Comparison of computerized methods for rapid eye movement sleep without atonia detection

Cesari, Matteo; Christensen, Julie Anja Engelhard; Kempfner, Lykke; Olesen, Alexander Neergaard; Mayer, Geert; Kesper, Karl; Oertel, Wolfgang H; Sixel-Döring, Friederike; Trenkwalder, Claudia; Sørensen, Helge Bjarup Dissing

Published in:
Sleep

Link to article, DOI:
10.1093/sleep/zsy133

Publication date:
2018

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Comparison of computerized methods for rapid eye movement sleep without atonia detection

Matteo Cesari¹*, Julie A. E. Christensen¹², Lykke Kempfner², Alexander N. Olesen¹³, Geert Mayer⁴, Karl Keser⁵, Wolfgang H. Oertel⁴, Friederike Sixel-Döring⁶, Claudia Trenkwalder⁶, Helge B. D. Sorensen¹ and Poul Jennum²

¹Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark, ²Department of Clinical Neurophysiology, Danish Center for Sleep Medicine, Rigshospitalet Glostrup, Denmark, ³Stanford Center for Sleep Medicine, Stanford University, Stanford, CA, ⁴Department of Neurology, Philipps University Marburg, Marburg, Germany, ⁵Department of Internal Medicine, Section Respiratory Diseases, Philipps University Marburg, Marburg, Germany and ⁶Paracelsus-Elena-Klinik, Kassel, Germany

Work Performed: Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark; Department of Clinical Neurophysiology, Danish Center for Sleep Medicine, Rigshospitalet Glostrup, Denmark

*Corresponding author. Matteo Cesari, Department of Electrical Engineering, Technical University of Denmark, Ørsteds Plads, Building 349, 2800 Kgs. Lyngby, Denmark. Email: maces@elektro.dtu.dk.

Abstract

Rapid eye movement (REM) sleep without atonia detection is a prerequisite for diagnosis of REM sleep behavior disorder (RBD). As the visual gold standard method is time-consuming and subjective, several automated methods have been proposed. This study aims to compare their performances: The REM atonia index (RAI), the supra-threshold-REM-activity metric, the Frandsen index, the short/long muscle activity indices, and the Kempfner index algorithms were applied to 27 healthy control participants (C), 25 patients with Parkinson’s disease (PD) without RBD (PD-RBD), 29 patients with PD and RBD (PD + RBD), 29 idiopathic patients with RBD, and 36 patients with periodic limb movement disorder (PLMD). The indices were calculated in various configurations: (1) considering all muscle activities; (2) excluding the ones related to arousals; (3) excluding the ones during apnea events; (4) excluding the ones before and after apnea events; (5) combining configurations 2 and 3; and (6) combining configurations 2 and 4. For each of these configurations, the discrimination capability of the indices was tested for the following comparisons: (1) (C, PD-RBD, PLMD) vs (PD + RBD, RBD); (2) C vs RBD; (3) PLMD vs RBD; (4) C vs PD-RBD; (5) C vs PLMD; (6) PD-RBD vs PD + RBD; and (7) C vs PLMD vs RBD. Results showed varying methods’ performances across the different configurations and comparisons, making it impossible to identify the optimal method and suggesting the need of further improvements. Nevertheless, RAI seems the most sensible one for RBD detection. Moreover, apnea and arousal-related movements seem not to influence the algorithms’ performances in patients’ classification.

Statement of Significance

Our study is the first one to provide a comparison of five different computerized methods for rapid eye movement (REM) sleep without atonia (RSWA) detection in a cohort including control participants, patients suffering from Parkinson’s disease (PD) without REM sleep behavior disorder (RBD), patients with PD and RBD, patients suffering from idiopathic RBD, and patients with periodic limb movement disorder. We analyze how well the different methods can classify different participant groups based on the automatically detected RSWA level. Moreover, we analyze the influence of movements related to respiratory events and arousals. By discussing strengths and weaknesses of the computerized methods, we believe that this study will help researchers in future development of new automated RSWA detection methods.

