

Validating an automated sleep spindle detection algorithm using an individualized approach

LAURA B. RAY¹, STUART M. FOGEL², CARLYLE T. SMITH¹ and KEVIN R. PETERS¹

¹Trent University, Peterborough, ON, Canada and ²Queens University, Kingston, ON, Canada

Accepted in revised form 11 September 2009; received 20 July 2009

SUMMARY The goal of the current investigation was to develop a systematic method to validate the accuracy of an automated method of sleep spindle detection that takes into consideration individual differences in spindle amplitude. The benchmarking approach used here could be employed more generally to validate automated spindle scoring from other detection algorithms. In a sample of Stage 2 sleep from 10 healthy young subjects, spindles were identified both manually and automatically. The minimum amplitude threshold used by the PRANA[®] (PhiTools, Strasbourg, France) software spindle detection algorithm to identify a spindle was subject-specific and determined based upon each subject's mean peak spindle amplitude. Overall sensitivity and specificity values were 98.96 and 88.49%, respectively, when compared to manual scoring. Selecting individual amplitude thresholds for spindle detection based on systematic benchmarking data may validate automated spindle detection methods and improve reproducibility of experimental results. Given that interindividual differences are accounted for, we feel that automatic spindle detection provides an accurate and efficient alternative approach for detecting sleep spindles.

KEYWORDS automated detection, individual differences, sleep, sleep spindles

INTRODUCTION

Sleep spindles are waxing and waning oscillations that are visually apparent in the ongoing scalp-recorded electroencephalogram (EEG) of non-rapid eye movement (NREM) sleep and fall usually within the frequency range of 12–16 Hz. They occur predominantly in, and are one of the defining parameters of, Stage 2 sleep, although sleep spindles persist throughout slow wave sleep (Rechtschaffen and Kales, 1968). The electrophysiological characteristics of sleep spindles are well understood in both humans (for review see De Gennaro and Ferrara, 2003) and animals (Destexhe and Sejnowski, 2001; Steriade, 1999; Steriade and Amzica, 1998) and can vary depending upon age (Principe and Smith, 1982), gender (Carrier *et al.*, 2001; Driver *et al.*, 1996) and scalp location

(Jobert *et al.*, 1992). Within an individual, sleep spindle characteristics remain relatively constant from night to night (De Gennaro *et al.*, 2005; Gaillard and Blois, 1981; Silverstein and Levy, 1975), assuming that the amount of learning from day to day has not been manipulated (Fogel and Smith, 2006; Gais *et al.*, 2002). De Gennaro *et al.* (2005) suggest that interindividual characteristics of sleep spindles are so reliable that they could act as an 'electrophysiological fingerprint' for an individual. However, there are large differences in sleep spindle characteristics between individuals, particularly with regard to the variations in sleep spindle peak frequency range (Werth *et al.*, 1997) and amplitude (Bódizs *et al.*, 2009).

Traditionally, sleep spindles have been scored visually in accordance with standard sleep scoring procedures (Rechtschaffen and Kales, 1968). This approach is labour-intensive and time-consuming. More recently, algorithm-based automated sleep spindle detectors have been developed in attempts to replace manual sleep spindle detection methods reliably and efficiently. One limitation with regard to both automated and

Correspondence: Kevin Peters PhD, Psychology Department, Trent University, 1600 West Bank Drive, Peterborough, ON, K9J 7B8, Canada. Tel.: 705 748 1011 ext 7795; fax: 705 748 1580; e-mail: kevinpeters@trentu.ca

