Towards a user-friendly sleep staging system for polysomnography part II: Patient-dependent features extraction using the SATUD system

Jade Vanbuis, Mathieu Feuilloy, Lucile Riaboff, Guillaume Baffet, Nicole Meslier, Frédéric Gagnadoux, Jean-Marc Girault

Version of Record: https://www.sciencedirect.com/science/article/pii/S2352914820306031
Manuscript_2451f35cec07dd5345d4f809bf72019e

Towards a user-friendly sleep staging system for polysomnography Part I: automatic classification based on medical knowledge

Jade Vanbuis[a,b,*], Mathieu Feuilloy[a,b], Guillaume Baffet, Nicole Meslier[c,d], Frédéric Gagnadoux[c,d] and Jean-Marc Girault[a,b]

[a]*ESEO, Angers, France*
[b]*LAUM, UMR CNRS 6613, Le Mans, France*
[c]*Angers sleep laboratory, University Hospital, Angers, France*
[d]*INSERM UMR 1063, University of Angers, Angers, France*

ARTICLE INFO

Keywords:
Automatic sleep staging for polysomnography
Decision support system
User-friendly and interpretable sleep scoring
Patient-dependent sleep scoring using the SATUD system
Respect of AASM Guidelines

ABSTRACT

Manual sleep scoring is a time-consuming task that requires a high level of medical expertise. For this reason, a number of automatic sleep scoring algorithms have recently been implemented. However, their use by physicians remains limited for various reasons: a lack of transparency of the approach used, insufficient heterogeneity among the patients used for testing, or a lack of practicality.
This paper presents a system for facilitated sleep scoring that will overcome these limitations. The proposed system, a user-friendly tool based on electrophysiological channels, was trained and tested on large datasets of 300 and 100 distinct recordings from patients with various sleep disorders. The method replicates the manual sleep scoring process, in accordance with the American Academy of Sleep Medicine (AASM) guidelines and generates patient-dependent sleep scoring (using the SATUD system). For an improved level of precision and confidence with regard to scoring, our approach also provides a table that gives indications about the confidence level of the algorithm when scoring sleep. In contrast to recent deep learning approaches, the algorithms used were chosen for their resilience and as they are easy to understand. Medical knowledge was included in the process as much as possible. Results showed that the system is consistent with manual scoring (mean Cohen's Kappa of 0.69 and accuracy rate of 77.8 %). It proves that a facilitated interpretation of the model, crucial in such fields as sleep diagnosis, can be provided when using automatic tools.
This new system thereby generates sleep scoring decision support tools, which should easily contribute to significant time-saving and help sleep specialists to perform sleep diagnosis.

1. Introduction

Sleep-disordered breathing (SDB) is a common health issue affecting approximatively a third of the population [1–3]. Symptoms often go unnoticed, since they are not specific to SDB and are quite common [4]. However, bad sleep quality can affect several vital functions, such as learning, memorization and adaptation, resulting in a deterioration of the quality of life.
Over the past few decades, there has been an increasing need for sleep diagnosis [5, 6]. The gold-standard procedure for SDB diagnosis, called polysomnography (PSG), involves the recording of electrophysiological (EP) and cardio-respiratory (CR) signals throughout an entire night [7]. Once recorded, signals are manually studied by a sleep specialist: respiratory events are identified using CR channels, and sleep is scored using EP channels. A diagnosis is reached by cross-checking this information with patient symptoms [7].
Sleep scoring is a time-consuming and complex task. It involves the assessment of each 30-second section's (called an epoch) degree of vigilance [7]. To do so, EP channels such as electroencephalograms (EEG), electrooculograms (EOG) and electromyograms (EMG) are visualized epoch by epoch. Each epoch is identified as belonging to the W stage (wakefulness), the N1 stage (light sleep), the N2 stage (also light sleep), the N3 stage (deep sleep) or the R stage (rapid eye movement sleep). The resulting succession of sleep stages is called a hypnogram. The American Academy of Sleep Medicine (AASM) manual for the scoring of sleep and associated events [7] describes each sleep stage's properties in detail, along with possible transitions between sleep stages. Despite the AASM guidelines, sleep staging remains time-consuming and complex, and the inter-scorer agreement rate hardly exceeds 80-90 % [8].

Lately, artificial intelligence and more specifically learning algorithms have proven their ability to solve complex problems in many healthcare sectors [9, 10]. New algorithms for automatic events or sleep analysis have emerged and been recognized by experts in the field for potentially helping to improve our understanding of sleep [11] and simplifying the scoring procedure [12].
A number of systems for automatic sleep scoring using EP signals have been developed. In such systems, algorithms are trained so they can classify each epoch into a sleep stage, using manual scoring as the reference. The algorithms developed can be broken down into three categories: deep learning [13], machine learning [14] and hybrid approaches [15]. Deep learning is generally applied directly on raw signals. In [16, 17], a combination of a Convolutional Neural Network (CNN) with a Recurrent Neural Network (RNN) is used to estimate relevant features and classify sleep, taking tempo-

*Corresponding author
jade.vanbuis@eseo.fr
ORCID(s): 0000-0001-6437-1597 (J. Vanbuis)

rality into consideration.
On the contrary, machine learning usually requires the extraction of descriptors, called features, before classification. Various machine learning classifiers were tested. Among the most widespread, we will mention Support Vector Machines (SVM) in [18–20], Multi-Layer Perceptrons (MLP) in [21, 22] and Random Forests (RF) in [23, 24].
Lastly, hybrid approaches combine deep learning or machine learning with expert knowledge. In [25, 26], symbolic fusion was applied so the features used for classification are qualitative, particularly close to the AASM guidelines.

