Clinical Neurophysiology 126 (2015) 1548–1556

Contents lists available at ScienceDirect

Clinical Neurophysiology

journal homepage: www.elsevier.com/locate/clinph

Inter-expert and intra-expert reliability in sleep spindle scoring

Sabrina L. Wendt [a,b], Peter Welinder [c], Helge B.D. Sorensen [d], Paul E. Peppard [e], Poul Jennum [b], Pietro Perona [c], Emmanuel Mignot [a], Simon C. Warby [a,f,*]

[a] Center for Sleep Science and Medicine, Stanford University, Palo Alto, CA, United States
[b] Danish Center for Sleep Medicine, Glostrup University Hospital, DK-2600 Glostrup, Denmark
[c] Computational Vision Laboratory, California Institute of Technology, Pasadena, CA, United States
[d] Dept. of Electrical Engineering, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark
[e] Department of Population Health Sciences, University of Wisconsin – Madison, Madison, WI, United States
[f] Center for Advanced Research in Sleep Medicine, Hôpital du Sacré-Coeur de Montréal, Department of Psychiatry, Université de Montréal, Montréal, Canada

ARTICLE INFO

Article history:
Accepted 29 October 2014
Available online 10 November 2014

Keywords:
Sleep spindles
Agreement
Reliability
Inter-rater
Inter-expert
Intra-rater
Intra-expert
Electroencephalography
Polysomnography
Event detection
Sleep staging
Sleep scoring

HIGHLIGHTS

- Spindle identification is a difficult task, and more than one sleep expert is needed to reliably score spindles in EEG data.
- The reliability of sleep staging may be improved by improving the reliability of spindle scoring, particularly for the discrimination of stage N1 and N2 sleep.
- Reliability of sleep spindle scoring can be improved by using qualitative confidence scores, rather than a dichotomous yes/no scoring system.

ABSTRACT

Objectives: To measure the inter-expert and intra-expert agreement in sleep spindle scoring, and to quantify how many experts are needed to build a reliable dataset of sleep spindle scorings.
Methods: The EEG dataset was comprised of 400 randomly selected 115 s segments of stage 2 sleep from 110 sleeping subjects in the general population (57 ± 8, range: 42–72 years). To assess expert agreement, a total of 24 Registered Polysomnographic Technologists (RPSGTs) scored spindles in a subset of the EEG dataset at a single electrode location (C3-M2). Intra-expert and inter-expert agreements were calculated as F_1-scores, Cohen's kappa (κ), and intra-class correlation coefficient (ICC).
Results: We found an average intra-expert F_1-score agreement of 72 ± 7% (κ: 0.66 ± 0.07). The average inter-expert agreement was 61 ± 6% (κ: 0.52 ± 0.07). Amplitude and frequency of discrete spindles were calculated with higher reliability than the estimation of spindle duration. Reliability of sleep spindle scoring can be improved by using qualitative confidence scores, rather than a dichotomous yes/no scoring system.
Conclusions: We estimate that 2–3 experts are needed to build a spindle scoring dataset with 'substantial' reliability (κ: 0.61–0.8), and 4 or more experts are needed to build a dataset with 'almost perfect' reliability (κ: 0.81–1).
Significance: Spindle scoring is a critical part of sleep staging, and spindles are believed to play an important role in development, aging, and diseases of the nervous system.

1. Introduction

Sleep spindles are discrete events observed in the scalp electroencephalogram (EEG) signal that are generated as a result of interactions between several regions of the brain including thalamic and cortico-thalamic networks (De Gennaro and Ferrara, 2003). They are observed as brief 11–16 Hz bursts that are distinct

* Corresponding author at: Université de Montréal, Center for Advanced Research in Sleep Medicine (CARSM), Sacré-Coeur Hospital of Montréal, 5400 Gouin Blvd. West, J-5000, Montréal, QC H4J 1C5, Canada.
E-mail address: simon.c.warby@umontreal.ca (S.C. Warby).

http://dx.doi.org/10.1016/j.clinph.2014.10.158
1388-2457/

from the background activity, typically last less than a second, are maximal at central scalp locations, and have a characteristic waxing and waning amplitude (Iber et al., 2007). Spindles are a defining EEG feature of non-REM stage 2 (N2) sleep, although they can also occur in N3 (Iber et al., 2007). The gold standard to detect spindles is visual scoring by a trained sleep technologist. However, the EEG is a noisy signal, making the process of identifying individual spindle events very time consuming and subjective. Spindle density (counts/min), amplitude and duration decrease with age (Crowley et al., 2002; Martin et al., 2013), which might make spindle identification a more difficult task in older subjects. The purpose of this study was to estimate intra-expert and inter-expert reliabilities of spindle scoring using EEG data from middle-to-older aged subjects in the general population.

Identification of sleep spindles is of great clinical and biological interest because they are believed to play an important role in development, aging, and diseases of the nervous system. Spindle density (Bodizs et al., 2005; Fogel et al., 2007), frequency (Geiger et al., 2011; Gruber et al., 2013), and activity (Schabus et al., 2006, 2008) have been correlated with both intelligence and general mental ability. Moreover, increased sleep spindle density following learning predicts improved memory consolidation (Bergmann et al., 2012; Eschenko et al., 2006; Gais et al., 2002; Genzel et al., 2012; Schabus et al., 2006, 2008; Tamminen et al., 2010; Wamsley et al., 2012). Pharmacological interventions that increase spindle density have been found to correlate with improvements in specific types of memory (Kaestner et al., 2013; Mednick et al., 2013) and spindle density has been associated with selective attention (Forest et al., 2007). Numerous studies have found alterations in sleep spindle density in patients with psychiatric (Ferrarelli et al., 2007, 2010; Limoges et al., 2005; Miano et al., 2004; Seeck-Hirschner et al., 2010; Wamsley et al., 2012) and neurologic disease (Comella et al., 1993; Emser et al., 1988; Montplaisir et al., 1995; Myslobodsky et al., 1982; Silvestri et al., 1995; Wiegand et al., 1991).

