Detailed Assessment of Sleep Architecture With Deep Learning and Shorter Epoch-to-Epoch Duration Reveals Sleep Fragmentation of Patients With Obstructive Sleep Apnea

Henri Korkalainen, Timo Leppänen, Brett Duce, Samu Kainulainen, Juhani Aakko, Akseli Leino, Laura Kalevo, Isaac O. Afara, Sami Myllymaa, and Juha Töyräs

***Abstract*—Traditional sleep staging with non-overlapping 30-second epochs overlooks multiple sleep-wake transitions. We aimed to overcome this by analyzing the sleep architecture in more detail with deep learning methods and hypothesized that the traditional sleep staging underestimates the sleep fragmentation of obstructive sleep apnea (OSA) patients. To test this hypothesis, we applied deep learning-based sleep staging to identify sleep stages with the traditional approach and by using overlapping 30-second epochs with 15-, 5-, 1-, or 0.5-second epoch-to-epoch duration. A dataset of 446 patients referred for polysomnography due to OSA suspicion was used to assess differences in the sleep architecture between OSA severity groups. The amount of wakefulness increased while REM and N3 decreased in severe OSA with shorter epoch-to-epoch duration. In other OSA severity groups, the amount of wake and N1 decreased while N3 increased. With the traditional 30-second epoch-to-epoch duration, only small differences in sleep continuity were observed between the OSA severity groups. With 1-second epoch-to-epoch duration, the hazard ratio illustrating the risk of fragmented sleep was 1.14 ($p = 0.39$) for mild OSA, 1.59 ($p < 0.01$) for moderate OSA, and 4.13 ($p < 0.01$) for severe OSA. With shorter epoch-to-epoch durations, total sleep time and sleep efficiency increased in the non-OSA group and decreased in severe OSA. In conclusion, more detailed sleep analysis emphasizes the highly fragmented sleep architecture in severe OSA patients which can be underestimated with traditional sleep staging. The results highlight the need for a more detailed analysis of sleep architecture when assessing sleep disorders.**

***Index Terms*—Deep learning, electroencephalography, obstructive sleep apnea, sleep fragmentation, sleep staging.**

Manuscript received May 24, 2020; revised November 6, 2020; accepted December 6, 2020. Date of publication December 9, 2020; date of current version July 20, 2021. This work was supported in part by the Research Committee of the Kuopio University Hospital Catchment Area for the State Research Funding (Projects 5041767, 5041768, 5041770, 5041776, 5041779, 5041780, 5041782, 5041794, and 5041797), in part by the Respiratory Foundation of Kuopio Region, the Research Foundation of the Pulmonary Diseases, Foundation of the Finnish Anti-Tuberculosis Association, the Päivikki and Sakari Sohlberg Foundation, the Finnish Cultural Foundation via the Post Docs in Companies program, the North-Savo Regional Fund, and the Central Fund, the Paulo Foundation, the Tampere Tuberculosis Foundation, the Erkko Foundation, and by Business Finland (decision 5133/31/2018) *(Corresponding author: Henri Korkalainen.)*

Henri Korkalainen, Timo Leppänen, Samu Kainulainen, Akseli Leino, Laura Kalevo, and Sami Myllymaa are with the Department of Applied Physics, University of Eastern Finland, 70210 Kuopio, Finland, and also with the Diagnostic Imaging Center, Kuopio University Hospital, 70210 Kuopio, Finland (e-mail: henri.korkalainen@uef.fi; timo.leppanen@uef.fi; samu.kainulainen@uef.fi; akseli.leino@uef.fi; laura.kalevo@uef.fi; sami.myllymaa@uef.fi).

Brett Duce is with the Department of Respiratory & Sleep Medicine, Sleep Disorders Centre, Woolloongabba, QLD 4102, Australia, and with the Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane City, QLD 4000, Australia (e-mail: brett.duce@health.qld.gov.au).

Juhani Aakko is with the CGI Suomi Oy, 00380 Helsinki, Finland (e-mail: juhani.aakko@cgi.com).

Isaac O. Afara is with the Department of Applied Physics, University of Eastern Finland, 70210 Kuopio, Finland, and also with the School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD 4072, Australia (e-mail: isaac.afara@uef.fi).

Juha Töyräs is with the Department of Applied Physics, University of Eastern Finland, Kuopio, Finland, and with the Diagnostic Imaging Center, Kuopio University Hospital, Kuopio, Finland, and also with the School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane 4072, Australia (e-mail: juha.toyras@uef.fi).

Digital Object Identifier 10.1109/JBHI.2020.3043507

I. Introduction

SLEEP is a restorative state with a multitude of functions such as memory consolidation and clearance of metabolic waste products from the brain [1], [2]. Sleep can be objectively assessed using electroencephalography (EEG), electrooculography (EOG), and electromyography (EMG) recorded during overnight polysomnography (PSG) and subjectively categorized into stages according to defined criteria [3]. The current practice of sleep staging utilizes a segmentation method of assigning a nominal sleep stage to each non-overlapping 30-second epoch from the onset of the recording [3]. This 30-second division is an arbitrary system that is a historical remnant of when EEG recordings were printed on paper and is not wholly based on physiological factors [4]–[8]. The 30-second epoch-based scoring was optimized for convenience and less labor rather than producing a more accurate representation of sleep [5]–[7].