Key Words: automated method; computerized analysis; electromyography; Parkinson’s disease; polysomnography; REM sleep behavior disorder; REM sleep without atonia
Introduction
Rapid eye movement (REM) sleep behavior disorder (RBD) is a parasomnia characterized by loss of muscle atonia during REM sleep and a history of recurrent dream enactment [1]. Idiopathic RBD is considered by far the strongest predictor of α-synucleinopathies, including Parkinson’s disease (PD) [2]. This is supported by several follow-up studies showing that patients with RBD are at great risk of developing PD [3, 4]. For example, a recent multicohort study showed that in 279 patients with RBD, 33.3 per cent of them had developed a neurodegenerative disease after 4 years and 41 per cent of them were at risk of developing it after 5 years [3]. Another study showed an 81 per cent conversion after 16 years [4]. Furthermore, neuropathological studies of patients with RBD showed presence of Lewy bodies in the brain [5].

In this context, the objectively confirmed, correct diagnosis of RBD gains importance as patients with RBD may be the target group in the development of potentially neuroprotective drugs. Moreover, RBD needs to be differentiated from other parasomnias and movement disorders unrelated to a neurodegenerative process, such as periodic limb movement disorder (PLMD). According to the American Academy of Sleep Medicine (AASM), the current diagnosis of RBD is based on the following criteria: (1) repeated episodes of sleep-related vocalization and/or complex motor behaviors; (2) these behaviors are documented by polysomnography (PSG) to occur during REM sleep or, based on clinical history of dream enactment, are presumed to occur during REM sleep; (3) PSG recording demonstrates REM sleep without atonia (RSWA); and (4) the disturbance is not better explained by another sleep disorder, mental disorder, medication, or substance use [6].

One of the most critical points in the diagnosis of RBD is the documentation of RSWA. The current gold standard in RSWA identification is based on visual analysis of the electromyographic (EMG) signals recorded in the PSG. Three main visual methods are currently used for RSWA detection: (1) the AASM method that recommends identifying tonic and phasic muscular activity in the chin and phasic activity in the limb EMG signals in >27 per cent of 30 s REM sleep epochs [6, 7]; (2) the Montréal method that calls for RSWA identification when tonic chin activity is seen in >30 per cent of 20 s REM sleep epochs or phasic activity is detected in >15 per cent of 20 s REM sleep epochs [8, 9]; and (3) the SINBAR method that has validated cut-off values for tonic, phasic, and any muscle activities in various muscles and muscle combinations using 3 s miniepochs as well as 30 s epochs [10]. The major drawbacks of these visual methods are their time-consumption and proneness to subjective interpretation, due to unclear definitions of resting EMG values and muscular events.

To overcome the weakness of manual RSWA detection, several computerized methods have been developed, including the following: the REM Atonia Index (RAI) [11, 12], the Suprathermal Threshold REM Activity Metric (STREAM) [13], the short and long Muscle Activity Index (sMAI and lMAI) [14], the Frandsen Index (FRI) [15], the Kempfner Index (KEI) [16, 17], and the computerized version of the SINBAR method [18]. These methods are not completely automated, because they require visual identification of REM stages, manual removal of artifacts, and muscle activities related to apneas and arousals [19].

Previous studies have focused on comparison of an automated method with visual methods [19, 20], or on comparison of several automated methods in cohorts of other patient groups [21]. So far, comprehensive and thorough comparisons of several methods in a cohort of controls, PD+/-RBD, RBD, and PLMD participants are lacking. Moreover, only one study has conducted a quantitative analysis on the influence of apnea and arousal-related movements in the algorithm outcomes [18]. Therefore, this study has two main aims: First, to provide a comparative analysis of the ability of different methods in distinguishing groups of patients based on the amount of automatically detected RSWA. Second, to analyze the influence of apneas and arousal-related movements in the capability of the different algorithms to distinguish participant groups.