However, automatic sleep staging faces several challenges. Firstly, the channels used for classification vary greatly from one patient to another. This high variability between subjects is caused by many factors including the subject's age, condition, drug intake, but also the positioning of sensors, the quality of signals (which can be altered by movements or sweating, for example) and the presence of accessories. It particularly complicates the learning process for machine learning approaches, where features are highly impacted by subject specific characteristics. This can be overcome if trained on a large number of subjects with various disorders. In [27], a review of 154 deep learning-based EEG analysis studies showed that half of them included less than 13 subjects, which is reported as being insufficient to illustrate human heterogeneity.
Another challenge regarding machine learning or deep learning approaches is acceptance by the medical community. A recent review paper by Fiorillo et al. [12] presented the barriers for the clinical use of automated scoring on a daily basis. The main limitation was the black box behavior of deep learning algorithms. Nowadays, a certain number of researchers attempt to improve the interpretability of their models [28], using, for example, hybrid approaches incorporating medical knowledge.
The final criteria is for the model to be easy for physicians to use. There is often a preliminary human action needed (for example, identification of artifacted epochs or partial scoring), meaning that methods cannot be applied immediately once the channels have been recorded.
In summary, the clinical use of automatic sleep scoring remains controversial because of three limitations:

a) lack of confidence in the developed approach (algorithms are often considered as a black box);
b) insufficient heterogeneity of the dataset, nonetheless necessary for assessing real-life performances;
c) lack of practicality of the developed approach, which sometimes requires human intervention before use.

The main aim of this study is to implement a user-friendly automatic sleep scoring system to overcome the three previous limitations. Unlike most of the recent studies, we chose to prioritize the understanding of the algorithm's operating mode, addressing issue *a)*. To provide an answer to limitation *b)*, the system was tested on a large dataset of patients with varying levels of SDB severity. Issue *c)* was also taken into consideration, since the system was designed to be used without any preliminary human action (for example partial sleep scoring or invalidation of epochs) and provides a probability table to further assist in scoring.
The work proposed here has been designed on EP channels obtained from polysomnographic recordings. Designed to assist physicians in their diagnosis, the developed system combines artificial intelligence and expert knowledge. Medical practitioners' concerns were considered and the result is a user-friendly tool providing scoring support to avoid sleep scorers spending excessive time on sleep staging.

The present article is the first of a two-part paper. In the following section (Section 2), the recordings used for training and testing are presented first of all. The algorithm architecture is then detailed and shows how the three limitations have been addressed. The SATUD algorithm, also introduced in this section, is detailed in the companion article [REF]. The elements for system evaluation are then presented. In Section 3, the system is compared with manual scoring and its performance is reported. Its results and impact on sleep diagnosis are discussed in Section 4. Finally, the conclusion is presented in Section 5.

2. Methods and materials

First of all, this section presents the database on which the algorithm was trained and tested. The methodology used in the system is then detailed, followed by a presentation of the system performance assessment.

2.1. Data acquisition

A total of 400 anonymous sleep recordings were included in this study thanks to the sleep cohort of Pays de La Loire. This cohort is operated under the aegis of the Institut de Recherche en Santé Respiratoire. Approval was obtained from the University of Angers ethics committee and the "Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé" (CCTIRS; 07.207bis). The recordings used in this study were acquired between 2012 and 2018, and all patients gave their written informed consent. The 400 recordings were divided into two datasets (D1 and D2). A random selection was made, ensuring equal representation of the severity of Obstructive Sleep Apnea (OSA) and the year of recording (see Table 1). D1 was made up of 300 recordings from 182 males and 117 females while D2 was made up of 100 recordings from 66 males and 34 females. Recordings from D1 were used as the training dataset and D2 as the test dataset. D1 dataset was made up of 329,911 epochs (W: 76,897 - R: 52,036 - N1: 24,283 - N2: 122,266 - N3: 54,429). As for D2 dataset, it was made up of 110,978 epochs (W: 24,172 - R: 18,194 - N1: 8,095 - N2: 41,095 - N3: 19,422).
The subjects, who were suspected of having OSA, underwent one-night's PSG in the sleep laboratory of Angers University Hospital (FRANCE). Sleep was recorded following the AASM guidelines [7]. The CID102L8D polysomnograph (CIDELEC St Gemmes-sur-Loire, FRANCE) used pro-

Table 1
OSA severity represented by age, using quartiles, evaluated for D1 and D2.

	D1 dataset				D2 dataset			
	Q1 19-43 y.o.	Q2 44-53 y.o.	Q3 54-62 y.o.	Q4 63-86 y.o.	Q1 19-41 y.o.	Q2 42-53 y.o.	Q3 54-63 y.o.	Q4 64-79 y.o.
No	32 %	18 %	9 %	5 %	44 %	8 %	8 %	8 %
Mild	31 %	23 %	23 %	25 %	32 %	28 %	20 %	24 %
Moderate	17 %	24 %	30 %	36 %	12 %	36 %	36 %	24 %
Severe	20 %	35 %	38 %	34 %	12 %	28 %	36 %	44 %

y.o. = years old

vided the usual electrophysiological (EP) and cardio-respiratory (CR) signals. It also included the PneaVoX® sensor, from which tracheal sounds and respiratory efforts are estimated to facilitate event scoring [29–31]. Once the signals were acquired, each sleep recording was manually scored by a single sleep specialist in accordance with AASM guidelines, although several sleep specialists were involved in this study. The hypnogram established by the sleep specialist was considered as our reference, and was referred to as $hypno_{ref}$ in the rest of this paper. Unlike other studies, epochs with artifacts were not discarded (neither manually nor automatically) to ensure the algorithm's efficiency in real-life conditions. The only epochs rejected were those with extremely bad quality signals preventing the manual scoring of events and/or sleep (for example epochs with missing signals). They were automatically invalidated by the CIDELEC user interface prior to scoring.

2.2. Algorithm structure

The algorithm presented in this section was implemented using Matlab® software to provide sleep scoring support tools. It was designed based on the scoring rules described in the AASM guidelines [7], and behaves similarly to the manual scoring process. Its inputs are EP signals and a priori medical knowledge (AASM guidelines). It classifies all epochs to provide an automatic hypnogram $hypno_{EP}$, with a probability table $probabilities_{EP}$. This table gives information about the confidence level of the algorithm when epochs were classified. Figure 1 illustrates the algorithm structure, described in this section. The architecture is composed of several main functions: **F1**, **F2**, **F3** and **F4**. Each function aims to reproduce one of the tasks realized by sleep specialists when scoring sleep. Firstly, an adaptation to each recording is achieved and provides patient-dependent features (**F1**). Using those features, a rough hypnogram is estimated (**F2**). Similarly to the manual scoring process, the hypnogram is then adjusted with regard to sleep patterns (identified in **F3**), surrounding epochs and transition rules (**F4**).

