | **1035** | 0 | 0 | 1149 |
| REM | 6 | 22 | 42 | 0 | **484** | 0 | 554 |
| Artifact | 3 | 0 | 0 | 0 | 0 | **11** | 14 |
| Sum | 369 | 315 | 1088 | 1146 | 572 | 21 | **3511** |
| SEN(%) | **79.7** | **54.3** | **80.4** | **90.1** | **87.4** | **78.6** | |
| PPV(%) | **72.4** | **58.8** | **82.1** | **90.3** | **84.6** | **52.4** | |
| Agreement (%) | **82.1** | | Cohen s Kappa | | | | **0.76** |

### CONSOLIDATED REFERENCE vs. ALGORITHM

| CREF \ AUT | W | N1 | N2 | N3 | R | A | Sum |
|---|---|---|---|---|---|---|---|
| Wake | **256** | 23 | 14 | 1 | 15 | 2 | 311 |
| NREM1 | 40 | **151** | 20 | 11 | 21 | 1 | 244 |
| NREM2 | 22 | 53 | **824** | 69 | 26 | 1 | 995 |
| NREM3 | 1 | 12 | 83 | **952** | 0 | 0 | 1048 |
| REM | 0 | 17 | 32 | 0 | **449** | 0 | 498 |
| Artifact | 3 | 0 | 0 | 0 | 0 | **11** | 14 |
| Sum | 322 | 256 | 973 | 1033 | 511 | 15 | **3110** |
| SEN(%) | **82.3** | **61.9** | **82.8** | **90.8** | **90.2** | **78.6** | |
| PPV(%) | **79.5** | **59.0** | **84.7** | **92.2** | **87.9** | **73.3** | |
| Agreement (%) | **85.0** | | Cohen s Kappa | | | | **0.80** |

## WELL-RESTED GROUP

### INITIAL REFERENT SCORING vs. ALGORITHM

| IREF \ AUT | W | N1 | N2 | N3 | R | A | Sum |
|---|---|---|---|---|---|---|---|
| Wake | **1048** | 146 | 45 | 4 | 109 | 13 | 1365 |
| NREM1 | 260 | **834** | 227 | 22 | 280 | 8 | 1631 |
| NREM2 | 20 | 561 | **5219** | 261 | 119 | 3 | 6183 |
| NREM3 | 15 | 55 | 402 | **2808** | 0 | 0 | 3280 |
| REM | 129 | 253 | 85 | 25 | **2861** | 0 | 3353 |
| Artifact | 12 | 3 | 0 | 0 | 0 | **107** | 122 |
| Sum | 1484 | 1852 | 5978 | 3120 | 3369 | 131 | **15934** |
| SEN(%) | **76.8** | **51.1** | **84.4** | **85.6** | **85.3** | **87.7** | |
| PPV(%) | **70.6** | **45.0** | **87.3** | **90.0** | **84.9** | **81.7** | |
| Agreement (%) | **80.8** | | Cohen s Kappa | | | | **0.74** |

### REPEATED REFERENT SCORING vs. ALGORITHM

| RREF \ AUT | W | N1 | N2 | N3 | R | A | Sum |
|---|---|---|---|---|---|---|---|
| Wake | **1046** | 147 | 26 | 3 | 102 | 11 | 1345 |
| NREM1 | 252 | **795** | 172 | 7 | 305 | 10 | 1541 |
| NREM2 | 35 | 599 | **5390** | 292 | 106 | 4 | 6426 |
| NREM3 | 21 | 92 | 346 | **2806** | 0 | 0 | 3265 |
| REM | 118 | 216 | 44 | 12 | **2846** | 0 | 3236 |
| Artifact | 12 | 3 | 0 | 0 | 0 | **106** | 121 |
| Sum | 1484 | 1852 | 5978 | 3120 | 3369 | 131 | **15934** |
| SEN(%) | **77.8** | **51.6** | **83.9** | **85.9** | **87.9** | **87.6** | |
| PPV(%) | **70.5** | **43.0** | **90.1** | **89.9** | **84.5** | **80.9** | |
| Agreement (%) | **81.5** | | Cohen s Kappa | | | | **0.75** |