Methods
Participants and recordings
The study cohort included 27 healthy control participants (C), 25 patients with PD without RBD (PD-RBD), 29 patients with PD and RBD (PD + RBD), 29 idiopathic patients with PD (RBD), and 36 patients with PLMD which were all recruited among participants evaluated at the Danish Center for Sleep Medicine in the period 2009–2015. Patient evaluation included a comprehensive medical and medication history and a full-night PSG study. All participants and patients were advised to discontinue medications interfering with PSG (i.e. hypnotics and antidepressants) 2 weeks before the sleep recording. Sleep disorder diagnoses were made according to current guidelines [6]. Table 1 presents demographic and sleep overview information of the participants studied. The study was accepted by the Danish Health Authority and the Data Protection Agency. All data were anonymized. The work was carried out in accordance with the Declaration of Helsinki.

Sleep was scored and evaluated by expert technicians according to the AASM criteria [7]. The chin, and left and right tibialis EMG signals were analyzed between lights off and lights on at a sampling frequency of 256 Hz with different amplifier systems, of which the lowest antialiasing cutoff frequency was 70 Hz. To keep the analysis uniform, we applied a low-pass filter with 3 dB cutoff at 70 Hz to all EMG signals. In addition, they were further filtered with a high-pass filter (3 dB cutoff at 10 Hz) and a notch filter at 50 Hz (implemented as a band-stop filter with 3 dB cutoffs at 48 and 52 Hz). Before lights off, clinical staff checked whether electrode impedances were lower than AASM-recommended values [7]. For all the EMG signals, saturation artifacts caused by electrode detachments were found by applying an algorithm that searched areas exceeding the empirically defined threshold of 4000 μV. Such identified artifacts and the preceding and following 5 s were removed from the analysis.

RSWA detection algorithms
We implemented different automated methods for RSWA detection. Briefly, RAI measures the percentage of 1 s REM miniepochs with atonia in the chin signal [11, 12]; STREAM measures the percentage of 3 s REM miniepochs with RSWA in the chin signal [13]; sMAI and lMAI are measures of the numbers of short and long movements per hour of REM sleep in the chin signal, respectively [14]; FRI measures the percentage of 3 s miniepochs in REM sleep with RSWA in the chin signal [15]; and KEI is a measure of the percentage of 3 s miniepochs in REM sleep
with RSWA obtained by analyzing the chin and tibialis muscles [16, 17]. We could not implement the computerized version of the SINBAR method, due to lack of important implementation aspects in its description [18]. A more comprehensive description of the implemented methods can be found in Supplementary Table S1. All algorithms were implemented in Matlab (R2016b, The MathWorks, Natick, MA, USA).

**Evaluation method**

For the evaluation of the impact of apneas and arousal-related movements in the outcome of the algorithms, we applied the algorithms in six different configurations, also illustrated in Figure 1:

1. **Configuration 1**: without excluding any muscular activity related to apneas and arousals
2. **Configuration 2**: excluding muscle activity located from 3 s before to 12 s after an arousal onset (Figure 1a)
3. **Configuration 3**: excluding muscle activity from 5 s before an apnea onset to 5 s after the same apnea end (Figure 1b)
4. **Configuration 4**: excluding muscular activity occurring before (5 s) and after (5 s) an apnea event (Figure 1c)
5. **Configuration 5**: combination of configurations 2 and 3
6. **Configuration 6**: combination of configurations 2 and 4.

For each of the described configurations, we applied the five algorithms to all the participants in the cohort. However, due to lack of manually scored REM sleep for 5 PD-RBD, 2 PD + RBD, 1 RBD, and 1 PLMD patients, we could not apply the algorithms to them. The resulting indices (RAI, STREAM, FRI, sMAI, lMAI, and KEI) were used to compare the performances of the algorithms in distinguishing groups in the following comparisons:

- **Comparison 1**: (C, PD-RBD, PLMD) vs (PD+RBD, RBD)
- **Comparison 2**: C vs RBD
- **Comparison 3**: PLMD vs RBD
- **Comparison 4**: C vs PD-RBD
- **Comparison 5**: C vs PLMD
- **Comparison 6**: PD-RBD vs PD+RBD
- **Comparison 7**: C vs PLMD vs RBD.