2.2.1. *F1 - The SATUD system*

Before scoring sleep stages, sleep specialists visualize all epochs to adapt their scoring to the patient's specific characteristics. This process was implemented automatically with the SATUD system, which does not require any training/test steps. The SATUD system was thus applied to all D1 and D2 recordings individually.
The SATUD system is fully presented in the companion paper [REF], in which its functioning and a simplified example of its use for sleep stage classification are detailed.
Briefly, the SATUD system extracted 41 patient-specific qualitative features from EP channels, using a priori medical knowledge obtained from the AASM guidelines. The 41 patient-

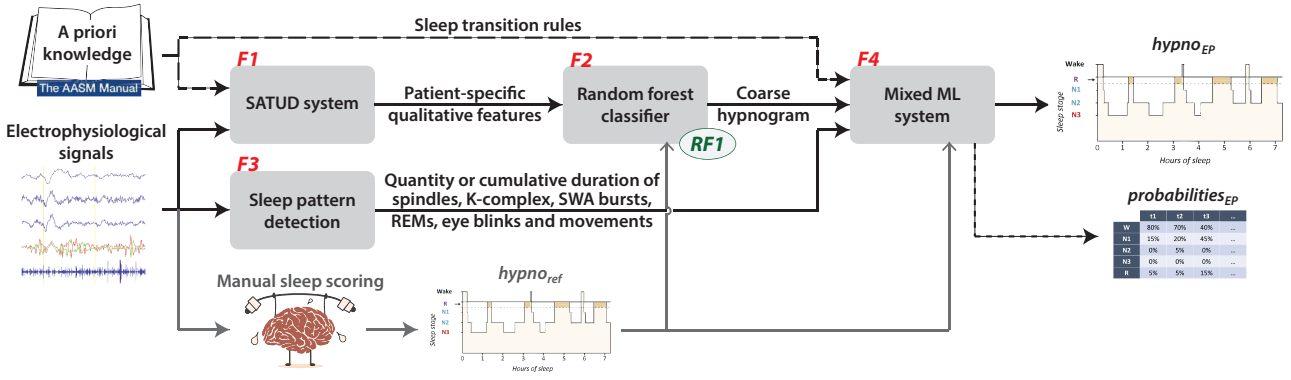

Figure 1: Functional architecture of the user-friendly automatic sleep staging system, composed of four main functions (F1, F2, F3 and F4). Medical knowledge from the AASM manual and electrophysiological signals are used as inputs. The outputs are a hypnogram $hypno_{EP}$ and the associated probability table $probabilities_{EP}$.

specific qualitative features were defined as the levels of various sleep stage descriptors (quantitative features), as presented in Table 2. For example, the AASM guidelines mention the importance of the EEG amplitude (1st line of Table 2) to score sleep stages. EEG amplitude is the highest in N3 and the lowest in N1. It is also generally higher in N2 than in W and R. However, EEG amplitude in R can increase when there are some artifacts (called Rapid Eye Movement artifacts). It can also be really high in W when the patient is agitated (due to movements). With regard to these elements, we decided to use 2 thresholds so we can associate EEG amplitude levels (or qualitative features) with stages as following: N3→High, N2→Mid or High, R→Low or Mid and N1→Low (nothing for W as it can be Low, Mid or High). As there was no need for a Mid EEG amplitude level alone, we did not compute it.

Table 2
List of the 41 sleep stages' qualitative features (right column) obtained using the SATUD system. Each qualitative feature corresponds to a level associated with a descriptor or quantitative feature (left column), identified as relevant for sleep stage scoring using the AASM guidelines.

Quantitative features	Th[a]	Qualitative features used	N[b]
EEG amplitude	2	Low | Low or Mid | Mid or High | High	4
EEG instability	1	No | Yes	2
Slow wave activity quantity	2	Low | Low or Mid | High	3
Alpha waves quantity	2	Low | Low or Mid | Mid | Mid or High | High	5
Beta waves quantity	2	Low | Mid | Mid or High	3
Delta waves quantity	2	Low | Mid or High	2
Theta waves quantity	2	Low | Low or Mid | Mid or High	3
Chin level	2	Low | Low or Mid | Mid or High | High	4
Chin instability	2	Low | Low or Mid | Mid or High | High	4
Summed EOG level	2	Low | Low or Mid | Mid or High	3
Summed EOG instability	2	Low | Low or Mid | Mid or High	3
Substracted EOG level	2	Low | Mid or High	2
Substracted EOG instability	2	Low | Mid or High | High	3
		Total	41

[a] Number of Thresholds used.
[b] Number of qualitative features used for each quantitative feature.

EEG waves (lines 3 to 7 of Table 2) are usually evaluated using frequency ranges with fixed boundaries. In this work, those ranges were redefined for each recording to adapt even further to each patient. In particular, the alpha waves frequency range during wakefulness was adjusted to avoid W overestimation[1].

[1] Alpha waves are representative of the W stage. For some patients, they can also occur in the R stage, or even throughout the entire recording in the case of alpha-delta sleep [32] patients.

2.2.2. F2 - Random forest classifier

Using the 41 patient-specific qualitative features obtained with the SATUD system (F1) as input, F2 aims to generate an initial hypnogram. This hypnogram was referred to as a 'coarse hypnogram' in the rest of this paper, since it will be used to obtain a more precise hypnogram (in F4). Because F2 required training/test steps, all D1 features were concatenated into a single large training matrix. In the same way, all D1 references $hypno_{ref}$ were concatenated.
The implemented classifier was a random forest [33], a machine learning algorithm that combines decision trees. This model was chosen because it is powerful, robust, and not opaque like deep learning methods. The random forest was developed using the TreeBagger function from Matlab®. This function bagged 100 classification trees using bootstrap samples of the data and randomly selecting a subset of 6 features at each node.
Once the model had been trained, it was saved under the name ***RF1*** to be reused for each test recording individually, resulting in one coarse hypnogram per D2 recording.

