

Investigation of an Automatic Sleep Stage Classification by Means of Multiscorer Hypnogram

V. C. Figueroa Helland¹; A. Gapelyuk^{2, 3}; A. Suhrbier³; M. Riedl²; T. Penzel⁴; J. Kurths^{2, 5}; N. Wessel^{2, 3}

¹Interdisciplinary Center for Dynamics of Complex Systems, University of Potsdam, Potsdam, Germany;

²Department of Physics, Humboldt-Universität zu Berlin, Berlin, Germany;

³Charité – Universitätsmedizin Berlin, Campus Berlin-Buch, Experimental and Clinical Research Center, Berlin, Germany;

⁴Department of Sleep Medicine, Charité – Universitätsmedizin Berlin, Campus Mitte, Berlin, Germany;

⁵Potsdam Institute for Climate Impact Research, Potsdam, Germany

Keywords

Sleep staging, polysomnogram, linear discriminant analysis, automatic classification

Summary

Objectives: Scoring sleep visually based on polysomnography is an important but time-consuming element of sleep medicine. Whereas computer software assists human experts in the assignment of sleep stages to polysomnogram epochs, their performance is usually insufficient. This study evaluates the possibility to fully automatize sleep staging considering the reliability of the sleep stages available from human expert sleep scorers.

Methods: We obtain features from EEG, ECG and respiratory signals of polysomnograms from ten healthy subjects. Using the sleep stages provided by three human experts, we evaluate the performance of linear discriminant analysis on the entire polysomnogram

and only on epochs where the three experts agree in their sleep stage scoring.

Results: We show that in polysomnogram intervals, to which all three scorers assign the same sleep stage, our algorithm achieves 90% accuracy. This high rate of agreement with the human experts is accomplished with only a small set of three frequency features from the EEG. We increase the performance to 93% by including ECG and respiration features. In contrast, on intervals of ambiguous sleep stage, the sleep stage classification obtained from our algorithm, agrees with the human consensus scorer in approximately 61%.

Conclusions: These findings suggest that machine classification is highly consistent with human sleep staging and that error in the algorithm's assignments is rather a problem of lack of well-defined criteria for human experts to judge certain polysomnogram epochs than an insufficiency of computational procedures.

the pattern found in consecutive 30-second-long epochs of the electroencephalography (EEG), electro-oculography (EOG), and electromyography (EMG) recordings [1, 2]. The resulting succession of discrete sleep stages is referred to as hypnogram and supports diagnostic decisions.

While automatic sleep stage classification is taken as the starting point for sleep stage scoring, its performance is usually insufficient so that the scoring ultimately requires visual inspection of the polysomnograms by expert human scorers. Visual examination of polysomnogram epochs constitutes not only a time-consuming procedure but further, the resulting hypnograms are strongly dependent on the particular human expert performing the analysis. It is known that there is a significant inter-scorer variability (about 70% agreements) [3]. Full automatization of sleep scoring would both increase the time efficiency and improve the reproducibility in the generating of hypnograms.

Correspondence to:

Niels Wessel
Department of Physics
Humboldt-Universität zu Berlin
Robert-Koch-Platz 4
10115 Berlin
Germany
E-mail: niels.wessel@charite.de

Methods Inf Med 2010; 49: 467–472

doi: 10.3414/ME09-02-0052

received: November 20, 2009

accepted: February 23, 2010

prepublished: July 20, 2010

1. Introduction

Sleep is an active and regulated process with an essential restorative function for physical and mental health. Sleep disorders can result not only in impairments in life quality but also in physiological dysfunc-

tions. Part of the diagnostic process is a quantitative sleep recording using polysomnography. Polysomnography is evaluated by a visual scoring of sleep stages. Sleep scoring is a fundamental aspect of sleep research and sleep medicine, and involves assigning one of six sleep stages to

2. Objectives

In this study we address the question as to why computerized sleep scoring [4, 5] has failed so far to produce sleep stage assignments that satisfy human experts. In principle the observed insufficiency of current computer-generated sleep stage scores may arise from a) an inability of the applied algorithms to reproduce the human scorer's sleep stages based on polysomnogram graphs or b) from inconsistency in the human polysomnogram scoring itself. To discriminate between the two possibilities

we apply classic linear discriminant analysis with stepwise feature selection to polysomnographic records from ten healthy subjects that have previously been scored by three human experts. Using this data set we can assess to what extent the machine is able to capture human assignments in the absence or presence of disagreement between human experts and in turn whether the assumed insufficiencies of machine learning algorithms are due to possibly oversimplified algorithms or rather are due to a lack of objective criteria based on which sleep stages can be assigned.