For each comparison, configuration, and index, we trained and tested a logistic regression model with a fivefold cross-validation scheme. Twenty runs of such training and testing scheme were repeated with 20 different random partitions of the data, thus obtaining 100 values of training and test sensitivity, specificity, and accuracy for each comparison, configuration, and index. We decided to use the 20 runs of fivefold cross-validation scheme in order to guarantee generalization, given the relatively small number of participants. Briefly, for each fold, the participants considered in the comparison analyzed are divided in training and test set in ratio 4:1. During the training, the logistic regression model finds the best threshold in distinguishing between the classes with a maximum-likelihood optimization approach. This optimal model is then tested on the test set and the resulting sensitivity, specificity, and accuracy are calculated. This approach works also for multiclass problems as in comparison 7 [22]. In particular, we calculated the test overall accuracy for all comparisons and the sensitivity and specificities relative to the following classes: (PD + RBD, RBD) for comparison 1, RBD for comparison 2, RBD for comparison 3, PD-RBD for comparison 4, and PLMD.
PLMD for comparison 5, PD + RBD for comparison 6, and RBD for comparison 7.

We finally evaluated whether the different configurations were leading to significantly different values of test sensitivity, specificity, and accuracy. The following statistical approach was applied [23]. Considering for example one index (i.e. RAI) and one performance measure (i.e. the test sensitivity), the average performance measure across the 20 runs of fivefold cross-validation was calculated for each comparison and configuration, thus obtaining a $7 \times 6$ table containing the average values of such performance measure for each comparison (rows) and configuration (columns). A Friedman test was then applied to evaluate the effect of the different configurations across the comparisons and the correspondent $p$-value was calculated. The same approach was repeated for all indices and for the three performance measures (i.e. test sensitivity, specificity, and

Figure 1. Illustrations of the excluded segments (highlighted in red) of the polysomnographic recording. (a) From 3 s before to 12 s after an arousal onset; (b) from 5 s before an apnea onset to 5 s after an apnea end event; (c) from 5 s before an apnea onset and from an apnea end to 5 s later. The choice of the length of segments excluded for the analysis was based on empirical analysis of the signals.
accuracy), thus obtaining 18 final p-values. Due to multiple comparisons, the significance level of 0.05 was corrected to 0.0028 with Bonferroni correction.

Results
For each configuration, we calculated the values of all the indices for the participants with manually scored REM sleep in the cohort. An example is represented in Figure 2, where the distributions of the indices across participant groups are shown for configuration 1. The distributions obtained for the remaining configurations are shown in Supplementary Figures S1–S5. Supplementary Table S2 shows the mean and standard deviation of all the distribution of indices across participant groups for each of the six configurations. The analysis of these distributions revealed that control participants are characterized by the lowest degree of muscular activity, whereas higher level of muscular activations was seen for patients with RBD and PD + RBD. PLMD

Figure 2. Indices values calculated by taking into account all movements (configuration 1). (a) REM atonia index (RAI), (b) supra-threshold REM activity metric (STREAM), (c) Frandsen index (FRI), (d) short muscle activity index (sMAI), (e) long muscle activity index (lMAI), and (f) Kempfner index (KEI) across participant groups shown as box plots with crosses denoting group means, and whiskers indicating the 99th percentile range. Post hoc statistical comparison performed with Mann–Whitney U-test with Tukey–Kramer correction *p < 0.05; **p < 0.01; ***p < 0.001. C = healthy controls; PD+/–RBD = patients with Parkinson’s disease with and without RBD; RBD = patients suffering from idiopathic REM sleep behavior disorder; PLMD = patients suffering from periodic limb movement disorder.
and PD-RBD groups were instead usually characterized by a level of muscular activity between the low level for C and the high level for PD + RBD and RBD. However, some exceptions to these patterns were seen throughout the indices and configurations.

The muscular activity indices were used to train 20 runs of logistic regression models with a fivefold cross-validation scheme for each index, configuration, and comparison, and Supplementary Tables S3–S11 hold the mean and standard deviation values across the runs and folds for train and test sensitivity, specificity, and accuracy for all the models. The mean and 25th–75th percentile values of test sensitivity, specificity, and accuracy for all the models are shown in Figures 3 and 4 and the main outcomes for each comparison are presented below:

- **Comparison 1** (Figures 3a and 4a and Supplementary Table S3): RAI, FRI, and STREAM showed the best results in terms of sensitivity, specificity and accuracy (average values around 70%) in distinguishing participants diagnosed without RBD versus the ones with RBD. This comparison can be seen as the one testing the capability of the algorithms in identifying RBD in a general population cohort, where different diseases and diagnoses are present.