A defining characteristic of the 30-second epoch staging system is that multiple sleep stages may be present in a single epoch but only a single stage can be assigned for each epoch. Therefore,

many transitions between sleep stages and between sleep and wakefulness remain overlooked and, for example, wake periods with a duration up to 30-seconds may be completely overlooked when divided over two consecutive epochs. This can cause overestimation of sleep quality and underestimation of sleep fragmentation and can also significantly affect the determination of sleep onset or the onset of REM sleep. Furthermore, the 30-second epoch-based sleep staging can cause large uncertainties in tests objectively assessing daytime sleepiness, for example Multiple Sleep Latency Test and Maintenance of Wakefulness Test, where the accurate identification of sleep onset would be paramount. Additionally, missing transitions from sleep to wakefulness can affect the estimation of the duration of continuous sleep periods and also affect the values of various clinical parameters used to illustrate the sleep architecture, for example, the total sleep time (TST), sleep efficiency (SE), and duration of wake after sleep onset (WASO). The current sleep staging practice with non-overlapping 30-second epochs is problematic especially when the sleep architecture is disturbed due to sleep disorders [5].

Obstructive sleep apnea (OSA) is one of the most common sleep disorders affecting over 900 million individuals [9]. OSA is characterized by recurrent obstructions of upper airways which often lead to arousals from sleep causing sleep fragmentation and multiple transitions between sleep stages and wakefulness [10], [11]. However, due to the current convention of sleep staging based on non-overlapping 30-second epochs, many of these transitions can be easily missed. Therefore, we hypothesize that the current sleep staging with non-overlapping 30-second epochs heavily underestimates the extent of sleep fragmentation in patients suffering from OSA. As deep learning-based methods have demonstrated remarkable accuracy in automatic sleep staging [12]–[17], we hypothesize that deep learning offers a unique possibility for providing a more feasible and accurate representation of sleep architecture beyond the non-overlapping 30-second epochs.

We have recently introduced a deep learning-based automatic sleep staging method [16] that surpassed previous state-of-the-art methods on a publicly available research dataset (Physionet Sleep-EDF [18], [19]). In a clinical dataset of patients with suspected OSA, the developed method reached at least similar inter-rater reliability (83.8% accuracy, $\kappa = 0.78$) [16] as between two manual scorers [20]–[22]. The deep learning-based sleep staging also succeeded in accurately identifying sleep stages from a single EEG channel [16] or even from a photoplethysmography signal [17]. However, the main advantage of the automatic sleep staging over manual scoring is the ability to always score the sleep stages consistently. Therefore, we aim to utilize this previously developed automatic method to assess sleep architecture in a more detailed manner. Furthermore, we aim to study how the sleep architecture of patients with varying degrees of OSA differs with more detailed sleep staging

II. Methods

A. Dataset

We have previously presented a deep learning-based automatic sleep staging model utilizing a clinical dataset of Type 1 polysomnographies (PSGs) [16]. The PSGs were conducted at the Princess Alexandra Hospital (Brisbane, Australia) for the clinical suspicion of OSA and recorded with the Compumedics Grael acquisition system (Compumedics, Abbotsford, Australia) between 2015 and 2017. The dataset comprised of 891 recordings out of which 717 were used to train the final model. In the present study, we retrained the model using half of the same population ($n = 445$). The remaining 446 recordings were left outside the retraining process and were included in the analyses conducted in the present study (Table I). The data collection was approved by the Institutional Human Research Ethics Committee of the Princess Alexandra Hospital (HREC/16/QPAH/021 and LNR/2019/QMS/54313).

TABLE I
Demographic Information of the Studied Population of Suspected Obstructive Sleep Apnea (OSA) Patients ($n = 446$)

	n **(% of the population)**
Non-OSA (AHI < 5)	71 (32%)
Mild OSA (5 ≤ AHI < 15)	125 (24%)
Moderate OSA (15 ≤ AHI < 30)	108 (28%)
Severe OSA (AHI ≥ 30)	142 (16%)
Female	199 (45%)
Male	247 (55%)
	Median (interquartile range)
Age (years)	56.1 (44.7–65.5)
Arousal index (1/h)	21.2 (14.3–33.4)
Apnea-hypopnea index, AHI (1/h)	18.4 (7.5–36.6)
Body mass index (kg/m^2)	35.0 (29.8–41.2)
Sleep efficiency (%)	67.0 (55.5–78.5)
Sleep latency (min)	19.0 (10.0-37.0)
Total recording time (min)	441.5 (405.0–476.5)
Total sleep time (min)	308.0 (252.3–354.0)
Wake after sleep onset (min)	136.5 (91.5–183.0)

B. Sleep Staging

The deep learning model comprised of a combined convolutional (CNN) and recurrent neural network (RNN) conducting the sleep staging in a sequence-to-sequence manner from sequences of hundred 30-second epochs (Python 3.6 with Keras API 2.24 and TensorFlow 1.13 backend). The CNN architecture consisted of six 1D convolutions each followed by batch normalization and a ReLU activation. A max-pooling layer was located after the first two and the two following convolutions. The final layer comprised a global average pooling layer. The complete network architecture consisted of a time distributed layer of the CNN described above followed by a gaussian dropout layer, a bidirectional long short-term memory (LSTM) layer, and a time distributed dense layer with softmax activation. An EEG channel (derivation F4-M1) and an EOG channel (derivation E1-M2) were used for the automatic sleep staging. No preprocessing was conducted on the signals aside from downsampling to 64 Hz from the original sampling frequency of 1024 Hz. The architecture of the model and the workflow for training the model is presented with more details in Korkalainen *et al.* [16].