2.2.3. F3 - Sleep pattern detection

Besides continuous features, so-called 'sleep patterns' are essential for sleep scoring. F3 aims to identify a majority of them within EP signals, using their description as mentioned in the AASM guidelines. Sleep spindles, K-complex, Slow Wave Activity (SWA) bursts, Rapid Eye Movements (REMs), eye blinks and movements were detected using signal processing algorithms previously implemented and not detailed in the present paper (filters, wavelets, empirical reasoning, etc.). For each sleep pattern, depending on its nature, its quantity or cumulative duration was computed for each half-epoch (for better respect of the AASM guidelines). The resulting features were computed for each recording individually (D1 and D2), and will be used to enhance the coarse hypnogram.

2.2.4. F4 - Mixed ML system

The last major element for sleep scoring are sleep transition rules. As described in the AASM guidelines, efficient sleep scoring requires the knowledge of surrounding epochs (especially for N2 and R sleep stages). F4 is a mixed machine learning system that combines the coarse hypnogram (obtained in F2), sleep pattern quantity or duration (obtained in F3) and a priori knowledge of sleep transition rules (obtained from the AASM guidelines). F4 outputs are the final hypnogram $hypno_{EP}$, along with the probability table $probabilities_{EP}$. Figure 2 illustrates F4 structure. Its architecture is composed of several main functions: **F4.1**, which estimates a hypnogram using the coarse hypnogram and sleep patterns by taking temporality into consideration, **F4.2** and **F4.3**, which adjust and correct the obtained hypnogram ensuring that there are no forbidden epoch sequences, and **F4.4**, which is an independent function using the coarse hypnogram and sleep patterns to provide a supplementary scoring support tool.

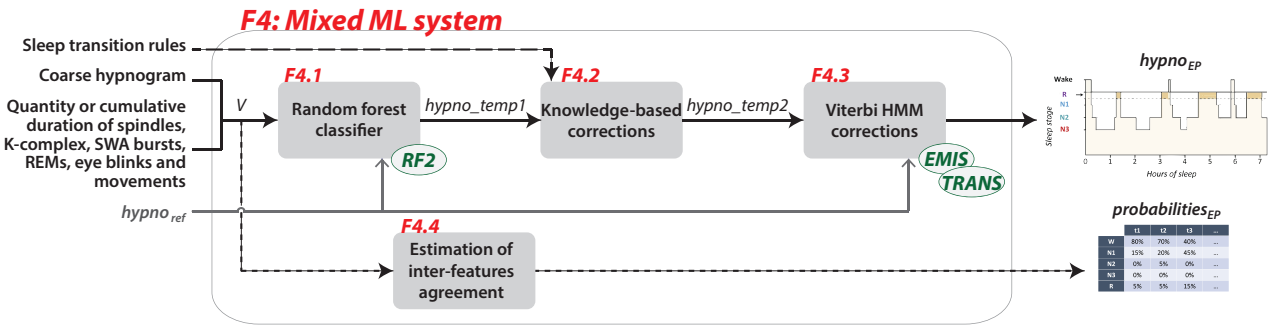

Figure 2: Functional architecture of the mixed ML system, composed of four main functions (F4.1, F4.2, F4.3 and F4.4). Inputs are sleep transition rules (from the AASM manual), the coarse hypnogram (from F2) and sleep patterns (from F3). The manual scoring $hypno_{ref}$ was used for training. The outputs are the hypnogram $hypno_{EP}$ and the associated probability table $probabilities_{EP}$.

F4.1 To address the timeline, nine features were identified within the coarse hypnogram (five features) and sleep patterns (four features). These features were extracted in regard to the current epoch, as shown in Figure 3.

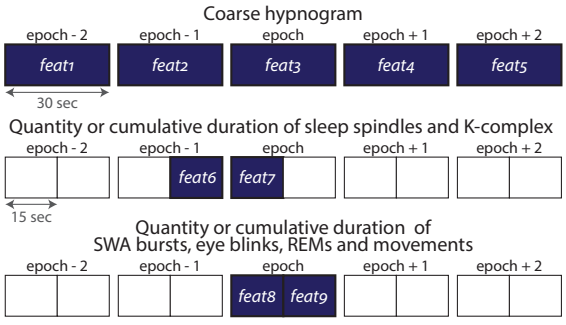

Figure 3: Features extracted from the coarse hypnogram and sleep patterns, in regard to the current epoch.

These features constitute the classifier input vector, noted V. As a consequence, the surrounding epochs are taken into consideration by the classifier. The chosen classifier was a random forest, developed with the same settings as in F2. It was trained concatenating all V vectors from D1 recordings as the input, and all D1 references $hypno_{ref}$ as the reference. The resulting hypnogram will be referred to as *hypno_temp1* in the rest of this paper. The trained model, called ***RF2***, was saved to be reused for each test recording individually, resulting in one *hypno_temp1* per D2 recording.

F4.2 *hypno_temp1* was smoothed using the transition rules described in the AASM guidelines. These rules, that define the possible or forbidden transitions between sleep stages, were implemented using the AASM guidelines but also empirically, by studying the errors that occurred the most. The smoothed hypnogram will be referred to as *hypno_temp2* in the rest of this paper.

F4.3 *hypno_temp2* was further smoothed using a Viterbi hidden Markov model [34], trained to identify and correct sequence mistakes [35, 36]. To do so, we used the hmmviterbi Matlab® function, which required the computation of two matrices:

- the emission probability matrix ***EMIS***, which corresponds to the probability of each sleep stage being emitted depending on the reference sleep stage. It was estimated using the confusion matrix between *hypno_temp2* and $hypno_{ref}$;
- the transition probability matrix ***TRANS***, which corresponds to the probability of transition between each sleep stage. It was estimated from $hypno_{ref}$.

Using these matrices, the Viterbi hidden Markov model smooths hypnogram *hypno_temp2*, resulting in the final one $hypno_{EP}$. ***EMIS*** and ***TRANS*** were both estimated from D1 and then saved to be reused for each test recording individually, resulting in one $hypno_{EP}$ per D2 recording.