One common limitation in research studies is that they focused only on spindle density, and ignored spindle characteristics like oscillation frequency, amplitude and duration, possibly because this information is more difficult to obtain. However, elegant modeling on how various neuronal networks are involved in the initiation, amplification, maintenance, or termination of sleep spindle bursts suggest that spindle characteristics may reflect an important role in the function of the spindle (Bazhenov et al., 2002; Bonjean et al., 2012, 2011; Fuentealba et al., 2005; Olbrich and Achermann, 2008). For example, specific types of memory consolidation have been associated with specific topographical locations (Martin et al., 2013) and oscillation frequencies (Fogel et al., 2012; Molle et al., 2011). The amplitude and duration of spindles also appears to be important for age-related changes (Nicolas et al., 2001), and are altered by benzodiazepines (Kaestner et al., 2013). The analysis of spindle characteristics requires precise determination of the beginning and end of spindle events in the EEG time series. Therefore, we previously tested several automatic sleep spindle detection algorithms and found their performance for detecting discrete spindle events to be significantly different from human experts. Further, the average inter-algorithm agreement was low (F_1-score = 32 ± 16%), suggesting that spindle detection was not consistent between automated detectors (Warby et al., 2014).

Identifying sleep spindles is also important because it is a critical part of sleep stage scoring. Spindles and K-complexes are the two EEG features that are used to differentiate stage 2 from stage 1. Despite detailed rules and guidelines however, inter-expert agreement for sleep stage scoring is not perfect. Studies from the last decade report an overall stage scoring agreement between observers of 76–82% (κ: 0.63–0.76) both in healthy subjects and patients with various sleep pathologies (Anderer et al., 2005; Danker-Hopfe et al., 2009, 2004; Magalang et al., 2013; Malinowska et al., 2009; Pittman et al., 2004). The agreement in scoring stage 2 is in the same range (κ: 0.60–0.72), whereas scoring stage 1 has considerably lower agreement (κ: 0.31–0.46) (Danker-Hopfe et al., 2009, 2004; Magalang et al., 2013). Furthermore, agreement in stage scoring has shown to worsen in subjects with increasing age and sleep disorder severity (Anderer et al., 2005). Improving the agreement of sleep spindle scoring, particularly in the transition of stage 1 to stage 2 in the EEG of older subjects may be important for improving the overall reliability of sleep stage scoring.

Very few studies have looked specifically at the agreement between human spindle scorers. In these studies, there were between 6 and 12 subjects (21–59 years old), and at most three experts were used to score spindles. In general, results were consistent with sleep stage scoring agreement, except that spindle scoring reliability in most cases deteriorated more rapidly with age and sleep pathologies. In healthy subjects, Huupponen et al. and Campbell et al. estimated 81% and 86% inter-expert agreement in sleep spindle identification, respectively (Campbell et al., 1980; Huupponen et al., 2007). Using three annotators, Zygierewicz et al. estimated an average agreement of 70 ± 8% in spindle identification in healthy subjects (Zygierewicz et al., 1999). In contrast, Devuyst et al. did not find the agreement in spindle scoring between two experts measured by F_1-score to be more than 46% when using slightly older patients with various sleep pathologies (Devuyst et al., 2011). To the best of our knowledge, no studies evaluating the intra-expert reliability in sleep spindle scoring have been reported.

The purpose of this study was to assess the intra-expert and inter-expert agreement of sleep spindle scoring averaged over multiple pairs of experts to find the mean pairwise reliability. In addition to measuring the reliability of identifying spindle events in the EEG signal, we also assess the reliability of estimating spindle characteristics of the events, such as duration. Finally, based on our calculation of mean inter-expert reliability, we estimate how many experts are needed to build a reliable dataset of spindle scorings in EEG of older subjects.

2. Methods

2.1. Subjects and the EEG dataset

The EEG data used in the study was 110 middle aged and older subjects (mean±SD: 57 ± 8 years, range: 42–72 years, 47% male). These subjects were selected as a random subset of the Wisconsin Sleep Cohort (Peppard et al., 2013), which is a representative sample of the general population. In-clinic overnight polysomnography (PSG) was collected on these subjects following standardized protocols (Peppard et al., 2009), including 18-channels in a referential montage to record sleep stage, breathing, heart rate and rhythm, leg movements, snoring, arterial oxygenation, and body position. EEG data were collected with a sampling frequency of 100 Hz and band-pass filtered between 0.3 and 35 Hz. Sleep staging was conducted using standard criteria according to Rechtschaffen and Kales (1968). In total, the dataset consisted of 400 randomly selected, artifact-free, 115 s segments of stage 2 sleep. Each 115 s segment was broken into 5 epochs of 25 s each, overlapping by 2.5 s. The segments were extracted from the 110 subjects in the following manner: 2 segments (10 epochs) were randomly selected from 100 subjects and 20 segments (100 epochs) from 10 subjects. We chose to sample a lot of data from few subjects and little data from many subjects to estimate both intra-subject and inter-subject spindle variations, thereby getting most information from a