In the present study, the model architecture presented previously in [16] was trained using only half ($n = 445$) of the complete population of 891 recordings. This was further split

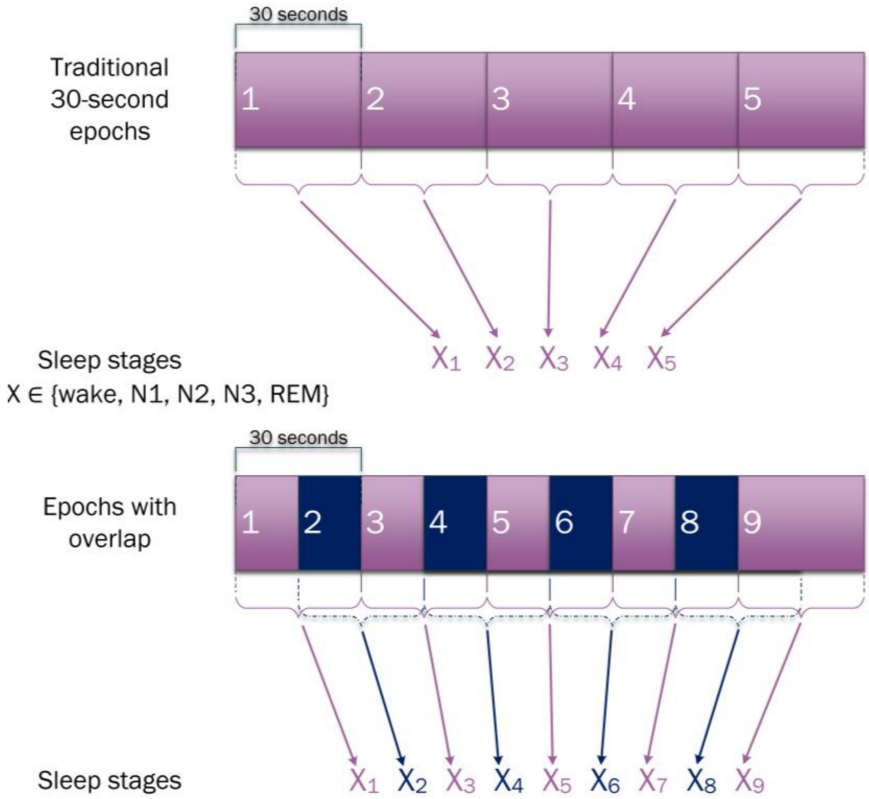

Fig. 1. Illustration of the sleep staging procedure with traditional sleep staging and when using overlap between the 30-second epochs. When using overlap, a sleep stage is identified for each epoch and the identified sleep stages are ordered according to the starting point of the epoch. In the figure, X_i illustrates an identified sleep stage (wake, N1, N2, N3, or REM) in an epoch i. Only the overlap with an epoch-to-epoch duration of 15 seconds is illustrated for clarity. The shorter epoch-to-epoch durations are treated similarly.

into a training set ($n = 400$) and a validation set ($n = 45$). The validation set was used in selecting the best performing model during training, i.e., the model with the lowest validation loss was selected. The remaining 446 recordings were not used in the training of the model and were included only in the further analyses. The retraining of the previously presented model was conducted to allow for a larger dataset to be used in the present study without having to rely on recordings that have been used during the training of the model.

After retraining the model, the study population not included in the training was reanalyzed with the deep learning-based sleep staging method. In addition to the traditional sleep staging with non-overlapping 30-second epochs, the sleep staging was conducted by allowing consecutive 30-second epochs to overlap with four different epoch-to-epoch durations: a new 30-second epoch taken every 15 seconds (50% overlap), every 5 seconds (83.3% overlap), every 1 second (96.7% overlap), or every 0.5 seconds (98.3% overlap). Each scoring then formed a time series of sleep stages (Fig. 1). The sleep architecture of the different OSA severity groups (non-OSA, mild OSA, moderate OSA, and severe OSA) were compared using three different approaches: 1) calculating the sleep stage percentages in each severity group, 2) calculating commonly used sleep parameters (total sleep time, sleep efficiency, and wake after sleep onset) for each individual in the groups, and 3) evaluating the continuity of sleep in each group based on survival analysis.

Three commonly used sleep parameters, total sleep time (TST), sleep efficiency (SE), and amount of wake after sleep onset (WASO), were calculated for each patient and mean and standard deviations were calculated for each OSA severity group. The statistical significance was evaluated using the Mann-Whitney *U* test when comparing the OSA groups to the non-OSA group and with Wilcoxon signed-rank test when comparing the sleep parameters between the more detailed sleep staging to the traditional sleep staging within the same OSA severity group.

Sleep continuity was evaluated based on survival analysis methodology. The rationale behind evaluating the continuity of sleep with survival analysis was previously presented by Norman *et al.* [23]. A continuous sleep period was defined as the interval between the transition to any sleep stage from wakefulness until the next epoch was scored as wake. The mean duration of sleep periods was calculated for each individual in the study population and was used as the time to event (transition to wake) in the survival analyses. The sleep continuity of the OSA groups and the non-OSA group were compared using Cox proportional hazards model with the hazard ratio illustrating the risk for fragmented sleep (i.e., short continuous sleep periods during the night). Furthermore, sleep continuity was studied with Kaplan-Meier survival curves. All statistical analyses were conducted with Matlab 2018b using the Statistics and Machine Learning Toolbox (The MathWorks, Natick, MA, USA).