F4.4 Independently from F4.1, F4.2 and F4.3, a table called $probabilities_{EP}$ was also computed. This table contains, for each epoch, the estimated probabilities of being in each sleep stage. It can be used to determine which epochs were more or less easily scored by the algorithm. $probabilities_{EP}$ computation is not detailed in this paper. In brief, it was established by measuring the agreement between sets of features which were selected as being representative of each specific sleep stage. $probabilities_{EP}$ thus reflects the algorithm's doubts when scoring sleep. Thanks to this, the medical practitioner knows which epochs the algorithm found difficult or easy to score.

2.3. System evaluation

Results were estimated from the recordings included in the test dataset D2. As a reminder, this dataset is made up of independent recordings not used during training. Each recording was processed using the previously trained models. Then, the resulting $hypno_{EP}$ and $probabilities_{EP}$ were gauged using the associated $hypno_{ref}$ (in 2.3.1 and 2.3.2, respectively). In both sections, reported scores correspond to the averaged individual scores.

2.3.1. Evaluation of $hypno_{EP}$ accuracy

For each recording, $hypno_{EP}$ was compared to $hypno_{ref}$ using a contingency table (see Table 3).

Table 3
Contingency table.

		Automatic analysis					
		W	N1	N2	N3	R	Total
Reference	W	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	$n_{1.}$
	N1	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	$n_{2.}$
	N2	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	$n_{3.}$
	N3	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	$n_{4.}$
	R	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}	$n_{5.}$
	Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{.5}$	n

The overall accuracy of the automatic scoring is assessed using Cohen's Kappa κ [37] and accuracy rate Acc.
κ is probably the most used index for automatic sleep scoring evaluation. Indeed, it measures the agreement between the reference and the automatic analysis by taking into consideration the random component of this agreement (expected agreement on the assumption that the manual and automatic analyses are totally independent). κ is calculated from the proportion of observed agreement P_o and the proportion of random agreement P_e:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

with $P_o = \frac{1}{n}\sum_{i=1}^{5} n_{ii}$ and $P_e = \frac{1}{n^2}\sum_{i=1}^{5} n_{i.} \times n_{.i}$.

It is usually interpreted using six ranges:

(i) $\kappa < 0.0$: Poor agreement
(ii) $0.0 \leq \kappa < 0.2$: Slight agreement
(iii) $0.2 \leq \kappa < 0.4$: Fair agreement
(iv) $0.4 \leq \kappa < 0.6$: Moderate agreement
(v) $0.6 \leq \kappa < 0.8$: Substantial agreement
(vi) $0.8 \leq \kappa$: Almost perfect agreement

Acc corresponds to the percentage of correctly scored epochs:

$$Acc(\%) = \frac{1}{n}\sum_{i=1}^{5} n_{ii} \times 100$$

Overall scores were also reported while considering subjects by age or OSA severity.

Furthermore, scores were estimated for each sleep stage individually. To do so, we adopted a one-vs.-rest approach where each stage is alternatively considered as the positive class and the others are combined into a single negative class. From the resulting true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), we estimated each stage's Cohen's Kappa, accuracy rate, sensitivity (defined as $\frac{TP}{TP+FN}$) and specificity (defined as $\frac{TN}{TN+FP}$).

2.3.2. Decision support with $probabilities_{EP}$

Since the probability table $probabilities_{EP}$ has no reference with which to be compared, we evaluated the mean probability given by $probabilities_{EP}$ when epochs are correctly versus erroneously scored in a specific sleep stage. The greater the difference between those mean probabilities, the better the probability table. Indeed, we want the mistakes in the algorithm to be limited to epochs where the features designated multiple sleep stages rather than a single one (suggesting manual scoring may also have been complex).

3. Results

3.1. Evaluation of $hypno_{EP}$ accuracy

Among the 100 recordings, 84 % obtained an overall Cohen's Kappa κ above 0.60, showing a substantial or almost perfect agreement with the manual scoring.
Table 4 gives, for D2 recordings, the mean value and the standard deviation of the overall Cohen's Kappa κ and accuracy rate Acc of $hypno_{EP}$. Scores per sleep stage are also reported. The overall κ and Acc were 0.69 and 77.8 %, respectively. If we consider each sleep stage detection individually, stage R obtained the best scores with a κ reaching 0.80 (almost perfect agreement with the reference). κ mean values indicate all other sleep stages got a substantial agreement with the manual scoring, except stage N1. Sensitivities were all above 82 %, except for N2 and N1 sleep stages (74.9 % and 20.8 %, respectively). Specificities were all above 94 %, except for N2 sleep stage (83.3 %).

Table 5 and Table 6 report the performances obtained depending on the subject's age (using quartiles) and OSA severity (based on a physician's diagnosis).
There was a substantial agreement with the manual scoring for all groups. For age groups, the lowest scores were obtained for 53-63 year old patients, and the better ones for patients above 63 years old. For OSA severities, κ were above 0.70 for patients with mild or moderate OSA, and below for patients with no or severe OSA.

Table 4
Overall and individual performances obtained from automatic sleep staging on D2 dataset.

D2 dataset	W	N1	N2	N3	R	All
Cohen's Kappa	0.74 ± 0.14	0.23 ± 0.11	0.64 ± 0.14	0.71 ± 0.20	0.80 ± 0.14	0.69 ± 0.10
Accuracy rate (%)	92.3 ± 4.5	92.1 ± 3.4	83.3 ± 6.2	92.7 ± 4.5	95.1 ± 3.0	77.8 ± 7.0
Sensitivity (%)	82.3 ± 15.5	20.8 ± 9.0	83.9 ± 9.5	74.9 ± 20.3	83.3 ± 16.5	N.A.
Specificity (%)	94.0 ± 5.7	97.8 ± 1.1	83.3 ± 8.0	96.5 ± 4.3	97.4 ± 2.2	N.A.

N.A. = Not Applicable

Table 5
Overall performances depending on the subject's age, using quartiles. Q1 = 19-41 years old, Q2 = 41-53 years old, Q3 = 53-63 years old and Q4 = 63-79 years old.

	Q1	Q2	Q3	Q4
κ	0.69 ± 0.10	0.69 ± 0.10	0.67 ± 0.10	0.70 ± 0.09
Acc	78.2 ± 6.6	77.9 ± 6.5	75.8 ± 7.1	78.9 ± 6.7

Table 6
Overall performances depending on the subject's OSA severity (obtained by the sleep expert).