3. Methods

3.1 Polysomnogram Data

The data utilized in this study was retrieved from the Siesta database and contains polysomnograms of ten healthy subjects suitable for multiscoring effects [6]. The polysomnograms contain six EEG channels with references (FP1-M2, FP2-M1, O1-M2, O2-M1, C4-M1, C3-M2, M1, and M2), one electrocardiogram channel (ECG modified II lead), a pulse measurement, two types of EMG (one m. submentalialis and one m. tibialis), nasal air flow, oxygen saturation and pulse rate, and respiratory movements of the thorax and of the abdomen. The corresponding hypnograms were generated by three experienced scorers. Two of the scorers assessed the data independently and the third one reached a consensus between the two assignments. In the following we refer to the third expert as the consensus scorer. All the scorers assigned sleep stages according to the guidelines of Rechtschaffen and Kales [1], with the stages: Wake, Stage 1 (S1), Stage 2 (S2), Stage 3 (S3), Stage 4 (S4), REM, and movement time. The sleep stages were thereby assigned for non-overlapping 30-second epochs.

3.2 Human Expert Scoring and Inter-scorer Variability Reduction

As pointed out in the introduction, the hypnograms obtained by two scoring experts do not always coincide. Such incon-

sistencies occur particularly often in sleep periods with rapid transitions between sleep stages and are possibly a major source for classification errors made by any classification algorithm. Therefore, to obtain sleep stage assignments with quantifiable consistency from the scores of the human experts we shift 3 min intervals in steps of 30 s epochs over the polysomnogram resulting in some 900 intervals per subject (corresponding to 7.5 hours of sleep recordings). Then, the following data sets, containing intervals with scores of increasing consistency, are generated:

- a) All 3 min intervals (8264) with sleep stage as assigned by the consensus scorer. The sleep stage of the covered 30 s epochs are thereby combined in such a way that if there is a particular sleep stage occurring in more than 50% of the 3 min window, we ascribe that sleep stage as joint assignment to the interval.
- b) Three min intervals for which the joint assignment is consistent for the three human scorers (scorers agree on what stage prevails in the interval).
- c) Three min intervals for which the consensus scorer assigned the same sleep stage to all six 30 s epochs (consensus scorer finds no sleep stage transitions).
- d) Three min intervals for which the three scorers all assign the same sleep stage to all six 30 s epochs (no scorer finds transitions and all scorers agree on every epoch)

Data sets (b) and (c) can be interpreted in terms of alternative filtering operations in that the joint assignment procedure in (b) smoothes the sleep stages of each scorer over time, whereas in (c) the consensus smoothes the stages across the scorers.

3.3 Polysomnogram Feature Extraction

We first compute several features for each 3 min interval of the polysomnogram. The EEG features are obtained from the C4-M1 channel after removing outlier values (beyond four standard deviations) followed by standardization (with respect to the full night signal mean and variance).