TABLE II
PERCENTAGE OF SLEEP STAGES WITH DIFFERENT EPOCH-TO-EPOCH DURATIONS IN OBSTRUCTIVE SLEEP APNEA (OSA) SEVERITY GROUPS

Epoch-to-epoch duration	Wake % of recording	N1 % of sleep	N2 % of sleep	N3 % of sleep	REM % of sleep
Non-OSA					
Manual: 30 s	28.8	8.1	48.2	24.5	19.2
Automatic: 30 s	31.2	8.5	49.5	22.2	19.7
15 s	32.6	9.0	49.5	22.4	19.1
5 s	30.9	6.9	50.7	24.8	17.6
1 s	29.4	5.9	51.3	25.7	17.0
0.5 s	28.9	6.0	51.1	25.4	17.5
Mild OSA					
Manual: 30 s	29.3	9.7	49.0	22.5	18.8
Automatic: 30 s	33.8	10.3	51.7	19.8	18.2
15 s	33.0	9.0	53.0	20.6	17.5
5 s	31.0	6.8	52.3	23.7	17.2
1 s	28.6	6.2	52.6	23.7	17.5
0.5 s	27.9	6.3	52.5	23.5	17.7
Moderate OSA					
Manual: 30 s	31.0	13.9	48.5	19.5	18.2
Automatic: 30 s	33.1	10.2	52.4	19.0	18.5
15 s	33.4	9.1	54.1	19.8	17.0
5 s	32.8	8.9	55.3	20.2	15.5
1 s	30.8	7.8	55.2	20.4	16.6
0.5 s	30.1	7.8	55.1	20.0	17.1
Severe OSA					
Manual: 30 s	39.3	22.7	47.0	14.7	15.7
Automatic: 30 s	38.2	14.0	50.7	16.4	18.8
15 s	38.2	12.9	52.3	17.2	17.6
5 s	40.1	13.2	55.0	15.9	15.8
1 s	39.7	13.5	55.5	16.3	14.7
0.5 s	38.8	13.7	55.2	16.3	14.8

III. RESULTS

A. Sleep Stages

The deep learning model reached a training accuracy (Cohen's kappa κ) of 89.2% ($\kappa = 0.85$) and a validation accuracy of 81.9% ($\kappa = 0.76$) during the retraining. In the current study population, the deep learning model had a sleep staging accuracy of 83.2% ($\kappa = 0.77$) when compared to the original manual analysis. This corresponded to accuracies of 91.7% for identifying wake, 41.4% for N1, 84.0% for N2, 83.4% for N3, and 90.9% for REM.

When comparing the deep learning-based sleep staging with traditional 30-second epochs and with varying overlap, the more detailed sleep staging decreased the amount of scored wake, N1, and REM and increased the amount of N2 and N3 in the non-OSA, mild OSA, and moderate OSA groups (Table II). In the mild OSA and moderate OSA groups, the amount of wake first decreased to the same level as with manual scoring with decreasing epoch-to-epoch durations but decreased even further with the shortest durations. In contrast, the amount of wake increased and the amount of N3 and REM decreased in patients with severe OSA with decreasing epoch-to-epoch duration. Examples of scored sleep stages with the traditional sleep staging and when decreasing the epoch-to-epoch duration are shown in Fig. 2.

B. Sleep Parameters

With the deep learning-based sleep staging, the TST and SE increased while WASO decreased in the non-OSA, mild OSA, and moderate OSA groups with decreasing epoch-to-epoch duration (Table III). In contrast, the TST and SE decreased while WASO increased in the severe OSA group with shorter epoch-to-epoch durations thus increasing the differences in the parameter values between severe OSA group and other groups.

TABLE III
MEAN (STANDARD DEVIATION) OF SLEEP PARAMETERS WITH DIFFERENT EPOCH-TO-EPOCH DURATIONS IN OBSTRUCTIVE SLEEP APNEA (OSA) SEVERITY GROUPS

Epoch-to-epoch duration	TST [min]	SE [%]	WASO [min]
Non-OSA			
Manual: 30 s	313.8 (78.8)	71.3 (16.8)	94.9 (60.5)
Automatic: 30 s	303.4 (80.0)	68.6 (16.9)	130.7 (65.3)
15 s	297.1 (78.1)	67.5 (16.7)	139.4 (75.9)
5 s	304.9 (72.3)	69.4 (15.5)	134.9 (72.7)
1 s	311.1 (77.1)	70.8 (16.5)	124.7 (74.4)
0.5 s	313.5 (78.2)	71.3 (16.9)	117.5 (75.7)
Mild OSA			
Manual: 30 s	312.9 (80.7)	70.7 (16.6)	98.5 (60.1)
Automatic: 30 s	292.7 (75.0)	66.4 (16.5)	140.8 (71.4)
15 s	296.4 (72.7)	67.1 (15.1)	138.5 (65.0)
5 s	305.4 (72.7)	69.2 (15.2)	133.7 (72.4)
1 s	315.8 (76.7)†	71.4 (15.7)†	122.5 (71.9)†
0.5 s	318.7 (78.5)†	72.1 (16.0)†	113.9 (66.1)†
Moderate OSA			
Manual: 30 s	306.8 (79.5)	69.0 (16.6)	106.3 (59.3)
Automatic: 30 s	297.3 (77.1)	66.8 (15.5)	135.7 (66.0)
15 s	296.3 (83.8)	66.5 (17.2)	142.4 (73.9)
5 s	298.9 (73.6)	67.2 (14.6)	143.0 (66.7)
1 s	307.8 (78.5)	69.2 (16.0)	132.2 (73.4)
0.5 s	310.8 (79.7)	69.9 (16.4)	120.5 (69.6)†
Severe OSA			
Manual: 30 s	267.6 (84.7)*	60.8 (18.6)*	139.8 (70.5)*
Automatic: 30 s	272.6 (84.6)*	61.8 (17.6)*	155.7 (74.7)*
15 s	272.3 (73.7)*	62.0 (16.3)*	159.7 (73.6)*
5 s	264.0 (75.5)*	60.1 (16.5)*	171.9 (75.9)*
1 s	266.0 (83.2)*	60.5 (18.4)*	171.0 (85.0)*
0.5 s	269.7 (84.2)*	61.3 (18.7)*	161.1 (78.7)*

TST = total sleep time, SE = sleep efficiency, WASO = wake after sleep onset. A statistically significant ($p < 0.05$) difference between an OSA severity group and the non-OSA group is denoted with an asterisk (*). A statistically significant ($p < 0.05$) difference between different epoch-to-epoch durations compared to automatic sleep staging with traditional 30-second epoch-to-epoch duration is denoted with a dagger (†).