	No	Mild	Moderate	Severe
κ	0.67 ± 0.12	0.70 ± 0.10	0.73 ± 0.06	0.65 ± 0.10
Acc	76.6 ± 7.7	79.1 ± 6.7	80.4 ± 4.7	74.6 ± 7.7

Figure 4 presents the confusion matrix related to Table 4. The quantity of epochs manually scored in each stage (W: 24,172 - R: 18,194 - N1: 8,095 - N2: 41,095 - N3: 19,422) should be borne in mind while interpreting the confusion matrix. Common mistakes appeared to be N1 and N3 epochs being automatically scored as N2 sleep stage.

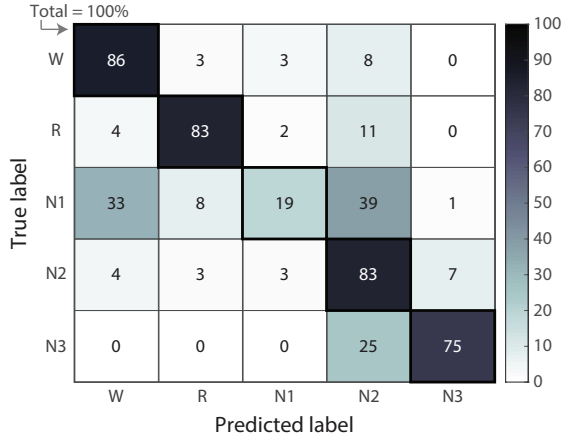

Figure 4: Confusion matrix (percentage over lines) obtained from automatic sleep staging on D2 dataset.

3.2. Decision support with $probabilities_{EP}$

$probabilities_{EP}$ is a supplementary scoring support tool that helps make the system more practical to use and improves physicians' confidence in the algorithm.
Table 7 reports the mean probabilities associated with epochs correctly and erroneously estimated by $hypno_{EP}$, depending on sleep stages. In this table, we can see that epochs correctly identified as N2 have a reported mean probability of 66 % of being N2 stage, whereas the epochs over-detected as N2 have a reported mean probability of only 44 % of being N2. It makes a difference of 22 % between actual N2 epochs and not. Algorithm confidence when scoring an epoch into N2 sleep stage is thus greater when it is an actual N2 epoch. Considering the mean probabilities reported for the other sleep stages, we can see that all differences are above 15 %, except N1 sleep stage.
Erroneous epochs are thus more likely to have small and uniform $probabilities_{EP}$ values than correctly classified ones. Indeed, the latter should show a clear superiority of the probability associated with their sleep stage compared with others. $probabilities_{EP}$ points out epochs which need to be checked as a priority, and manually corrected if necessary.

Table 7
Mean probabilities per sleep stage while the reference agrees or disagrees with the stage detected by the system.

	$hypno_{ref}$ agrees	$hypno_{ref}$ disagrees
$hypno_{EP}$ = W	79 %	60 %
$hypno_{EP}$ = N1	65 %	56 %
$hypno_{EP}$ = N2	66 %	44 %
$hypno_{EP}$ = N3	87 %	71 %
$hypno_{EP}$ = R	69 %	53 %

4. Discussion

The main goal of this study was to implement a user-friendly automatic sleep scoring system.

Several factors have helped to overcome the three limitations presented above:

a) the developed system was designed to be as easy as possible to interpret. To do so, its construction replicates manual scoring and uses medical knowledge extracted from the AASM guidelines. Firstly, the channels chosen were the same ones as when manually scoring sleep. Secondly, we know that medical practitioners quickly visualize the recording before scoring it, to become familiar with its specific characteristics and score accordingly. This step, replicated in the SATUD system (set out in the companion paper [ref]), causes the automatic hypnogram to be patient-dependent. Thirdly, most of the elements described in the AASM guidelines were included in the system: continuous features, sleep patterns, surrounding knowledge and transition rules. Lastly, the methodologies implemented in the system were established using interpretable algorithms or medical knowledge;
b) algorithm performances were evaluated on a hundred independent recordings, from patients with and without sleep-disordered breathing;
c) the system works in real-life conditions and does not require any previous human intervention. There is no need for invalidation of epochs or event scoring. $probabilities_{EP}$ also improves system practicality by pointing out epochs that should be checked as a priority.

Despite these restrictions, the method showed it could get results comparable to those obtained with manual scoring, reaching Cohen's Kappa values around 0.69 and accuracy rates around 78 % (see Table 4). There was no significant impact of age on performance (see Table 5). As for OSA severity (see Table 6), performance was the lowest for patients with severe OSA syndrome. Surprisingly,

Table 8
Performance of 5-stage classification using electrophysiological channels, compared with related works.

	Number of one-night PSG (training and test)	Subject diagnosis	Approach	Overcome limitations*	Acc (%)	κ
Zokaeinikoo et al. 2016 [24]	20 (LOOCV**)	Healthy only	ML	c	74	
Biswal et al. 2018 [16]	10,000 (9,000-1,000)	Healthy and SDB	DL	b and c	**88**	0.81
Zhang et al. 2019 [17]	5,804 (5,213-580)	Healthy and SDB	DL	b and c	87	**0.82**
Chen al. 2019 [26]	16***	Healthy and SDB	Interpretable	a and b	80	0.72
This work	400 (300-100)	Healthy and SDB	Interpretable	a, b and c	78	0.69

ML = Machine Learning, DL = Deep Learning
* Limitations overcome from our point of view.
Limitation a: model opacity, b: insufficient heterogeneity of dataset and c: lack of practicality
** Leave-One-Out Cross-Validation: one by one, each subject's recording was selected as the test dataset, and the others were combined into the training dataset. Final results are provided by the model with the highest scores.
*** Semi-automated method that requires the manual scoring of 5 % of each recording.

the group with the second worst performance was patients with no OSA syndrome. Nonetheless, the results obtained for all groups were acceptable (the lowest κ was 0.65), indicating the algorithm is robust to more or less fragmented sleep recordings. Considering each sleep stage individually (see Figure 4 and Table 4), it seems that the errors were mainly N3 epochs misclassified as N2 sleep stage. Also, sleep stage N1 had very low performance compared to other stages. This is not surprising, since sleep stage N1 is a transitional stage representing approximately 5 % of the night with a very likely overlap with W and N2 stages. In fact, inter-scorer agreement for N1 sleep stage is the lowest [8]. Table 7 showed that misclassified epochs have lower probability values in the returned table $probabilities_{EP}$, compared with correctly scored epochs. Using $probabilities_{EP}$, some of the mistaken epochs could thus be identified and rescored by the manual scorer. A possible strategy would be to highlight epochs with no obvious superiority of one stage probability among the others. The scorer could consider reviewing only those epochs.