### CONSOLIDATED REFERENCE vs. ALGORITHM

| CREF \ AUT | W | N1 | N2 | N3 | R | A | Sum |
|---|---|---|---|---|---|---|---|
| Wake | **993** | 130 | 14 | 3 | 77 | 10 | 1216 |
| NREM1 | 172 | **577** | 159 | 4 | 145 | 10 | 1057 |
| NREM2 | 18 | 476 | **4929** | 244 | 87 | 3 | 5757 |
| NREM3 | 9 | 43 | 321 | **2761** | 0 | 0 | 3134 |
| REM | 114 | 116 | 11 | 6 | **2796** | 0 | 3043 |
| Artifact | 12 | 3 | 0 | 0 | 0 | **106** | 121 |
| Sum | 1316 | 1334 | 5452 | 3018 | 3105 | 129 | **14328** |
| SEN(%) | **81.6** | **54.6** | **85.6** | **88.1** | **91.9** | **87.6** | |
| PPV(%) | **75.4** | **43.2** | **90.9** | **91.5** | **90.0** | **82.2** | |
| Agreement (%) | **84.9** | | Cohen s Kappa | | | | **0.80** |

**Figure 2.**

Agreement between the manual scoring reference and automatic scoring of the $F_{p1}$–$F_{p2}$ signal in the groups of sleep-deprived and well-rested volunteers. The referent scoring of the polysomnography montage is in rows, and the automatic scoring of the frontopolar derivation is in columns.