Then we compute the power within the following frequency bands: *delta* (0.5–4 Hz), *theta* (4–8 Hz), *alpha* (8–12 Hz), and *beta* (12–30 Hz). The measure *P* refers to the power of these bands within the 3 min epoch. Subsequently, we compute the following quantities: *beta/delta*, *alpha/delta*, *theta/delta*, *beta/theta*, *alpha/theta*, *beta/alpha*, *beta/P*, *alpha/P*, *theta/P*, and *delta/P*. This selection of frequency band quantification we take from the sleep stage definitions given in the sleep scoring manual [1]. As in previous studies taking into account properties from cardiorespiratory signals for computer classification [7, 8] we also extend our feature pool beyond information from the EEG. From the respiratory and electromyogram signals we compute the mean, standard deviation, median, and root mean square standard deviation. Finally, for the quantification of the ECG signal, we compute several heart rate variability (HRV) parameters [9]: mean heart rate, the standard deviation (sdNN) as well as the coefficient of variation (cvNN = sdNN over mean beat-to-beat-interval). A detailed description and classification of dynamic changes using standard HRV measures is often not sufficient, therefore, we are also applying nonlinear methods based on symbolic dynamics: POLVAR20, the probability of low variability below 20 ms quantifies laminar phases in the time series which have proven to predict life-threatening arrhythmias as early as 10 min before the event [10, 11]. FWRENYI4, the Renyi-entropy of order 4 of the word distribution – quantifying the global heart rate dynamics which has been shown to identify risk patients after myocardial infarction [12, 13]. Finally as the third symbolic dynamics parameter we calculated WSDVAR: the word variability – quantifying the dynamical changes and which has been proven very helpful to quantify complexity in physiological time series. Using WSDVAR were top scorers in the Computers in Cardiology 2002 challenge [14] and applied it also successfully to animal models [15, 16]. For details we refer to our previous papers [10, 17, 18].

The above measures result in 74 feature values for each 3 min interval of the polysomnogram.

Table 1

Percentage of sleep stage intervals retained in each dataset

Dataset	Wake	S1	S2	S3 & S4	REM	Total
(a)	100	100	100	100	100	100
(b)	82	49	83	82	86	82
(c)	71	33	84	84	84	81
(d)	61	13	67	70	68	65

Table 2

Overlap of datasets (b) and (c)

Number of intervals			
	(b)	(c)	Overlap
Wake	527	456	417
S1	180	120	79
S2	3586	3621	3242
S3 & S4	1126	1164	1050
REM	1355	1314	1176
Total	6774	6675	5964

3.4 Stepwise Linear Discriminant Analysis and Cross-validation

Linear discriminant analysis provides a framework to classify an object based on the feature values which describe it. By means of linear discriminant functions it is possible to combine these features in such a way that objects from different classes are as distinguishable as possible [19]. The features from which the discriminant functions are constructed must be chosen based on their suitability for performing accurate classification. As a method for fast feature selection we apply forward stepwise selection with Wilk's lambda as optimality criterion. Then we evaluate the performance of this classification method beyond the training examples: we generate one linear discriminant model by stepwise feature selection on 2/3 of the 3 min intervals (training set) and then use the model to classify the remaining 1/3 of the intervals (validation set). This choice for the size of randomly chosen training and validation sets is based on a trade-off between having a training set as large as possible, while leaving a validation set large enough for a solid estimate of classification performance measures. The training-validation procedure is repeated 100 times, yielding 100 linear discriminant models. This collection of models can be used to assess the reliability of the selected features as successful candidates for sleep stage classification.

We accounted for the number of times (among the 100) a feature was selected as a good sleep stage predictor. We assess the performance of each linear discriminant model on the validation set by computing accuracy, sensitivity (true positive rate) and specificity (true negative rate).

4. Results

4.1 Extent of Scorer Disagreement over Epoch Classification

Classification procedures (human or machine-performed) require the notion of "correct category of the object" in order to allocate objects to a class. To derive a machine rule that classifies sleep stages, the source of correct sleep stages is human expert scorers. Therefore, we analyze to what extent the notion of correct classification exists in practice, namely, to what degree the scorers agree in their assignment of sleep stages to polysomnogram intervals. To this end, we define four sets of intervals with increasing stringency regarding the consistency of sleep stage class assignments of human experts (see Methods). Except from sleep stage S1, we find that between 82% and 86% of intervals have the same joint assignment (dataset (b), ▶ Table 1). This indicates that intervals containing sleep stage S1 are subject to greater contro-

Table 3 Feature selection. The table shows the list of relevant features (according to stepwise feature selection, see Methods). The rightmost column shows the frequency with which the features got selected as good sleep stage predictors (among 100 training-validation executions).

Selection of variables according to % subsets for which Wilk's lambda is optimized		
EEG	delta	100
	relative delta	99
	signal power	78
	relative theta	78
	theta	78
	beta/alpha	38
	beta/theta	30
	theta/delta	29
ECG	beta/delta	36
	cvNN	49
	POLVAR20	60
	sdNN	24
	FWRENYI4	26
Respiration	WSDVAR	30
	meanNAF	21

versy among human experts than those containing other sleep stages.