C. Sleep Continuity

With 30-second epoch-to-epoch duration in the deep learning-based sleep staging, the differences in the sleep fragmentation between the groups were small based on the Cox proportional hazards model or Kaplan-Meier survival curves (Table IV, Fig. 3). When the sleep staging was conducted in more detail with shorter epoch-to-epoch durations, differences between the non-OSA and the OSA severity groups began to increase. The hazard ratios illustrating the risk of fragmented sleep with 1-second epoch-to-epoch duration were 1.14 ($p = 0.39$), 1.59 ($p < 0.01$), and 4.13 ($p < 0.01$) in mild, moderate and severe OSA groups, respectively. The obtained hazard ratio for the severe OSA group even surpassed the value obtained with the manual sleep staging indicating that the deep learning-based sleep staging with short epoch-to-epoch duration reveals larger differences

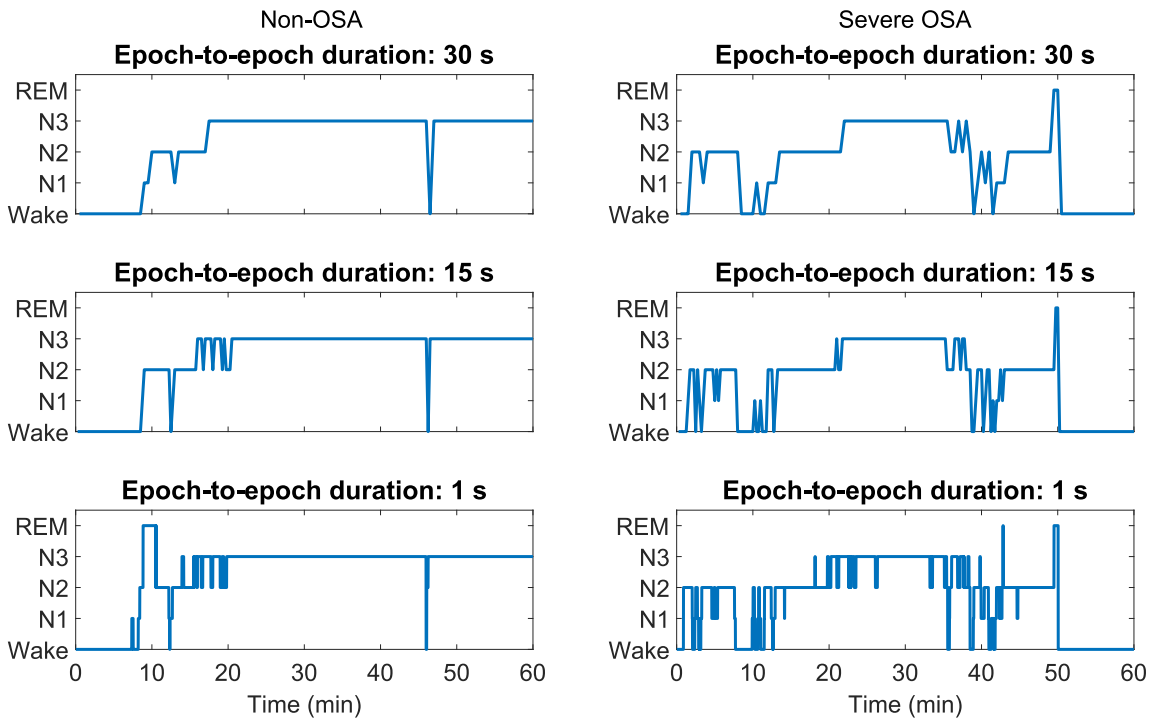

Fig. 2. Examples of an hour of scored sleep stages for a patient with no obstructive sleep apnea (OSA, left, a 64-year-old female with an apnea-hypopnea index of 2.2) and with severe OSA (right, a 65-year-old male with an apnea-hypopnea index of 36.1) with the automatic sleep staging based on different epoch-to-epoch durations.

between the non-OSA and severe OSA group. Similar differences between groups with decreasing epoch-to-epoch duration can be seen in the Kaplan-Meier survival curves (Fig. 3).

IV. Discussion

In this study, we introduced a novel method to analyze sleep in a more detailed manner using deep learning-based automatic sleep staging. Our results reveal that reducing the epoch-to-epoch duration between consecutive 30-second epochs considerably affects the evaluated sleep architecture and can provide greater insights into the sleep architecture beyond the traditional 30-second epoch-to-epoch duration. The results further reveal the highly fragmented sleep architecture of patients suffering from severe OSA. Overall, the results suggest that based on more detailed sleep analysis, severe OSA patients have considerably less REM sleep and slightly less N3 sleep than estimated via traditional epochs while the amount of N2 and wakefulness during the night is higher. Similarly, total sleep time and sleep efficiency decreased with shorter epoch-to-epoch durations. Finally, the results with our detailed sleep staging approach expose larger differences in the sleep continuity between individuals without OSA and individuals in different OSA severity categories.