manual sleep spindle scoring is that there are varying definitions of the sleep spindle reported in the literature, making the criterion used for spindle scoring inconsistent across studies. Another limitation is that interscorer reliability for manual sleep spindle scoring is approximately 86% (Campbell *et al.*, 1980), suggesting that there is variability between scorers possibly due to subjectivity or expertise of the scorer. Consequently, the performance of each automated detection system is being compared to varying 'gold standards'. Interscorer reliability is reduced commonly by having more than a single independent and experienced scorer detect spindles manually in the same data, and the spindles are counted only when the scorers are in agreement (Campbell *et al.*, 1980; Morrow and Casey, 1986). Despite these challenges, several automated sleep spindle detection methods are being currently used (Acir and Güzeliş, 2004; Devuyst *et al.*, 2006; Held *et al.*, 2004; Huupponen *et al.*, 2007; Olbrich and Achermann, 2005; Schabus *et al.*, 2004; Schimicek *et al.*, 1994; Schonwald *et al.*, 2006; Ventouras *et al.*, 2005). Some of these methods rely upon a fixed amplitude threshold, regardless of the variability in sleep spindle amplitudes across subjects. There have been previous attempts to take into account the variability in spindle amplitudes when detecting spindles (Bódizs *et al.*, 2009; Huupponen *et al.*, 2000, 2007).

The goal of the current investigation was to validate systematically the accuracy of an automated sleep spindle detection method that takes into consideration individual differences in sleep spindle amplitude, and that uses a reliable manually scored gold standard for comparison.

METHODS

Subjects and data acquisition

The subjects were 10 (five males; five females) young volunteers (mean age = 20.14 ± 2.25 years; range = 17–24). At the time of the recordings, all subjects indicated that they were in good health. Exclusion criteria included individuals with irregular or unusual sleeping patterns, sleep disorders and medications that would disrupt sleep architecture. All subjects gave informed consent to participate, and data acquisition procedures were approved by the Trent University Research Ethics Board.

Polysomnograms for 10 individuals were selected randomly from an existing data set recorded previously with the Suzanne™ (Tyco Healthcare Group LP, Mansfield, MA, USA) portable polysomnographic system at 120 samples s^{-1} . The initial recordings included a bipolar submental electromyogram (EMG) channel and monopolar C3, C4, left- and right electro-oculogram (EOG) channels referenced to electrodes placed on the contralateral mastoid bones (A1 and A2). Data acquisition filters were set at a low-pass cut-off at 30 Hz and a high-pass cut-off of 0.3 Hz. The high-pass filter for the EMG channel was set at 10 Hz and no low-pass was used. Sleep stages were scored using 30 s epochs according to standard criteria (Rechtschaffen and Kales, 1968) using

SANDMAN™ software (version 7.2; Melville Diagnostics, Ottawa, ON, Canada).

Manual sleep spindle scoring

Working separately, two expert technologists, both with a minimum 10 years' experience of sleep scoring, identified and marked individual sleep spindles for the first 25 epochs of consolidated Stage 2 sleep from each quartile of the night. Sleep spindles were identified in 30 s epochs containing a C3-A2 channel with filters as described above and a duplicate channel with filters set to show only the 12–16 Hz frequency range. The criteria used in the manual scoring were that sleep spindles were defined as 12–16 Hz frequency range, at least 0.5 s in duration, and resembled the typical fusiform spindle morphology. Only those sleep spindles that were identified by both technologists were considered 'true' sleep spindles.

Minimal amplitude threshold determination and automated sleep spindle scoring

A single expert technologist identified and recorded the peak amplitudes of the first 15 sleep spindles in Stage 2 sleep from both the first and second halves of the night (see Fig. 1).

This procedure was used to determine the mean and standard deviation (SD) of the peak sleep spindle amplitudes for each subject. The minimal amplitude criterion was