Similar results are obtained for data set (c) where we include only those intervals where the consensus scorer found no transitions between sleep stages (▶ Table 1 row c). Such regions likely contain the most accurate sleep stage assignments. Finally, data set (d) consists only of those intervals for which the three scorers all assign the same sleep stage to all the covered epochs. We find that only 65% of intervals fall into this category of highest reliability. This suggests that one third of the data is cause for ambiguities in sleep stage assignment.

4.2 Performance of Linear Discriminant Models with Different Number of Features on Consistent Data

Supervised learning classification algorithms require data with predefined cat-

egory, although other machine-learning techniques could be applied otherwise [20]. Therefore, here we first supply the algorithm with only those polysomnogram intervals for which the three human experts assigned the same stage to all contained epochs (data set (d)). The training and validation sets generated from this data are optimally suited as they contain only the intervals with the least ambiguous sleep stage. A number of classification algorithms are available [21], here we choose linear discriminant functions based on the reliability of this long tested framework rather than the sophistication of newer ones, as we are interested in evaluating the principle behind the possibility of fully automatic sleep stage classification.

We take the features which, among 100 training-validation iterations, are most often selected as good predictors (see Methods and ▶Table 3) and specify the number of features by observing if the results are satisfactory for the five sleep stages. For comparison we point to the improvement of taking 3 or 15 features: the

overall performance (agreement with human consensus scorer) increases from 90% to 93%, the true positive classifications increase, for stage Wake from 66% to 83%, for S2 from 93% to 97%, for deep sleep from 93% to 94%, and for REM from 84% to 90% (▶Fig. 1). We observe the general tendency that the classification accuracy improves with more features. However, an analysis of true positive classifications per sleep stage shows that the inclusion of more features benefits the classification of some sleep stages but deteriorates it for others. We find that for stages Wake, S2, and deep sleep (S3 and S4) performance is optimal at around 9, 6, 2 features respectively. For stage REM there is a drop in performance already at two features (although the maximum occurs with 18 features).

True negative classifications are high (for all stages above 90%), with the lowest value occurring for S2. This result holds even using very few features (observe right panel in ▶Fig. 1).

4.3 Quality of Machine Classification Depending on Consistency of Human Expert Scorers

Having established that linear discriminant models can reproduce over 90% of sleep stages uniformly assigned by all three human expert scorers we now assess how well our automatic classification agrees with the human consensus scorer on the datasets including intervals on which the three human scorers disagreed with each other (intervals of less certain sleep stage).

When applying the linear discriminant classification to only those 2892 intervals with least reliable sleep stage (dataset (a) – dataset (d)) we obtain a performance of only $\approx 61\%$. For dataset (a) (which includes all 8264 intervals) the sleep stages assigned by the algorithm agree with those of the consensus scorer in 80% of cases ($0.90 \times 0.65 + 0.61 \times 0.35 = 0.80$). This represents a drop in performance of about 10% compared to what has been achieved on the intervals with most reliable sleep stage (see datasets (a) and (d) in ▶Fig. 1).