In the non-OSA group, the amount of wakefulness, N1 sleep and REM decreased with shorter epoch-to-epoch durations. At the same time, the amount of N2 and N3 increased. Similar behavior was observed in the population with mild or moderate OSA. Conversely, the severe OSA group differed from the other groups: the amount of wakefulness and N1 increased while the amount of REM and N3 decreased along decreasing epoch-to-epoch duration. This illustrates that the sleep architecture of severe OSA patients is more disrupted than estimated with

TABLE IV
Hazard Ratios, Confidence Intervals (95% CI), and *p*-Values of Fragmented Sleep in Obstructive Sleep Apnea (OSA) Severity Groups

Epoch-to-epoch duration	Hazard ratio	95% CI	*p*
	Mild OSA		
Manual: 30 s	1.20	0.90-1.61	0.21
Automatic: 30 s	1.36	1.02-1.83	0.04
15 s	1.16	0.86-1.56	0.32
5 s	1.03	0.77-1.39	0.83
1 s	1.14	0.85-1.53	0.39
0.5 s	1.17	0.87-1.57	0.29
	Moderate OSA		
Manual: 30 s	1.68	1.24-2.28	<0.01
Automatic: 30 s	1.13	0.84-1.52	0.43
15 s	1.22	0.90-1.65	0.21
5 s	1.38	1.02-1.87	0.04
1 s	1.59	1.17-2.15	<0.01
0.5 s	1.65	1.21-2.24	<0.01
	Severe OSA		
Manual: 30 s	3.73	2.78-5.00	<0.01
Automatic: 30 s	2.00	1.50-2.67	<0.01
15 s	1.54	1.16-2.06	<0.01
5 s	2.64	1.97-3.54	<0.01
1 s	4.13	3.06-5.57	<0.01
0.5 s	3.99	2.96-5.38	<0.01

the traditional 30-second epoch-to-epoch duration, with more wakefulness and N1 sleep present during the night. A similar effect can be observed in the total sleep time and sleep efficiency, which decreased in the severe OSA group along with decreasing epoch-to-epoch durations.

Based on these results, the traditional approach based on 30-second epoch-to-epoch duration may be suitable for a healthy population but does not provide the necessary detail for assessing the sleep of patients suffering from sleep disorders. Our more detailed sleep staging approach would appear to provide a more realistic representation of the highly disrupted sleep architecture which is easily overlooked when using the traditional non-overlapping 30-second epochs. This could ultimately lead to a more informed diagnosis of various sleep disorders and their effects on sleep architecture. Moreover, further studies linking detailed sleep architecture to daytime symptoms, cardiovascular risks, therapeutic outcomes, and perceived sleep quality are warranted.

In our sleep continuity analyses, only small differences between the healthy population and different OSA groups were seen using the traditional non-overlapping 30-second epoch approach. In contrast, the more detailed sleep staging approach revealed larger differences in the sleep continuity between the OSA groups and the healthy population, even surpassing manual scoring. For example, with 1-second epoch-to-epoch duration, the hazard ratio illustrating the risk of fragmented sleep was 1.14 ($p = 0.39$) for mild OSA, 1.59 ($p < 0.01$) for moderate OSA, and 4.13 ($p < 0.01$) for severe OSA. This shows that the risk of fragmented sleep increases with increasing OSA severity. Similarly, the Kaplan-Meier survival curves (Fig. 3) show that differences between all OSA groups become more apparent with more detailed sleep staging. However, it must be noted that with decreasing epoch-to-epoch duration, the mean duration of continuous sleep decreased in all the groups, as can be seen from the Kaplan-Meier curves. This is expected as the overlapping epochs provide a way to assess sleep architecture in a more detailed manner capturing more transitions to wakefulness during the night. Moreover, decreasing the epoch-to-epoch duration even further to 0.5 seconds produced slightly smaller differences between the OSA severity groups. This may be due to too small differences between adjoining epochs or due to the small uncertainty always related to sleep staging; that is, epochs on the verge of being scored to wake may falsely be scored as such with short epoch-to-epoch durations. Investigating this effect and finding the optimal epoch-to-epoch duration warrants further studies.

We hypothesized that the current sleep staging procedure underestimates the degree of sleep fragmentation caused by OSA. These results support our hypothesis in severe OSA patients and only to some extent in mild and moderate OSA patients. The percentage of sleep stages, total sleep time, sleep efficiency, and wake after onset were similar in non-OSA patients and patients with mild or moderate OSA. However, the survival analysis-based assessment of sleep continuity also revealed differences between non-OSA patients and patients with mild or moderate OSA. This can be seen both in the Kaplan-Meier survival curves (Fig. 3), and the Cox regression (Table IV). A similar effect was also seen by Norman *et al.* who reported that no significant differences exist between the normal and mild OSA groups in traditional sleep parameters and differences only emerge when considering the sleep continuity with survival analysis [23]. However, it has to be noted that the division into OSA severity groups is highly artificial and simplistic and the severity assessment with AHI might not sufficiently reflect the physiological effects of OSA [24]–[26]. Therefore, it could be beneficial to study how sleep fragmentation varies when defining the OSA severity differently or even attempt to define the severity of OSA by using sleep fragmentation as a metric. Regardless, a more detailed analysis of sleep provides more insight into the sleep architecture and could be highly useful when assessing the sleep of OSA patients, and could supplement the evaluation of disease severity.