Table 8 presents the obtained results, compared with the literature. Only studies using EEG, EOG and EMG, compared with hypnograms manually scored following the AASM guidelines and indicating the number of patients and training/test repartitions were considered for comparison. Deep learning approaches, which reached better scores, do not overcome limitation a). Chen et al. approach [26], which is also interpretable, do not overcome limitation c) since it is semi-automated. Our method showed that despite the inclusion of transparency, hybrid methods can still reach adequate scores.

The proposed system provides resilient tools to facilitate sleep scoring, thus assisting sleep experts in diagnosing sleep disorders. As we are aware of sleep specialists' mistrust in automatic approaches, we identified three limitations to their use and designed the system to overcome them. To go even further, several perspectives are considered. Firstly, we would like to evaluate this methodology using recordings from other sleep laboratories. Indeed, our recordings were all provided by one sleep laboratory and, even if several manual scorers established the references, local scoring practices may have influenced the algorithms. Secondly, micro-arousals[2] (used when manually scoring sleep as they are required by some transition rules), were not detected in this system. Work is currently being done to identify them. Thirdly, the automatic sleep scoring impact on sleep diagnosis should be evaluated. The Apnea Hypopnea Index (AHI) resulting from the current system should be compared with the one resulting from manual scoring. Lastly, since sleep diagnosis is sometimes performed using devices that do not record EP channels, it would be interesting to see how good automatic sleep scoring from cardio-respiratory channels could be.

5. Conclusion

In this paper, a new approach for automatic sleep staging was presented. Its architecture was designed to reproduce step-by-step the manual scoring tasks realized by sleep experts: adaptation to each recording's specific characteristics, study of temporal and spectral content, identification of sleep patterns and classification regarding surrounding epochs and transition rules. As it is easy to understand, this model is well suited to the medical community which lacks confidence in the models generally implemented.
The method was evaluated on 100 patients with and without sleep-disordered breathing, and results showed the automatic hypnogram made a good performance regardless of subjects' age or OSA severity. With a mean Cohen's Kappa and accuracy rate of 0.69 and 77.8 %, respectively, the algorithm obtained a high agreement with the manual scorer.
The proposed method was put together as a scoring support tool for sleep scoring and is thus useable immediately after the polysomnographic recording, without the need for any preliminary human action. Besides the automatic hypnogram, it also points out epochs which should be checked and rescored if necessary.
Given that sleep scoring is a time-consuming and complex task, the presented user-friendly tool should greatly support sleep specialists in their diagnosis.

[2] Short awakenings, markers of sleep disruption

Funding

This study was supported by grants from the Institut de Recherche en Santé Respiratoire des Pays de La Loire.

Acknowledgements

The authors would like to thank Christelle Gosselin and Jean-Louis Racineux, from the Institut de Recherche en Santé Respiratoire des Pays de La Loire, and Margaux Blanchard, from the ESEO. Thanks to Alain Le Duff and Lucile Riaboff, previously from the ESEO. We thank Julien Godey, Laetitia Moreno and Marion Vincent, sleep technicians in the Department of Respiratory and Sleep Medicine of Angers University Hospital.

References

[1] R. Heinzer, S. Vat, P. Marques-Vidal, H. Marti-Soler, D. Andries, N. Tobback, V. Mooser, M. Preisig, A. Malhotra, G. Waeber, P. Vollenweider, M. Tafti, J. Haba-Rubio, Prevalence of sleep-disordered breathing in the general population: the HypnoLaus study, The Lancet Respiratory Medicine 3 (2015) 310–318. doi:10.1016/S2213-2600(15)00043-0.

[2] J. B. Croft, CDC's Public Health Surveillance of Sleep Health, 2017.

[3] C. V. Senaratna, J. L. Perret, C. J. Lodge, A. J. Lowe, B. E. Campbell, M. C. Matheson, G. S. Hamilton, S. C. Dharmage, Prevalence of obstructive sleep apnea in the general population: A systematic review, Sleep Medicine Reviews 34 (2017) 70–81. doi:10.1016/j.smrv.2016.07.002.

[4] AASMTaskForce, Sleep-related Breathing Disorders in Adults: Recommendations for Syndrome Definition and Measurement Techniques in Clinical Research, Sleep 22 (1999) 667–689. doi:10.1093/sleep/22.5.667.

[5] K. A. Franklin, E. Lindberg, Obstructive sleep apnea is a common disorder in the population- a review on the epidemiology of sleep apnea, Journal of Thoracic Disease 7 (2015) 1311–1322. doi:10.3978/j.issn.2072-1439.2015.06.11.

[6] P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, K. M. Hla, Increased Prevalence of Sleep-Disordered Breathing in Adults, American Journal of Epidemiology 177 (2013) 1006–1014. doi:10.1093/aje/kws342.

[7] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, R. M. lloyd, S. F. Quan, M. M. Troester, B. V. Vaughn, The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, number 2.4 in American Academy of Sleep Medicine, Darien IL, 2017.

[8] R. S. Rosenberg, S. Van Hout, The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring, Journal of Clinical Sleep Medicine (2013). doi:10.5664/jcsm.2350.

[9] A. L. Fogel, J. C. Kvedar, Artificial intelligence powers digital medicine, npj Digital Medicine 1 (2018). doi:10.1038/s41746-017-0012-2.

[10] E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature Medicine 25 (2019) 44–56. doi:10.1038/s41591-018-0300-7.