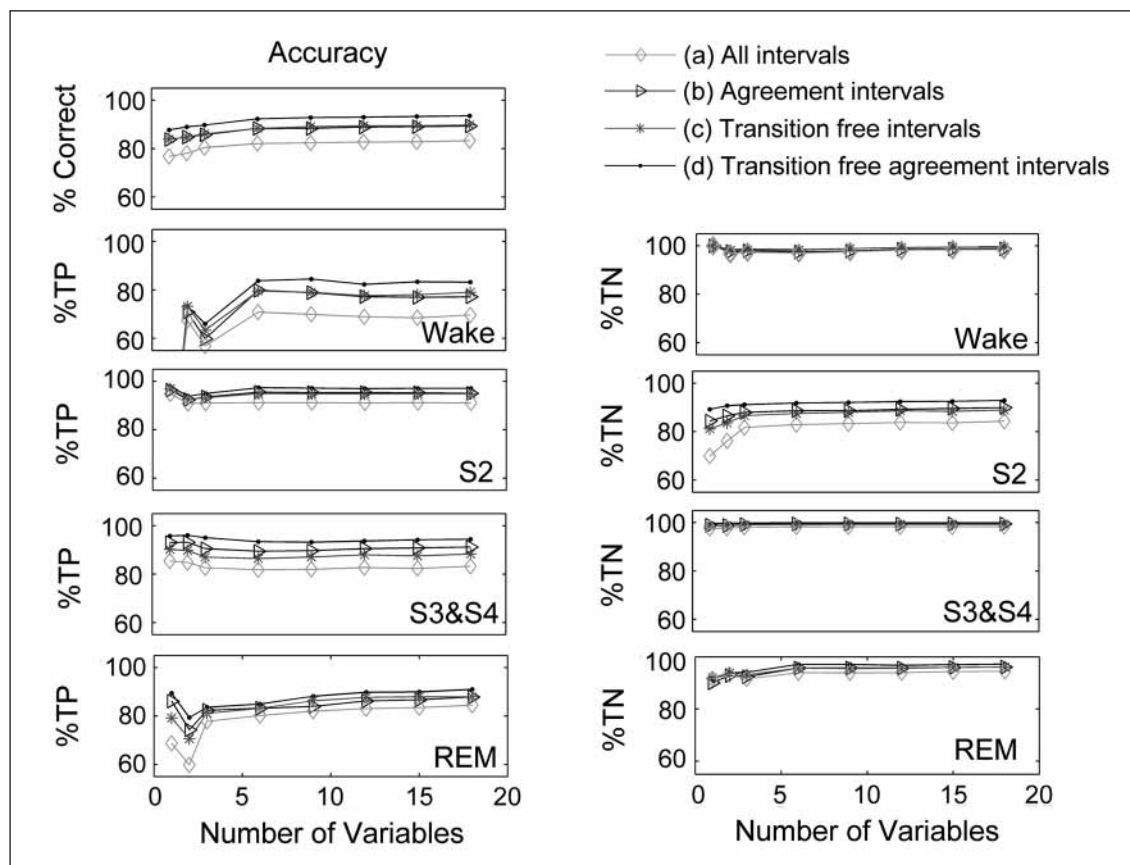


Fig. 1 Performance of linear discriminant models on the four consistency datasets. Each curve corresponds to the agreement of the linear discriminant model with the human consensus scorer on the indicated dataset. Left panels show the percentage of true positives (%TP) for increasing number of features included in the linear discriminant model. Right panels show the corresponding percent of true negatives (%TN).

Interestingly, dataset (b) (where scorers agree on what stage is prevalent in the interval) and dataset (c) (where the consensus finds no transitions) overlap to a large extent (approximately 90%, ▶Table 2). This indicates that if the scorers agree in their joint assignment, then the consensus likely scored no transitions (and vice versa), confirming that disagreements occur mostly in determining transitions. Concordantly, the performance of the linear discriminant analysis on dataset (c) resembles that on dataset (b) (▶Fig. 1).

The largest drop in true positive classifications ($\approx 14\%$) is observed in Wake indicating that the classification of this stage is most strongly contaminated by controversial scoring (among stages REM, S2 and deep sleep). The classification of deep sleep also improves considerably by removing intervals of ambiguous sleep stage (observe the drop of $\approx 11\%$ in the true positive classifications).

5. Discussion

We have shown that linear discriminant analysis with only three features (EEG low frequency) can reproduce over 90% of sleep stages when these are unambiguously assigned by three human experts. Ignoring the problem of ambiguities in sleep staging, the automatic classification agrees with the human consensus scorer in 80% of the intervals (83% using nine features). As pointed out in the Results section, the true positive classification curves have different optima for different sleep stages, and using 15 features may come at the cost of slightly overfitting some sleep stages. This result is nevertheless higher than the 70% agreement between human experts as reported in other studies [22, 23]. On this matter we point out that our analysis, because it is based on 3 min intervals using scores available at higher resolution (of 30 s), makes it possible to distinguish heterogeneous dynamics within the interval. Through the joint sleep stage assignment procedure such local information is combined to the effect of smoothing sleep stage variability in the longer interval. Human experts also perform such smoothing in visually scoring the data by assigning the sleep stage that

dominates in the 30 s epoch, whenever they observe characteristics of several sleep stages.