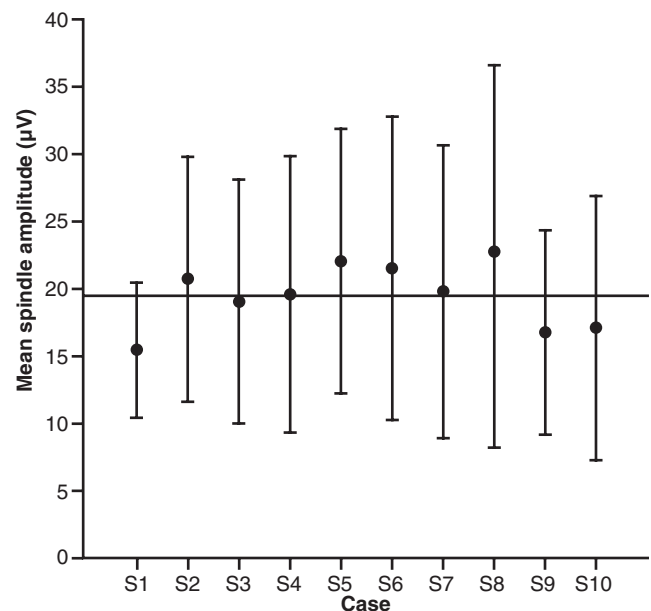


Figure 1. Mean sleep spindle peak amplitudes (filled circles) \pm 1.96 standard deviations (bars) for each of the 10 subjects. The range of mean sleep spindle amplitudes across all 10 subjects was 15.49–22.77 μV . To help illustrate the advantage of our approach, consider two other possible approaches. First, if one were to simply use the mean peak sleep spindle amplitude across subjects (19.49 μV), a number of sleep spindles would be missed, particularly in S1, S9 and S10. On the other hand, if one used an arbitrary criterion, such as 25 μV , all the sleep spindles would be missed in S1, and a considerable number would be missed in S2, S3, S4, S7, S9 and S10.

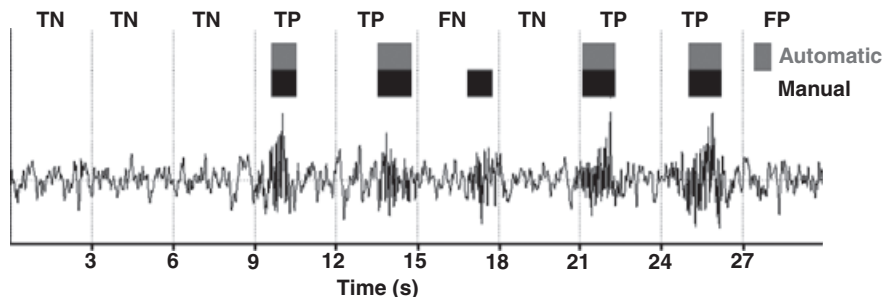


Figure 2. Procedure for sleep spindle identification. Electroencephalogram was divided into 3-s epochs. In each 3-s epoch, a decision was made as to the agreement between the manual and automated scoring methods. Spindles were identified and classified in the 3-s epoch containing spindle onset. When both manual and automatic methods identified the same sleep spindle correctly the epoch was categorized as true positive (TP). When both scoring methods did not score spindles within the 3-s epoch, the epoch was categorized as true negative (TN). Because we chose to use the manual scoring method as the ‘gold standard’, epochs in which a sleep spindle was identified by the automated method but not by the manual method were scored as false positives (FP). Similarly, epochs in which a sleep spindle was identified by the manual method but not the automated method were scored as false negatives (FN).

calculated by subtracting the mean peak amplitude by 1.96 SD for each subject. This number was chosen based on preliminary testing which identified that calculating the minimal amplitude criterion using a SD either less than or > 1.96 resulted in a reduction in sensitivity or specificity. To illustrate this approach, consider the following example: if the sleep spindle mean peak amplitude for a given individual was $20.75 \mu\text{V}$ and the SD was $4.65 \mu\text{V}$, the minimal amplitude criterion for this individual would be $20.75 - (4.65 \times 1.96) = 11.64 \mu\text{V}$. Thus, the amplitude of the sleep spindle had to be equal to or greater than $11.64 \mu\text{V}$ for the event to be counted as a sleep spindle by the automatic detector.