[11] T. Penzel, R. Conradt, Computer based sleep recording and analysis, Sleep Medicine Reviews 4 (2000) 131–148. doi:10.1053/smrv.1999.0087.

[12] L. Fiorillo, A. Puiatti, M. Papandrea, P.-L. Ratti, P. Favaro, C. Roth, P. Bargiotas, C. L. Bassetti, F. D. Faraci, Automated sleep scoring: A review of the latest approaches, Sleep Medicine Reviews 48 (2019). doi:10.1016/j.smrv.2019.07.007.

[13] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444. doi:10.1038/nature14539.

[14] C. M. Bishop, Pattern Recognition and Machine Learning, 2006.

[15] S. Wermter, R. Sun, An Overview of Hybrid Neural Systems, in: G. Goos, J. Hartmanis, J. van Leeuwen, S. Wermter, R. Sun (Eds.), Hybrid Neural Systems, volume 1778, Springer Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 1–13. doi:10.1007/10719871_1, series Title: Lecture Notes in Computer Science.

[16] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, M. T. Bianchi, Expert-level sleep scoring with deep neural networks, Journal of the American Medical Informatics Association 25 (2018) 1643–1650. doi:10.1093/jamia/ocy131.

[17] L. Zhang, D. Fabbri, R. Upender, D. Kent, Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks, Sleep 42 (2019). doi:10.1093/sleep/zsz159.

[18] S. Enshaeifar, S. Kouchaki, C. C. Took, S. Sanei, Quaternion Singular Spectrum Analysis of Electroencephalogram With Application in Sleep Analysis, IEEE Transactions on Neural Systems and Rehabilitation Engineering 24 (2016) 57–67. doi:10.1109/TNSRE.2015.2465177.

[19] T. Lajnef, S. Chaibi, P. Ruby, P.-E. Aguera, J.-B. Eichenlaub, M. Samet, A. Kachouri, K. Jerbi, Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines, Journal of Neuroscience Methods 250 (2015) 94–105. doi:10.1016/j.jneumeth.2015.01.022.

[20] S. Mahvash Mohammadi, S. Kouchaki, M. Ghavami, S. Sanei, Improving time-frequency domain sleep EEG classification via singular spectrum analysis, Journal of Neuroscience Methods 273 (2016) 96–106. doi:10.1016/j.jneumeth.2016.08.008.

[21] S. Charbonnier, L. Zoubek, S. Lesecq, F. Chapotot, Self-evaluated automatic classifier as a decision-support tool for sleep/wake staging, Computers in Biology and Medicine 41 (2011) 380–389. doi:10.1016/j.compbiomed.2011.04.001.

[22] G. Garcia-Molina, F. Abtahi, M. Lagares-Lemos, Automated NREM sleep staging using the Electro-oculogram: A pilot study, in: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, IEEE, 2012, pp. 2255–2258. URL: http://ieeexplore.ieee.org/abstract/document/6346411/.

[23] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, H. Dickhaus, Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier, Computer Methods and Programs in Biomedicine 108 (2012) 10–19. doi:10.1016/j.cmpb.2011.11.005.

[24] M. Zokaeinikoo, Automatic Sleep Stages Classification, Ph.D. thesis, 2016. URL: http://trace.tennessee.edu/utk_gradthes/4088/.

[25] A. Ugon, Fusion Symbolique et Données Polysomnographiques, Ph.D. thesis, 2015.

[26] C. Chen, A. Ugon, C. Sun, W. Chen, C. Philippe, A. Pinna, Towards a Hybrid Expert System Based on Sleep Event's Threshold Dependencies for Automated Personalized Sleep Staging by Combining Symbolic Fusion and Differential Evolution Algorithm, IEEE Access 7 (2019) 1775–1792. doi:10.1109/ACCESS.2018.2887082.

[27] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, J. Faubert, Deep learning-based electroencephalography analysis: a systematic review, Journal of Neural Engineering 16 (2019). doi:10.1088/1741-2552/ab260c.

[28] F. Doshi-Velez, B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, arXiv:1702.08608 [cs, stat] (2017). ArXiv: 1702.08608.

[29] M. Glos, A. Sabil, K. S. Jelavic, C. Schöbel, I. Fietze, T. Penzel, Characterization of Respiratory Events in Obstructive Sleep Apnea Using Suprasternal Pressure Monitoring, Journal of Clinical Sleep Medicine 14 (2018) 359–369. doi:10.5664/jcsm.6978.

[30] T. Penzel, A. Sabil, The use of tracheal sounds for the diagnosis of sleep apnoea, Breathe 13 (2017) e37–e45. doi:10.1183/20734735.008817.

[31] T. Penzel, A. Sabil, Physics and Applications for Tracheal Sound Recordings in Sleep Disorders, in: K. N. Priftis, L. J. Hadjileontiadis, M. L. Everard (Eds.), Breath Sounds, Springer International Publishing, Cham, 2018, pp. 83–104. doi:10.1007/978-3-319-71824-8_6.

[32] P. Hauri, D. R. Hawkins, Alpha-delta sleep, Electroencephalography and Clinical Neurophysiology 34 (1973) 233–237. doi:10.1016/

0013-4694(73)90250-2.

[33] T. K. Ho, Random Decision Forests (1995). doi:10.1109/ICDAR.1995.598994.

[34] L. E. Baum, T. Pietrie, Statistical Inference for Probabilistic Functions of Finite State Markov Chains, The Annals of Mathematical Statistics (1966) 1554–1563. doi:10.1214/aoms/1177699147.

[35] J. Yang, Toward physical activity diary: motion recognition using simple acceleration features with mobile phones, in: Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics - IMCE '09, ACM Press, Beijing, China, 2009, p. 1. doi:10.1145/1631040.1631042.

[36] L. Riaboff, S. Poggi, A. Madouasse, S. Couvreur, S. Aubin, N. Bédère, E. Goumand, A. Chauvin, G. Plantier, Development of a methodological framework for a robust prediction of the main behaviours of dairy cows using a combination of machine learning algorithms on accelerometer data, Computers and Electronics in Agriculture 169 (2020). doi:10.1016/j.compag.2019.105179.

[37] J. Cohen, A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement 20 (1960) 37–46. doi:10.1177/001316446002000104.