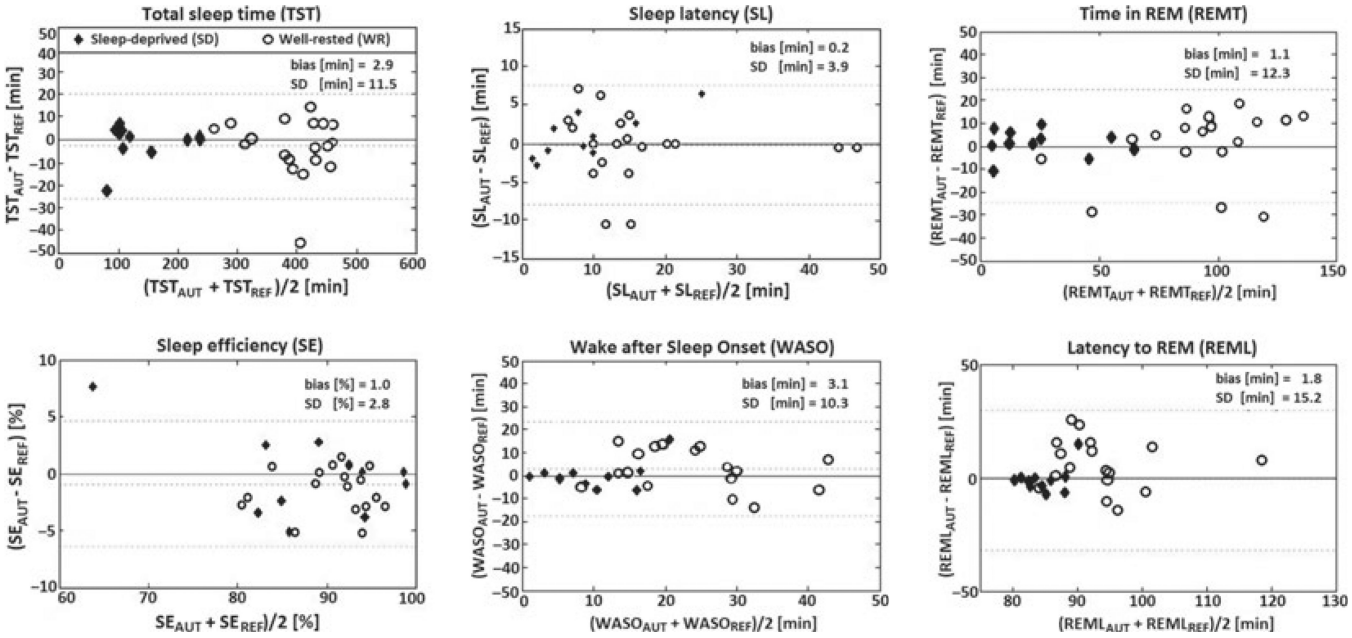


**Figure 3.**

Bland–Altman plots of selected variables calculated from the automatic (AUT) and referent scoring (REF). Biases, standard deviations (SD) and intra-class correlation coefficients (ICCs) are reported for pooled data from both validation groups, but the groups are differentiated graphically because of the large between-group differences in ranges of some variables. Dotted lines mark biases and limits of agreements (± 1.96*SD).

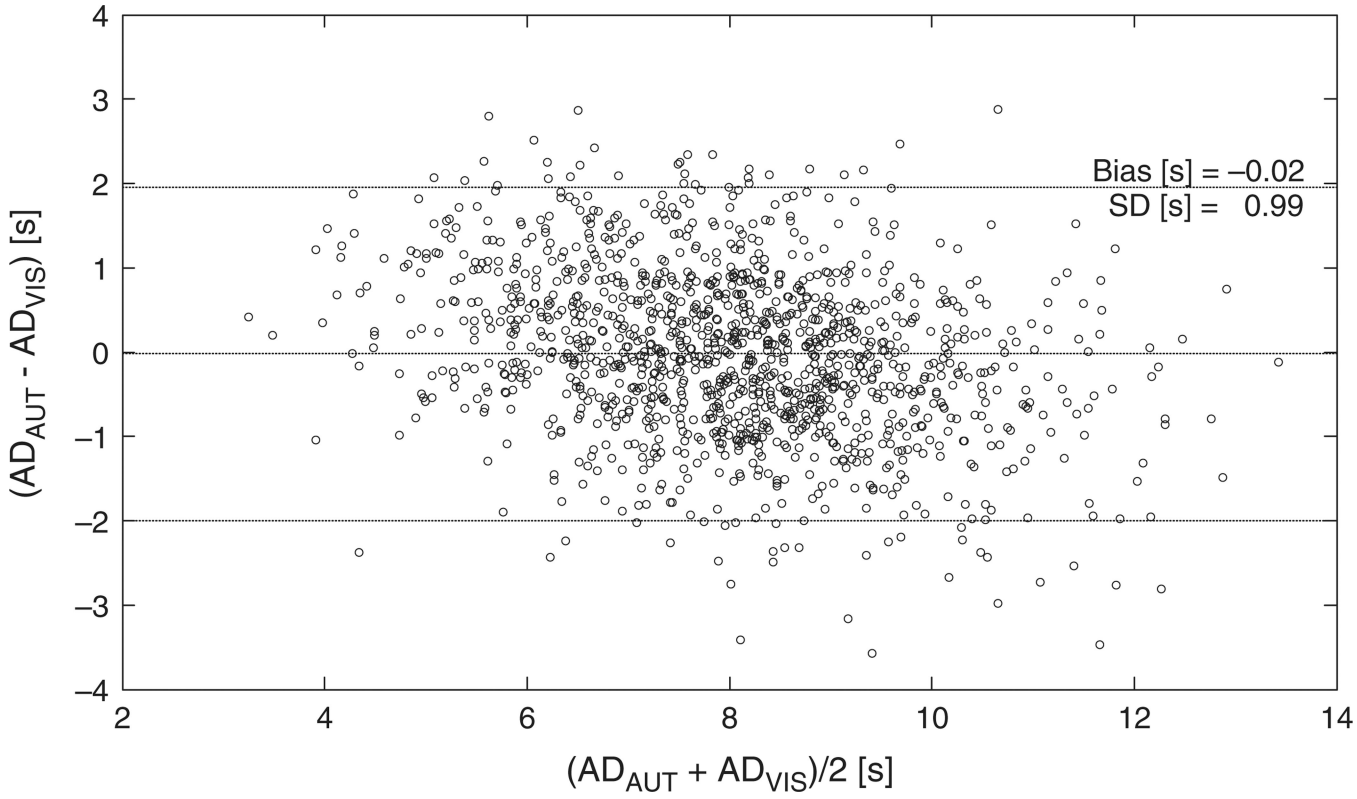


**Figure 4.**

Bland–Altman plot of the cortical arousals determined by the algorithm ($AD_{AUT}$) versus duration of arousals marked by the referent scorer ($AD_{REF}$).