Prior to automated sleep spindle detection, high-frequency muscle artefact visible in the C3 scalp electrode was identified and excluded from further analyses. The PRANA[®] software (PhiTools) sleep spindle detection algorithm was employed to detect sleep spindles automatically using the minimal amplitude criterion described above. The PRANA[®] software sleep spindle detection algorithm is based upon the method published by Schimicek *et al.* (1994), with some minor enhancements for onset/offset detection and additional criteria for candidate selection. Stated simply, the algorithm initially band-pass filters the raw EEG to the user-defined frequency range. Next, the algorithm compares the user-defined amplitude criterion to the actual peak-to-peak amplitude. Lastly, the algorithm calculates the duration of the time epochs where the amplitudes exceed the user-defined criteria. If the user-defined duration criteria are met then sleep spindles are detected within the epoch. All other detection parameters were user-defined as: 12–16 Hz low-pass and high-pass digital filtering; minimum duration of 0.5 s; maximal duration of 3 s; and a minimum interspindle interval of 0.1 s.

Statistical procedure

For comparisons between manual and automated sleep spindle scoring, 3-s epochs were used. True positive (TP), true negative

Table 1 Mean (\pm SD) of true positive, false positive, true negative and false negative epochs

	Automated detections	
	Positive	Negative
Manual detections		
True	162.40 (± 86.48)	739.70 (± 123.28)
False	96.20 (± 51.20)	1.70 (± 1.77)

(TN), false positive (FP) and false negative (FN) epochs were identified visually (see Fig. 2). Test sensitivity [TP/(TP + FN)], specificity [TN/(TN + FP)] and FP rates [FP/(FP + TN)] were calculated (Huupponen *et al.*, 2007).

RESULTS

The mean peak amplitude across all 10 subjects ranged from 15.49 to 22.77 μV (see Fig. 1). The average peak amplitude across all 10 subjects was 19.49 μV , SD = 5.55 μV . A total of 1000 3-s epochs of Stage 2 sleep for each subject were used to compare the manual and automated sleep spindle scoring methods. The vast majority of epochs (73.97%) were categorized as TN, as would be expected for an event such as the sleep spindle that occurs typically three to five times per minute of Stage 2 sleep in a healthy young adult (Evans and Richardson, 1995). For the remaining epochs, 16.24% were scored as TP, 9.62% were scored as FP and only 0.17% were classified as FN. Table 1 presents descriptive data on the number of TP, FP, TN and FN for all subjects. Overall sensitivity of 98.96%, specificity of 88.49% and false positive rates of 11.51% were achieved when the automated scoring method was compared to the manual scoring method.

DISCUSSION

We present a systematic method to validate the accuracy of an automated sleep spindle detection algorithm using a subject-specific minimum amplitude threshold. Fixed amplitude crite-

ria do not take into account the variability in sleep spindle amplitudes across subjects. The importance of tuning the spindle detection amplitude threshold is also supported by previous work (Bódizs *et al.*, 2009; Huupponen *et al.*, 2000, 2007). We identified a sensitivity value of 98.96% and a specificity value of 88.49% when comparing manual scoring to the PRANA[®] software sleep spindle detection algorithm. This method could be used in the validation of any automated sleep spindle detection algorithm utilizing an amplitude criterion. In the first step, sleep spindles were identified independently by two expert scorers in the same series of epochs distributed evenly throughout the collection. Only those sleep spindles that were identified by both scorers were considered ‘true’ sleep spindles. In step 2, the minimum amplitude threshold was determined. For each subject, a sample of sleep spindles was identified from across the collection. The peak amplitudes were noted and were used to determine each subject’s mean peak amplitude. The minimal amplitude threshold was equal to each subject’s mean peak amplitude minus 1.96 SD. In the third step, the automated detection algorithm was employed using the subject-specific minimal amplitude threshold determined in step 2. In the final step, comparisons were made between the manual and automatic detection methods. Epochs of 3-s duration were used to identify the occurrence of TP, TN, FP and FN epochs.