- **Comparison 2** (Figures 3b and 4b and Supplementary Table S4): when the methods were evaluated in distinguishing C and RBD, FRI outperformed the other methods with average values of sensitivity around 90 per cent, specificity and accuracy around 80 per cent. Also IMAI and KEI showed good performances, with average sensitivity, specificity, and accuracy around 80 per cent.

- **Comparison 3** (Figures 3c and 4c and Supplementary Table S5): RAI was the best index in differentiating RBD from PLMD, with average values of sensitivity, specificity, and accuracy around 70 per cent.

- **Comparison 4** (Figures 3d and 4d and Supplementary Table S6): in this comparison, the best index was considered the one with the lowest capability in distinguishing C from PD-RBD, because theoretically both groups do not suffer from RSWA. It was therefore expected that the indices were not able to correctly distinguish the two participant groups and STREAM presented the most coherent results with this expectation: around 0 per cent sensitivity, 50 per cent specificity, and 50 per cent accuracy.

- **Comparison 5** (Figures 2e and 3e and Supplementary Table S7): KEI showed higher sensitivity, specificity, and accuracy than the other methods (average values of sensitivity, specificity, and accuracy in the range 60–80%, excluding configuration 2) in distinguishing C and PLMD groups.

- **Comparison 6** (Figures 3f and 4f and Supplementary Table S8): STREAM, FRI, and sMAI showed the highest performances in distinguishing PD-RBD and PD + RBD based on the amount of detected RSWA (around 80% sensitivity, 60% specificity, and 60–70% accuracy for STREAM and FRI, and around 70% sensitivity, 100% specificity, and 70% accuracy for sMAI).

- **Comparison 7** (Figures 3g and 4g and Supplementary Table S9): in the discrimination of C, RBD, and PLMD groups, RAI showed the highest values in terms of sensitivity and specificity for detecting RBD (average specificity around 80% and sensitivity 60–70%). Considering the overall accuracy, RAI, and KEI presented comparable results (around 50%).

Table 2 shows the p-values obtained from each Friedman test used to evaluate the effect of the different configurations across the comparison for each index and test sensitivity, specificity, and accuracy. Due to the corrected significance criteria, none of the calculated p-values is significant, thus meaning that the test sensitivity, specificity, and accuracy are not affected by the different configurations. Only IMAI presents values close to the significance for sensitivity and accuracy, which might indicate that this index is the one that is mostly affected by different configurations.

**Discussion**

Quantification and detection of RSWA is fundamental in order to identify participants suffering from idiopathic RBD and therefore in high risk of developing synucleinopathies [24]. Current gold standard methods for RSWA detection are based on visual analysis of EMG signals and they do not define objectively resting EMG and muscular events [7–10]. For this reason, a robust and interclinically validated computerized method for RSWA quantification is desirable for clinical practice. In this study, we compared the performances of five automated methods in distinguishing different participant groups based on the amount of automatically detected RSWA and we also investigated the impact of including and excluding movements related to apneas and arousals. Two main outcomes came from our analysis: First, we observed that the performances of the methods vary depending on the comparison considered; thus, there is not an automated method that can be elected as the optimal in RSWA detection. Second, we observed that the performances of the different computerized methods in group classification are not influenced by movements related to apneas and arousals, thus supporting their robustness to these sleep events. In the subsections titled “Participants, Computation of RSWA indices, Comparison of the computerized methods, Analysis of apnea and arousal-related movements, and Limitations,” we will discuss more in detail the various aspects of our study.