Our approach for detailed sleep analysis was based on overlapping 30-second epochs, which is both a strength and a limitation in the present study. The developed approach benefits from decades of clinical practice of sleep staging, is easily applicable to daily work, and the results can be interpreted similarly as with traditional manual sleep staging. However, the detailed analysis was still based on identifying a sleep stage for each epoch and does not provide a continuous scale of sleep depth in that sense. However, this approach allows comparison with the traditional 30-second epoch-based sleep staging and eases the interpretation of results over an arbitrary, continuous scale of sleep depth which has been previously attempted based on EEG frequency content [4], [6]. The main advantage of the developed method over traditional sleep staging is the capability to observe the transitions between the discrete stages with better temporal resolution. However, further studies are warranted to conduct similar approaches using shorter epoch durations without overlap to gain a deeper understanding of sleep microstructure. Furthermore, we only investigated the transitions between sleep stages and did not consider arousals from sleep. Arousals were discarded as the reliability of arousal scoring can be relatively low [27] and to focus solely on the sleep staging process and on how it could be improved. Therefore, future studies investigating the effect of arousals alongside the more detailed sleep staging are warranted.

The deep learning-based automatic sleep staging method was trained using manual sleep stage scoring. This manual scoring material can understandably suffer from human error and differences between scoring traditions of different scorers. However, all the manual scorers involved in scoring the study material participate regularly in intra- and inter-laboratory scoring concordance activities. Furthermore, in a previous study on inter-rater reliability at the sleep laboratory, the mean Cohen's kappa (standard error of the mean) of sleep staging was 0.74 (0.02) [28] illustrating high reliability between the scorers. Furthermore, the use of the deep learning approach can be considered as one of the biggest strengths of our study. In contrast to manual scoring, the deep learning-based sleep staging is always conducted consistently and all the scorings are highly comparable. This is also the rationale behind why the comparison between traditional and detailed sleep staging was possible. The scoring of the same recordings with different approaches would

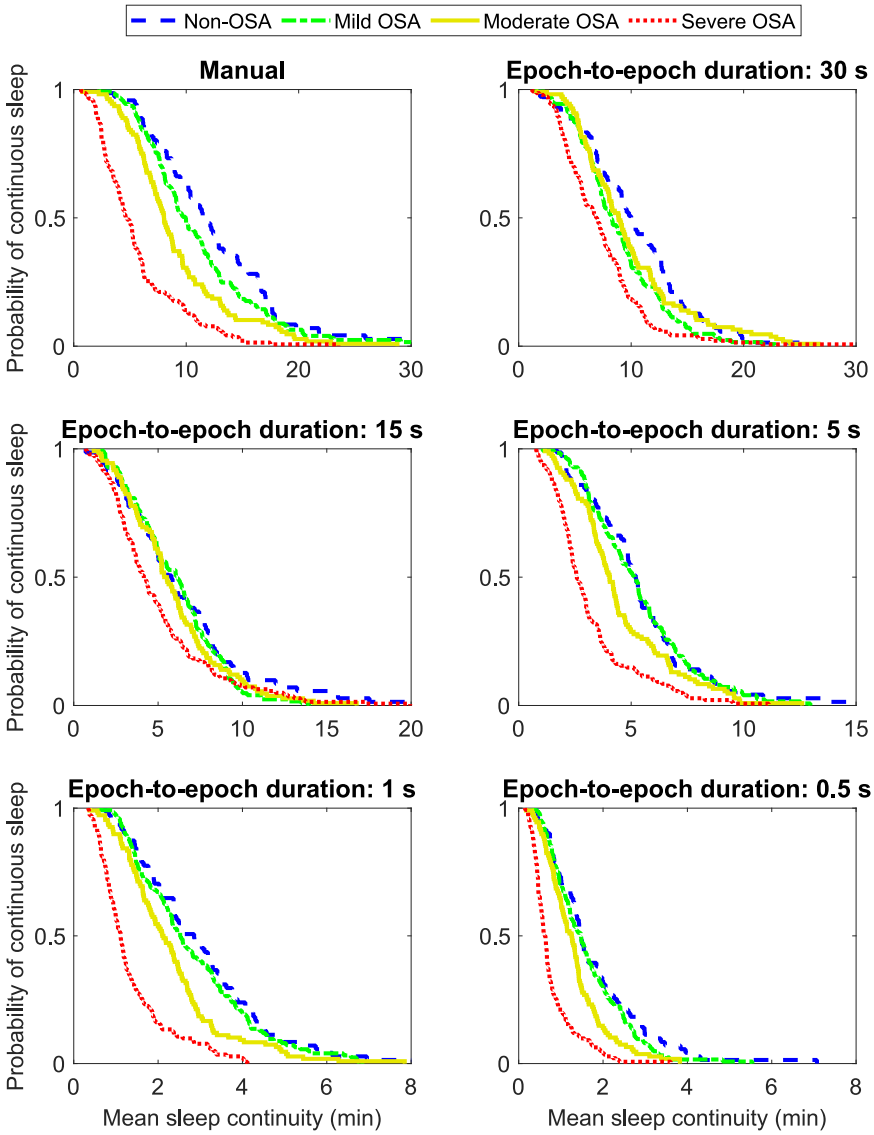

Fig. 3. Kaplan-Meier survival curves of mean continuous sleep durations for each obstructive sleep apnea (OSA) severity group with manual scoring and automatic scoring using different epoch-to-epoch durations.

have been highly biased and laborious with manual scoring. Furthermore, our method can alleviate the biggest limitations of the manual sleep staging with a fast, and easily applicable method requiring no increase in working hours spent currently in clinical practice. Automatic sleep staging could reduce the workload and simultaneously produce the traditional sleep staging alongside the more detailed representation with overlapping epochs in a timely manner, generally in less than a minute.

V. Conclusion

More detailed sleep staging using a deep learning-based automatic method is a highly promising approach to gain further insight into the characteristics of the fragmented sleep architecture of patients suffering from sleep disorders. The detailed sleep staging emphasized the highly disrupted sleep architecture of patients with severe OSA which can be vastly underestimated with traditional sleep staging. These results highlight the need for a more detailed analysis of sleep architecture in daily clinical practice.