We have performed our analysis on polysomnograms from healthy subjects, however, the scoring of data from individuals suffering from sleep disorders poses a greater challenge to both human scoring experts and to computerized sleep staging procedures. In such setting the application of our classification method is expected to be less accurate and the study of multi-scoring effects is likely to be of even greater relevance.

The removal of sleep stage transitions is an oversimplification of sleep dynamics. These intervals represent a problem both in the context of classification algorithms and in the context of humans performing the scoring. Even if the transitions would be unequivocally defined by the human experts, an interval containing transitions is only partially classifiable (it is not a stage, but rather a mixture of stages). We find that approximately 20% of the intervals contain sleep stage transitions according to the consensus scorer.

The aim of having unambiguous definitions of sleep stages is to separate polysomnogram segments accordingly. Therefore, unclear sleep scoring criteria result in difficulty for experts to agree about successive data epochs as being “different” stages (and if so, which stages), that is, in their scoring of transitions. This lack of definiteness is evidenced by the fraction of intervals containing transitions (▶Table 1) and the overlap between datasets (b) and (c) (▶Table 2). Datasets (b) and (c) largely coincide, so an interval from (c), namely, one with no transitions (according to the consensus) likely results in an unanimous joint assignment. Therefore, disagreements between experts occur mostly where at least one scorer (here, the consensus) finds transitions (and for S1 disagreement occurs even despite lack of transitions). For these intervals there is no sleep stage that can serve as a rule for training our classifier.

Sources of controversy in sleep staging arising from the definition of sleep stage leave some space for individual interpretation [22]. Hence, in order to achieve a higher agreement rate, a better refined sleep stage definition is needed. Actually

this was one of the reasons for the recent revision of the guidelines for sleep stages [2].

We have evaluated the EEG with features that assume stationarity. Applying our analysis to 3 min intervals with transitions introduces error. Using 30 s windows does not remove the error introduced by classifying transition segments because the experts have provided no information of the varying dynamics within intervals shorter than 30 s. With longer intervals we can segregate in dataset (d) the intervals for which sleep stage stationarity best holds in the sense that all scorers agree that the 3 min interval consists of a steady sleep stage. In intervals with transitions, separating error due to stationarity-based methods on instationary data vs. error due to scorer disagreement requires more careful analysis.

More sophisticated techniques for polysomnogram quantification, classification and feature selection are candidates for improving the present classification. We have not taken into account the possibly complex interaction between features. An outlook on improvements includes adaptive quantification of the EEG (for coping with transition segments), branch and bound feature selection and support vector machines or fuzzy logic as alternative automatic classifiers.

6. Conclusions

We conclude that a classification algorithm based on linear discriminant analysis can to a very large extent reproduce the judgment of sleep scoring human experts. Only three frequency features from the EEG suffice to accomplish an accuracy of 90% if the intervals are such that no disagreement arises between experts and are all free from sleep state transitions. On such epochs we can increase the accuracy to 93% by including features from the ECG and respiratory signal parameters, mainly to the advantage of improving the classification of Wake and REM. Removing sources of sleep stage ambiguity improves classification considerably: 10% overall. In contrast, on intervals of ambiguous sleep stage, the agreement between the automatic classification and

the human expert is approximately only 61%. The problem of ambiguous scoring affects the classification stages Wake and deep sleep more than S2 and REM. With these findings we conclude that fully automatic sleep staging is achievable through resolving ambiguities in the assignment of sleep stages.

Acknowledgments

We acknowledge financial support by the Deutsche Forschungsgemeinschaft Grants No. KU-837/20-2, No. KU-837/23-1 and PE 628/4-1 as well as by the EU Network of Excellence BioSim, contract no. LSHB-CT-2004-005137.