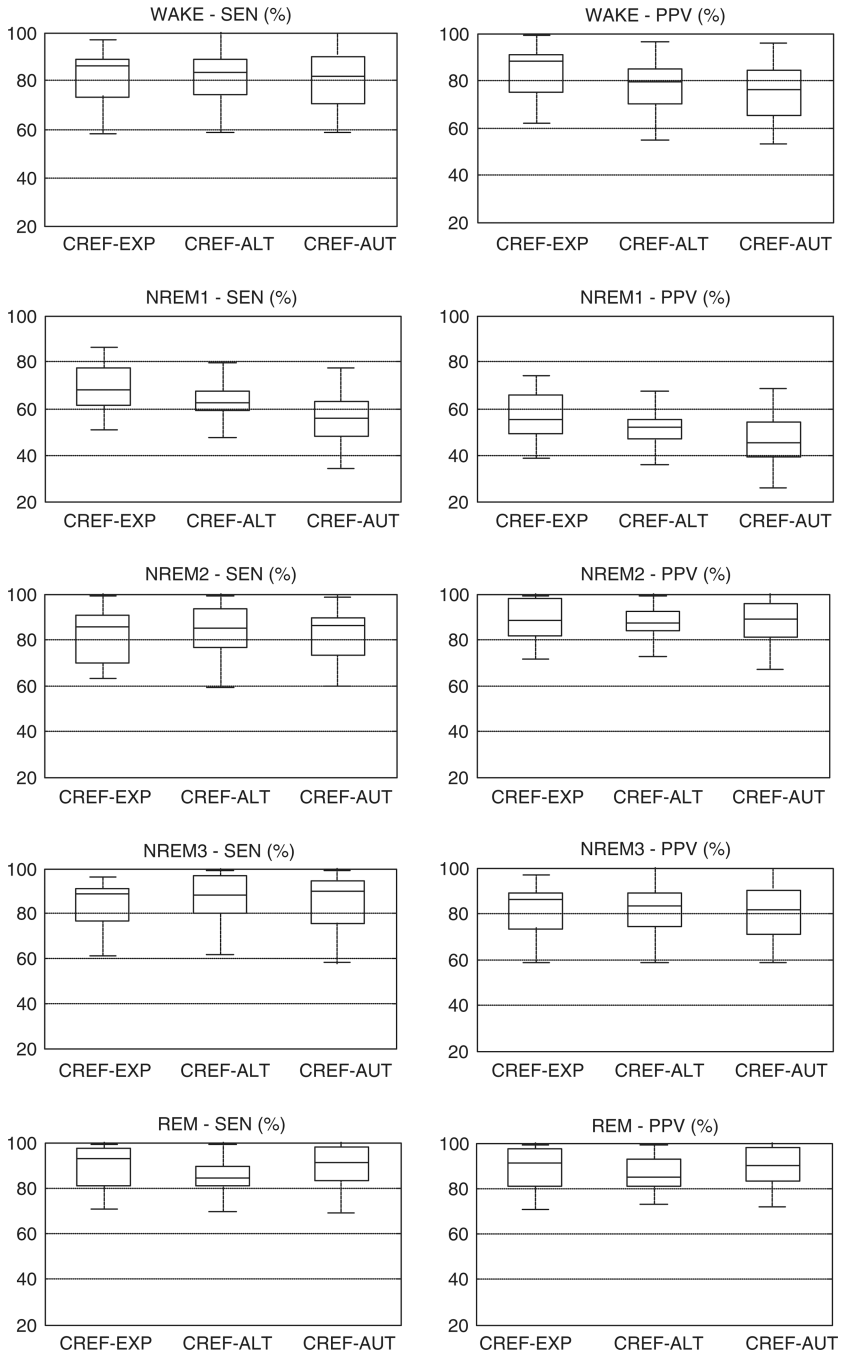


**Figure 5.**

Distributions of the stage-specific sensitivity (SEN) and positive predictive value (PPV) of the expert (EXP), alternative (ALT) and automatic scoring (AUT) as compared with the consolidated reference (CREF).