**Participants**

The analysis of Table 1 reveals that the participants included did not differ significantly for age and gender distribution. Moreover, the apnea–hypopnea index was not significantly different among the groups, but significant difference was instead found when separating into the rate of apneas during REM and nonrapid eye movement (NREM). Similarly, the rate of arousals in REM and NREM was found to be significantly different between the groups. The periodic limb movement series (PLMS) index and the rate of PLM in REM and NREM sleep were calculated as the number of limb movements (LM) included in a PLM series per hour of sleep, REM sleep, and NREM sleep, respectively [25]. All these measures were found to be significantly different between the groups. On the other hand, the rate of LM not included in PLM series (nPLM) during the entire night, and NREM sleep were found not to significantly differ between the groups. Significant difference was instead found in the rate of nPLM during REM sleep. A deeper analysis of the PLMs index and the rates of PLM during REM and NREM shows that patients with PD-RBD show lower values than other groups, which seems to contradict previous findings [26]. However, a closer look to the recordings showed that patients with PD-RBD present increased limb muscular activity, mainly characterized...
by twitches and movements longer than 10 s, which were not classified as LM according to the standards [7]. These observations open the question whether a more comprehensive and data-driven definition of LM might be required in the future, as recently done for PLM [27]. In another study, we have attempted such an approach for patients with PLMD [28], but more investigations are needed for patients with PD, who might show even more complex movements.

Figure 3. Sensitivity (SE) and specificity (SP) values obtained during test for each index and configuration for comparison 1–7 (a–g). SE and SP values are calculated for the following classes: (a) PD+/−RBD, RBD; (b) RBD; (c) RBD; (d) PD−RBD; (e) PLMD; (f) PD+/−RBD, and (g) RBD. Values are shown as mean value with the whiskers indicating the 25th–75th percentiles through the 20 runs and 5 folds used in the classification scheme. C = healthy controls; PD+/−RBD = patients with Parkinson’s disease with and without RBD; RBD = participants suffering from idiopathic REM sleep behavior disorder; PLMD = participants suffering from periodic limb movement disorder; RAI = REM atonia index; STREAM = supra-threshold REM activity metric; FRI = Frandsen index; sMAI = short muscular activity index; lMAI = long muscular activity index; KEI = Kempfner index.
Computation of RSWA indices

In the implementation of the different algorithms here presented, we did not introduce any difference with respect to the original methods except the filtering in the preprocessing of the EMG signal. In fact, RAI was originally calculated in chin signals filtered between 10 and 100 Hz [11, 12] and sMAI/lMAI between 10 and 120 Hz [14]. Because of technical specification of the amplifiers available for recording, we had to filter the signals.

Figure 4. Accuracy (ACC) values obtained during test for each index and configuration for comparison 1–7 (a–g). Values are shown as mean value with the whiskers indicating the 25th–75th percentiles through the 20 runs and 5 folds used in the classification scheme. C = healthy controls; PD+/−RBD = patients with Parkinson’s disease with and without RBD; RBD = patients suffering from idiopathic REM sleep behavior disorder; PLMD = patients suffering from periodic limb movement disorder; RAI = REM atonia index; STREAM = supra-threshold REM activity metric; FRI = Frandsen index; sMAI = short muscular activity index; IMAI = long muscular activity index; KEI = Kempfner index.

Computation of RSWA indices
between 10 and 70 Hz. This filtering has already been applied to RAI [29], sMAI, and IMAI [21] before and we assume that it does not alter significantly the outcomes of the algorithms, but further studies should investigate this aspect.

When comparing the computed indices with the original ones obtained for the same groups of participants [11–17], it is noticed that their average values are generally in the same range, with only average sMAI values higher in our cohort compared with the ones presented in the original study [14] and in a recent validation study [30]. RAI values computed for our cohort present higher variances than the ones presented in the original studies for controls, RBD, PD-RBD, and PD + RBD patients [12, 20]. The variance of STREAM values computed for our controls and RBD participants is very similar to the original ones [13] and the same is observed for FRI values computed for controls, RBD, and PD patients [15]. SMAI and IMAI variances calculated for our controls and RBD patients are higher than the ones shown in the original study [14] and the later validation [30]. Kempfner et al. [17] used a low number of participants for testing their method; therefore, a comparison is not reasonable. In summary, it can be concluded that there are important analogies concerning mean values for the same groups of participants (excluding sMAI), but the variances for some indices are higher in our cohort than in previous studies. The higher sMAI values found in our study might be caused by the fact that our recordings have high ECG interference in the chin signal, maybe incorrectly identified as short movements by the implemented sMAI algorithm. ECG artifact correction with adaptive filtering might solve this problem, but at the same time, it might also include a bias due to changes in the frequency content of the EMG signal. On the other hand, the higher variances might be the consequence of interclinical and inter scorer variability, as well as artifacts that we have not taken into account such as snoring. The generally high comparability of our index values with the original indices is a good indication that the implementation of the algorithms was performed correctly.