The accuracy values reported here are comparable to, or greater than, those reported in previous studies utilizing other computer-based sleep spindle detection methods identifying sensitivities from 70 to 94.6% and specificities from 80.6 to 98.6% (Acir and Güzeliş, 2004; Huupponen *et al.*, 2007; Schonwald *et al.*, 2006; Ventouras *et al.*, 2005), although details of the analysis methods (e.g. epoch length, recording parameters, varying gold standards, user-defined spindle parameters) used to compare manual and automated detections vary between studies. More specifically, using a Matching Pursuit algorithm for detection of sleep spindles, Schonwald *et al.* (2006) reported a sensitivity of 80.6% and specificity of 80.6% for Stage 2 sleep. An average sensitivity of 94.6% and an average false detection rate of 4.0 were also obtained using an artificial neural network-based detection system (Acir and Güzeliş, 2004).

Although the current investigation analysed sleep spindles at a central scalp location during Stage 2 sleep, this procedure should be repeated for other scalp locations and for other NREM sleep stages (i.e. Stages 3 and 4). In addition, we used a fixed-frequency band (12–16 Hz) including both slow and fast spindles. This parameter was user-defined, and could be modified to analyse slow (e.g. 12–14 Hz) and fast spindles (e.g. 14–16 Hz) using an individualized approach in separate analyses for each type of spindle. Furthermore, future studies could apply this approach to the detection of sleep spindles in other age groups, such as adolescents (12–16 years of age) or seniors (60–85 years of age).

In conclusion, we have validated an automatic sleep spindle detection algorithm using a method that takes individual differences in spindle amplitude into account. While one of the

potential drawbacks of this procedure is that two expert scorers must first identify manually a sample of sleep spindles, the individualized approach outlined here could be implemented into existing computerized algorithms. A semi-automated approach would be the ideal compromise between a fully automated method that does not take into account individual differences and a completely manual scoring method, which is highly labour-intensive and time-consuming. We suggest that this approach may be useful for validating other methods and improve reproducibility of experimental results. Once the initial validation procedure is complete for a given algorithm, the procedure can be simplified to a single expert scorer identifying a sample of 30 sleep spindles from across the collection. The subject’s minimum amplitude threshold can be determined based upon the mean peak amplitude of the 30 spindles minus 1.96 SD. At this point the automated detection algorithm can be employed using the subject-specific minimal amplitude threshold.

ACKNOWLEDGEMENTS

This study was supported by Natural Sciences and Engineering Research Council (NSERC) of Canada and the Canadian Institutes of Health Research (CIHR).

REFERENCES

- Acir, N. and Güzeliş, C. Automatic recognition of sleep spindles in EEG by using artificial neural networks. *Exp. Syst. Appl.*, 2004, 27: 451–458.
- Bódizs, R., Körmendi, J., Rigó, P. and Lázár, A. The individual adjustment method of sleep spindle analysis: methodological improvements and roots in the fingerprint paradigm. *J. Neurosci. Methods*, 2009, 178: 205–213.
- Campbell, K., Kumar, A. and Hofman, W. Human and automatic validation of a phase-locked loop spindle detection system. *Electroencephalogr. Clin. Neurophysiol.*, 1980, 48: 602–605.
- Carrier, J., Land, S., Buysse, D. J., Kupfer, D. J. and Monk, T. H. The effects of age and gender on sleep EEG power spectral density in the middle years of life (ages 20–60 years old). *Psychophysiology*, 2001, 38: 232–242.
- De Gennaro, L. and Ferrara, M. Sleep spindles: an overview. *Sleep Med. Rev.*, 2003, 7: 423–440.
- De Gennaro, L., Ferrara, M., Vecchio, F., Curcio, G. and Bertini, M. An electroencephalographic fingerprint of human sleep. *Neuroimage*, 2005, 26: 114–122.
- Destexhe, A. and Sejnowski, T. J. *Thalamocortical Assemblies: How Ion Channels, Single Neurons and Large-Scale Networks Organize Sleep Oscillations*. Oxford University Press, New York, 2001.
- Devuyst, S., Dutoit, T., Didier, J. F. *et al.* Automatic sleep spindle detection in patients with sleep disorders. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, 2006, 1: 3883–3886.
- Driver, H. S., Dijk, D. J., Werth, E., Biedermann, K. and Borbély, A. A. Sleep and the sleep electroencephalogram across the menstrual cycle in young healthy women. *J. Clin. Endocrinol. Metab.*, 1996, 81: 728–735.
- Evans, M. and Richardson, N. E. Demonstration of a 3–5-s periodicity between the spindle bursts in NREM sleep. *J. Sleep Res.*, 1995, 4: 196–197.
- Fogel, S. M. and Smith, C. T. Learning-dependent changes in sleep spindles and Stage 2 sleep. *J. Sleep Res.*, 2006, 15: 250–255.