## Table 1

Demographic characteristics and sleep structure in the training and validation groups (all fields show mean values with the range in brackets)

| | Training group (sleep-deprived) | Validation group 1 (sleep-deprived) | Validation group 2 (well-rested) |
|---|---|---|---|
| Number of subjects (m/f) | 10 (6/4) | 11 (6/5) | 18 (8/10) |
| Type of data | Daytime naps | Daytime naps | Nocturnal sleep |
| Age (years) | 27 (20–63) | 25 (18–40) | 27 (18–45) |
| RT (min) | 160 (120–197) | 175 (112–260) | 450 (314–502)* |
| RDI ($h^{-1}$) | 1.5 (0–3) | 1.8 (0–4) | 1.9 (0–4) |
| Arousal index ($h^{-1}$) | 11 (1–30) | 12 (2–25) | 10 (1–20) |
| Wake (% of RT) | 10 (3–21) | 11 (2–26) | 9 (2–18) |
| REM (% of RT) | 15 (2–29) | 14 (1–26) | 19 (8–26) |
| NREM1 (% of RT) | 9 (4–24) | 10 (1–23) | 11 (5–21) |
| NREM2 (% of RT) | 36 (22–45) | 35 (10–53) | 41 (34–50) |
| NREM3 (% of RT) | 30 (2–45) | 30 (3–53) | 20 (14–41)† |

*$P < 0.0001$.

†$P < 0.1$, marginally significant.

NREM, non-rapid eye movement; RDI, respiratory disturbance index; REM, rapid eye movement; RT, recording time.

## Table 2

Inter–rater and between-montage agreement

| Group | Sleep-deprived | Well-rested |
|---|---|---|
| Inter–rater % agreement (kappa) | | |
| IREF–EXP | 83.8 (0.78) | 83.2 (0.77) |
| RREF–EXP | 83.7 (0.78) | 82.8 (0.77) |
| CREF–EXP | 88.5 (0.84) | 86.2 (0.81) |
| Between-montage % agreement (kappa) | | |
| IREF–ALT | 82.9 (0.76) | 80.4 (0.74) |
| RREF–ALT | 81.5 (0.75) | 80.5 (0.74) |
| CREF–ALT | 85.4 (0.80) | 83.8 (0.78) |

ALT, alternative scoring; CREF, consolidated reference; EXP, independent expert scoring; IREF, initial referent scoring; RREF, repeated referent scoring.

**Table 3**

Summary sleep measures

| | IREF | RREF | EXP | ALT | AUT | Significance |
|---|---|---|---|---|---|---|
| TST (min) | 305.7 ± 12.8 | 306.3 ± 12.9 | 305.5 ± 13.2 | 304.5 ± 11.7 | 303.3 ± 12.0 | $\chi^2 = 1.74$, df = 4, ns |
| SE (%) | 90.8 ± 3.6 | 91.0 ± 3.7 | 90.5 ± 3.6 | 89.9 ± 3.4 | 89.4 ± 3.4 | $F_{4,28} = 1.89$, ns |
| SL (min) | 11.2 ± 3.1 | 10.9 ± 2.9 | 10.6 ± 2.8 | 12.3 ± 3.0 | 12.9 ± 3.5 | $F_{4,28} = 2.25$, ns |
| WASO (min) | 18.3 ± 8.8 | 18.0 ± 9.0 | 19.1 ± 9.4 | 18.4 ± 8.2 | 19.0 ± 9.1 | $\chi^2 = 1.42$, df = 4, ns |
| AI ($h^{-1}$) | 12.4 ± 3.4 | 12.9 ± 3.2 | 12.7 ± 3.3 | 12.6 ± 2.9 | 13.4 ± 3.1 | $F_{4,28} = 1.07$, ns |
| REMT (min) | 67.2 ± 9.8 | 65.4 ± 10.4 | 67.7 ± 10.3 | 66.9 ± 9.7 | 67.9 ± 9.5 | $\chi^2 = 2.31$, df = 4, ns |
| REML (min) | 88.7 ± 12.2 | 89.4 ± 13.1 | 86.0 ± 11.7 | 89.1 ± 12.8 | 87.3 ± 13.4 | $F_{4,28} = 0.94$, ns |
| SWST (min) | 76.2 ± 13.4 | 76.1 ± 13.2 | 75.3 ± 12.1 | 76.2 ± 12.8 | 73.6 ± 12.4 | $\chi^2 = 2.23$, df = 4, ns |
| SWSL (min) | 50.5 ± 5.7 | 51.0 ± 5.9 | 49.2 ± 6.2 | 47.6 ± 5.5 | 46.1 ± 6.1 | $F_{4,28} = 4.29$, $P < 0.01$* |