### Comparison of the computerized methods

From the analysis illustrated in Figures 3 and 4 and Supplementary Tables S3–S9, it is not possible to identify the optimal automated algorithm for RSWA detection, because the algorithm performances change depending on the groups considered. When all comparisons are considered, FRI and KEI show promising results, but still need further validation before applying them in clinical settings. Moreover, it should be kept in mind that these two indices were developed in our sleep clinic and it has been shown that their results are highly correlated [15]. An external validation of these methods is needed in order to further confirm their robustness in a multicohort scenario.

In addition, it should be noted that RAI seems to be the most sensible index in RBD detection, because of its high performances in the comparisons C vs RBD, PLMD vs RBD, and C vs RBD vs PLMD. In particular, it achieved mean sensitivity and specificity of around 60%–80% in these comparisons. These observations confirm that RAI is a useful method in detecting RBD, both in idiopathic cases [11], and as a comorbidity to narcolepsy [31] and PD [20]. Furthermore, the results show good performances of RAI in distinguishing between patients with PLMD and RBD.

The analysis of some single comparisons reveals interesting aspects. For example, in comparison 4, it was expected that the indices showed poor performances in distinguishing the groups due to the theoretical absence of RSWA in both C and PD-RBD. At the same time, it was expected that in comparison 6, the indices were able to well distinguish PD-RBD from PD + RBD due to the presence of RSWA in the latter group. However, these expectations were not met in particular by RAI and IMAI, which show moderate capability of distinguishing PD-RBD from C and poor performances in distinguishing PD-RBD from PD + RBD. In previous studies, it has been shown that a large amount of patients with PD are characterized by REM behavioral events, which are defined as motor behaviors and/or vocalization with a purposeful component [32, 33]. Among the other methods, RAI and IMAI might capture such motor events, therefore leading to the unexpected performances before described. Moreover, in comparison 5, KEI has the highest performances when compared with other algorithms, which could simply be due to the fact that this method is the only one which also uses tibialis EMG signals in the analysis.

The average test sensitivity and specificity reached by the algorithms in all group comparisons were found to be generally below 80 per cent, suggesting that these methods can be used as supportive tools and not as stand-alone automated diagnostic tools. The varying results across indices, configurations, and comparisons suggest that another more robust and generalized algorithm is needed. Ideally, such an algorithm should perform fully automated RSWA detection, therefore including automatic REM sleep detection as well, which might help to make RSWA detection more objective. To the best of our knowledge, only one study has previously integrated REM and RSWA detection [17], and although it has shown promising results, it cannot be easily applied in a clinical environment, because of its computational complexity. In this context, it should be mentioned that an automated REM sleep detector should follow gold standard rules for REM sleep identification, but at present these criteria are dubious for RBD, as scoring of REM sleep requires that EMG is characterized by muscle atonia [7]. Moreover, changes in electroencephalographic and electrooculographic patterns have been observed in RBD when compared with controls [34, 35], making it even more difficult to identify REM sleep. Abnormal sleep patterns have also been seen in patients with neurodegenerative disease [36], and this reflects in low inter scorer variability [37, 38]. From this, it can be concluded that a new fully automated algorithm for RSWA detection might be developed independently of the gold standard rules, with the goal of effectively distinguishing different types of parasomnia.