- Gaillard, J. M. and Blois, R. Spindle density in sleep of normal subjects. *Sleep*, 1981, 4: 385–391.
- Gais, S., Molle, M., Helms, K. and Born, J. Learning-dependent increases in spindle density. *J. Neurosci.*, 2002, 22: 6830–6834.
- Held, C. M., Causa, L., Estévez, P. *et al.* Dual approach for automated sleep spindles detection within EEG background activity in infant polysomnograms. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, 2004, 1: 566–569.
- Huupponen, E., Värri, A., Himanen, S. L., Hasan, J., Lehtokangas, M. and Saarinen, J. Optimization of sigma amplitude threshold in sleep spindle detection. *J. Sleep Res.*, 2000, 9: 327–334.
- Huupponen, E., Gomez-Herrero, G., Saastamoinen, A., Varri, A., Hasan, J. and Himanen, S. Development and comparison of four sleep spindle detection methods. *Artif. Intell. Med.*, 2007, 40: 157–170.
- Jobert, M., Poiseau, E., Jähnig, P., Schultz, H. and Kubicki, S. Topographical analysis of sleep spindle activity. *Neuropsychobiology*, 1992, 26: 210–207.
- Morrow, T. J. and Casey, K. L. A microprocessor device for the real-time detection of synchronized alpha and spindle activity in the EEG. *Brain Res. Bull.*, 1986, 16: 439–442.
- Olbrich, E. and Achermann, P. J. Analysis of oscillatory patterns in the human sleep EEG using a novel detection algorithm. *Sleep Res.*, 2005, 14: 33–346.
- Principe, J. C. and Smith, J. Sleep spindle characteristics as a function of age. *Sleep*, 1982, 5: 73–84.
- Rechtschaffen, A. and Kales, A. *A Manual of Standardized Terminology, Techniques and Scoring Systems for Sleep Stages in Human Subjects*. Government Printing Office, Washington, DC, 1968.
- Schabus, M., Gruber, G., Parapatics, S. *et al.* Sleep spindles and their significance for declarative memory consolidation. *Sleep*, 2004, 27: 1479–1485.
- Schimicek, P., Zeitlhofer, J., Anderer, P. and Saletu, B. Automatic sleep-spindle detection procedure: aspects of reliability and validity. *Clin. Electroencephalogr.*, 1994, 25: 26–29.
- Schonwald, S., de Santa-Helena, E. L., Rossatto, R., Chaves, M. and Gerhardt, G. Benchmarking matching pursuit to find sleep spindles. *J. Neurosci. Methods*, 2006, 156: 314–321.
- Silverstein, L. D. and Levy, C. M. The stability of the sigma sleep spindle. *Electroencephalogr. Clin. Neurophysiol.*, 1975, 40: 666–670.
- Steriade, M. Coherent oscillations and short-term plasticity in corticothalamic networks. *Trends Neurosci.*, 1999, 22: 337–345.
- Steriade, M. and Amzica, F. Coalescence of sleep rhythms and their chronology in corticothalamic networks. *Sleep Res. Online*, 1998, 1: 1–10.
- Ventouras, E., Monoyiou, E., Ktonas, P. *et al.* Sleep spindle detection using artificial neural networks trained with filtered time-domain EEG: a feasibility study. *Comput. Methods Prog. Biomed.*, 2005, 78: 191–207.
- Werth, E., Achermann, P., Dijk, D. J. and Borbély, A. Spindle frequency activity in the sleep EEG: individual differences and topographic distribution. *Clin. Neurophysiol.*, 1997, 103: 535–542.