References

[1] S. Diekelmann and J. Born, "The memory function of sleep," *Nature Rev. Neurosci.*, vol. 11, no. 2, pp. 114–126, 2010.
[2] N. E. Fultz *et al.*, "Coupled electrophysiological, hemodynamic, and cerebrospinal fluid oscillations in human sleep," *Science*, vol. 366, no. 6465, pp. 628–631, 2019.
[3] R. B. Berry *et al.*, "The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications," Version 2.5, American Academy of Sleep Medicine, Darien, IL, USA, 2018.
[4] M. Younes *et al.*, "Odds ratio product of sleep EEG as a continuous measure of sleep state," *Sleep*, vol. 38, no. 4, pp. 641–654, 2015.
[5] S. L. Himanen and J. Hasan, "Limitations of Rechtschaffen and Kales," *Sleep Med. Rev.*, vol. 4, no. 2, pp. 149–167, 2000.
[6] M. H. Asyali, R. B. Berry, M. C. K. Khoo, and A. Altinok, "Determining a continuous marker for sleep depth," *Comput. Biol. Med.*, vol. 37, no. 11, pp. 1600–1609, Nov. 2007.

[7] J. Pardey, S. Roberts, L. Tarassenko, and J. Stradling, "A new approach to the analysis of the human sleep/wakefulness continuum," *J. Sleep Res*, vol. 5, no. 4, pp. 201–210, 1996.
[8] A. Rechtschaffen and A. Kales, *A Manual of Standardized Terminology, Techniques and Scoring System of Sleep Stages in Human Subjects*. Los Angeles, CA, USA: University of California, Brain Information Service/Brain Research Institute, 1968.
[9] A. V. Benjafield *et al.*, "Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis," *Lancet Respir. Med*, vol. 7, no. 8, pp. 687–698, 2019.
[10] AASM, "Sleep–related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research," *Sleep*, vol. 22, pp. 667–689, 1999.
[11] R. J. Kimoff, "Sleep fragmentation in obstructive sleep apnea," *Sleep*, vol. 19, no. suppl_9, pp. S61–S66, 1996.
[12] H. Phan, F. Andreotti, N. Cooray, O. Chén, and M. de Vos, "SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 400–410, Jan. 2019.
[13] A. Patanaik, J. L. Ong, J. J. Gooley, S. Ancoli-Israel, and M. W. L. Chee, "An end-to-end framework for real-time automatic sleep stage classification," *Sleep*, vol. 41, no. 5, pp. 1–11, 2018.
[14] S. Biswal, H. Sun, B. Goparaju, M. Brandon Westover, J. Sun, and M. T. Bianchi, "Expert-level sleep scoring with deep neural networks," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 12, pp. 1643–1650, 2018.
[15] A. Malafeev *et al.*, "Automatic human sleep stage scoring using deep neural networks," *Front. Neurosci.*, vol. 12, Nov. 2018, p. 781.
[16] H. Korkalainen *et al.*, "Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea," *IEEE J. Biomed. Heal. Inform.*, vol. 24 no. 7, pp. 2073–2081, Jul. 2020.
[17] H. Korkalainen *et al.*, "Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea," *Sleep*, vol. 43, no. 11, 2020, Art. no. zsaa098.
[18] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberyé, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, Sep. 2000.
[19] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e220, 2000.
[20] H. Danker-Hopfe *et al.*, "Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard," *J. Sleep Res*, vol. 18, no. 1, pp. 74–84, 2009.
[21] U. J. Magalang *et al.*, "Agreement in the scoring of respiratory events and sleep among international sleep centers," *Sleep*, vol. 36, no. 4, pp. 591–596, 2013.
[22] R. G. Norman, I. Pal, C. Stewart, J. A. Walsleben, and D. M. Rapoport, "Interobserver agreement among sleep scorers from different centers in a large dataset," *Sleep*, vol. 23, no. 7, pp. 901–908, 2000.
[23] R. G. Norman, M. A. Scott, I. Ayappa, J. A. Walsleben, and D. M. Rapoport, "Sleep continuity measured by survival curve analysis," *Sleep*, vol. 29, no. 12, pp. 1625–1631, 2006.
[24] H. Korkalainen, J. Töyräs, S. Nikkonen, and T. Leppänen, "Mortality-risk-based apnea–hypopnea index thresholds for diagnostics of obstructive sleep apnea," *J. Sleep Res*, vol. 28, no. 6, pp. 1–7, 2019.
[25] S. Kainulainen *et al.*, "Severity of desaturations reflects OSA-related daytime sleepiness better than AHI," *J. Clin. Sleep Med.*, vol. 15, no. 8, pp. 1135–1142, 2019.
[26] S. Kainulainen *et al.*, "Severe desaturations increase PVT-based median reaction time and number of lapses in OSA patients," *Eur. Respir. J.*, vol. 55, no. 4, 2020, Art. no. 1901849.
[27] M. J. Drinnan, A. Murray, C. J. Griffiths, and G. J. Gibson, "Interobserver variability in recognizing arousal in respiratory sleep disorders," *Amer. J. Respir. Crit. Care Med.*, vol. 158 pp. 358–362, 1998.
[28] B. Duce, C. Rego, J. Milosavljevic, and C. Hukins, "The AASM recommended and acceptable EEG montages are comparable for the staging of sleep and scoring of EEG arousals," *J. Clin. Sleep Med.*, vol. 10, no. 7, pp. 803–809, 2014.