*Automatic estimates significantly different from the referent (IREF and RREF) and independent expert scoring.

AI, arousal index; ALT, alternative scoring; AUT, automatic scoring; EXP, independent expert scoring; IREF, initial referent scoring; REML, latency to rapid eye movement; REMT, time spent in rapid eye movement; RREF, repeated referent scoring; SE, sleep efficiency; SL, sleep latency; SWSL, latency to non-rapid eye movement; SWST, time spent in non-rapid eye movement; TST, total sleep time; WASO, wake time after sleep onset.

**Table 4**

Performance of FP-STAGER and similar algorithms that operate on simplified montages

| | Algorithms | | | Performance (Cohen's kappa) | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Algorithm** | **Signal** | **Study characteristics** | **W** | **N1** | **N2** | **N3** | **R** | **All** |
| FP-STAGER | Summary results for all 29 subjects | $F_{p1}$–$F_{p2}$ | No. stages: 5; reference: single scorer (REF) | 0.72 | 0.43 | 0.77 | 0.85 | 0.82 | 0.75 |
| | | | No. stages: 5; reference: single scorer (EXP) | 0.69 | 0.34 | 0.72 | 0.82 | 0.79 | 0.71 |
| | | | No. stages: 5; reference: consensus (REF and EXP) | 0.74 | 0.32 | 0.76 | 0.86 | 0.83 | 0.76 |
| | | | No. stages: 4; reference: consensus (REF and EXP) | 0.74 | 0.80 | | 0.86 | 0.83 | 0.79 |
| | | | No. stages: 3; reference: consensus (REF and EXP) | 0.74 | | 0.82 | | 0.83 | 0.81 |
| Other algorithms | ASEEGA | Cz–Pz | Sample: 15 healthy subjects; no. stages: 5; reference: consensus scoring | 0.83 | 0.19 | 0.73 | 0.80 | 0.83 | 0.76 |
| | EOG-2ch | $E_1$–$M_2$ $E_2$–$M_1$ | Sample: 131 healthy subjects and patients with obstructive sleep apnoea; no. stages: 5; reference: single scorer | 0.69 | 0.24 | 0.64 | 0.70 | 0.67 | 0.62 |
| | EOG-1ch | $E_1$–$E_2$ | Sample: 131 healthy subjects and patients with obstructive sleep apnoea; no. stages: 4; reference: single scorer | 0.62 | 0.54 | | 0.70 | 0.56 | 0.59 |
| | ZEO | $F_{p1}$–$F_{p2}$ | Sample: 26 healthy subjects; no. stages: 4; reference: consensus scoring | 0.69 | 0.69 | | 0.66 | 0.75 | 0.70 |
| | ARES | $F_{p1}$–$F_{p2}$ | Sample: 20 healthy subjects and patients with obstructive sleep apnoea; no. stages: 3; reference: consensus scoring | 0.62 | | 0.67 | | 0.65 | 0.64 |

EOG, electrooculogram; EXP, independent expert scoring; REF, referent scoring.