### Table 2. p-Values obtained as outputs of the Friedman tests to evaluate the effect of different configurations across the seven comparisons on test sensitivity, specificity, and overall accuracy

<table>
<thead>
<tr>
<th>Index</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAI</td>
<td>0.7575</td>
<td>0.8032</td>
<td>0.3304</td>
</tr>
<tr>
<td>STREAM</td>
<td>0.2031</td>
<td>0.5882</td>
<td>0.0685</td>
</tr>
<tr>
<td>FRI</td>
<td>0.0681</td>
<td>0.0897</td>
<td>0.0897</td>
</tr>
<tr>
<td>sMAI</td>
<td>0.4844</td>
<td>0.9463</td>
<td>0.5297</td>
</tr>
<tr>
<td>IMAI</td>
<td>0.0150</td>
<td>0.0870</td>
<td>0.0155</td>
</tr>
<tr>
<td>KEI</td>
<td>0.1667</td>
<td>0.9189</td>
<td>0.3820</td>
</tr>
</tbody>
</table>

RAI = REM atonia index; STREAM = supra-threshold REM activity metric; FRI = Frandsen index; sMAI = short muscular activity index; IMAI = long muscular activity index; KEI = Kempfner index.
Analysis of apnea and arousal-related movements

The p-values shown in Table 2 lead us to the conclusion that muscular activities related to apneas and arousals do not influence the capability of the algorithms in distinguishing between the groups investigated, thus showing that the analyzed computerized methods seem to be robust to muscular artifacts related to apneas and arousals. This result is particularly relevant considering the significantly different number of arousals and apneas per hour of REM sleep in the groups included in the study (Table 1).

This result is in contrast with many studies using computerized methods, where it is highly recommended to remove respiration and arousal-related movements [18, 19]. It should be noticed that the participants here studied are not affected by severe apneas, which has previously been stated to be a confounder in RBD detection [39]. An accurate analysis of the outcomes of these methods in patients with severe apnea should be carried out in future to understand the effect of frequent apnea events.

Limitations

This study has several limitations. First, we did not include the computerized SINBAR method because the available description of the algorithm [18] is ambiguous and not detailed enough to be accurately reproduced. The SINBAR group claims that their computerized algorithm achieves higher performances after manual correction of arousal and respiration-related movements [18], an aspect contradicting our results. We think that a future comparative study including the computerized SINBAR will be highly interesting. Second, the fraction of men and age of the control group are not significantly different from the other groups, but more homogenous groups might be desirable to further reduce eventual confounders in the analysis. In this context, a larger sample size might also be desirable for future analyses and comparison of methods. Third, the chin EMG signal might be affected by snoring artifact, and we have not checked the influence of snoring in the signals, which might therefore affect the indices. Fourth, we chose empirically the length of segments to exclude in correspondence to arousals and sleep apneas. The choice was made in accordance with expert technicians and by visually inspecting some recordings. Different segment lengths might influence the analysis. Fifth, we have used logistic regression for distinguishing the different groups, and we cannot exclude that different classification methods can lead to different results. Finally, healthy controls, PD-RBD, PD + RBD, RBD, and PLMD patients have been selected on a 1:1 ratio, as done in the original studies where the computerized methods were proposed [11–17]. A future general population study where the prevalence of diseases [40–43] is taken into account will reveal a better estimate of the positive predictive value and specificity of the computerized methods. This study can therefore be seen as a prelude for future improvements and developments of automated RSWA detection algorithms that should be tested in the general population.

Conclusions

This study presents a comparative analysis of different computerized methods for detection of RSWA. In particular, the six indices calculated (RAI, STREAM, FRI, sMAI, lMAI, and KEI) were tested for their ability to distinguish between different patient groups. We observed high variability in performances across the comparisons analyzed; thus, it is not possible to elect one of the methods as the optimal automated RSWA index. All the indices generally present accuracy values that are not high enough for using them as stand-alone in clinical practice, but rather as supportive tools. Moreover, it was observed that the performances of the algorithms in distinguishing patient groups are not influenced by movements related to apneas and arousals. We think that this study can contribute in the further improvement and development of computerized methods for RSWA detection.

Supplementary Material

Supplementary material is available at SLEEP online.

Acknowledgments

We would like to thank Anders Stockmarr from the Department of Applied Mathematics and Computer Science of the Technical University of Denmark for his precious help in the statistical analysis.

Conflict of interest statement. Wolfgang H. Oertel is a Hertie-Senior-Research-Professor supported by the Charitable Hertie Foundation, Frankfurt/Main, Germany. All the other authors report neither financial nor nonfinancial disclosures.

References