References

1. Rechtschaffen A, Kales A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Washington DC: US Government Printing Office, US Public Health Service; 1968.
2. Iber C, Ancoli-Israel S, Chesson A, Quan SF. The AASM Manual for the scoring of sleep and associated events: Rules, terminology and technical specifications. American Academy of Sleep Medicine; 2007.
3. Danker-Hopfe H, Anderer P, Zeitlhofer J, Boeck M, Dorn H, Gruber G, et al. Interrater reliability (IRR) for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J Sleep Res* 2009; 18 (1): 74–84.
4. Anderer P, Gruber G, Parapatics S, Wörtz M, Miazhyńska T, Klosch G, et al. An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24×7 utilizing the Siesta database. *Neuropsychobiology* 2005; 51 (3): 115–133.
5. Caffarel J, Gibson GJ, Harrison JP, Griffiths CJ, Drinnan MJ. Comparison of manual sleep staging with automated neural network-based analysis in clinical practice. *Med Biol Eng Comput* 2006; 44 (1–2): 105–110.
6. Klösch G, Kemp B, Penzel T, Schlögl A, Rappelsberger P, Trenker E, et al. The SIESTA project polygraphic and clinical database. *IEEE Eng Med Biol Mag* 2001; 20 (3): 51–57.
7. Welch AJ, Richardson PC. Computer sleep stage classification using heart rate data. *Electroen Clin Neuro* 1973; 34 (2): 145–152.
8. Redmond SJ, Chazal P, O'Brien C, Ryan S, McNicholas WT, Heneghan C. Sleep staging using cardiorespiratory signals. *Somnologie – Schlaf-forschung und Schlafmedizin* 2007; 11 (4): 245–256.
9. Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology: Heart rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use. *Eur Heart J* 1996; 17 (3): 354–381.
10. Wessel N, Malberg H, Bauernschmitt R, Kurths J. Nonlinear methods of cardiovascular physics and their clinical applicability. *Int J Bif Chaos* 2007; 17 (10): 3325 – 3371.
11. Wessel N, Ziehmann C, Kurths J, Meyerfeldt U, Schirdewan A, Voss A. Short-term forecasting of life-threatening cardiac arrhythmias based on symbolic dynamics and finite-time growth rates. *Phys Rev E* 2000; 61 (1): 733–739.
12. Voss A, Kurths J, Kleiner HJ, Witt A, Wessel N. Improved analysis of heart rate variability by methods of non-linear dynamics. *J Electrocardiol* 1995; 28 (Suppl): 81–88.
13. Voss A, Hnatkova K, Wessel N, Kurths J, Sander A, Schirdewan A, et al. Multiparametric analysis of heart rate variability used for risk stratification among survivors of acute myocardial infarction. *Pacing Clin Electrophysiol* 1998; 21 (1 Pt 2): 186–192.
14. Wessel N, Malberg H, Meyerfeldt U, Schirdewan A, Kurths J. Classifying simulated and physiological heart rate variability signals. *Comput Cardiol* 2002; 29: 133–135.
15. Wessel N, Malberg H, Heringer-Walther S, Schultheiss HP, Walther T. The angiotensin-(1–7) receptor agonist AVE0991 dominates the circadian rhythm and baroreflex in spontaneously hypertensive rats. *J Cardiovasc Pharmacol* 2007; 49 (2): 67–73.
16. Wessel N, Bauernschmitt R, Wernicke D, Kurths J, Malberg H. Autonomic cardiac control in animal models of cardiovascular diseases. I. Methods of variability analysis. *Biomed Tech (Berl)* 2007; 52 (1): 43–49.
17. Wessel N, Voss A, Malberg H, Ziehmann C, Voss HU, Schirdewan A, et al. Nonlinear analysis of complex phenomena in cardiological data. *Herzschr Elektrophys* 2000; 11 (3): 159–173.
18. Voss A, Kurths J, Kleiner HJ, Witt A, Wessel N, Saparin P, et al. The application of methods of non-linear dynamics for the improved and predictive recognition of patients threatened by sudden cardiac death. *Cardiovasc Res* 1996; 31 (3): 419–433.
19. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer; 2001.
20. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons; 1990.
21. Penzel T, Conradt R. Computer based sleep recording and analysis. *Sleep Med Rev* 2000; 4 (2): 131–148.
22. Penzel T, Behler PG, von Buttler M, Conradt R, Meier M, Möller A, et al. Reliabilität der visuellen Schlafauswertung nach Rechtschaffen und Kales von acht Aufzeichnungen durch neun Schlaflabore. *Somnologie* 2003; 7 (2): 49–58.
23. Danker-Hopfe H, Kunz D, Gruber G, Klösch G, Lorenzo JL, Himanen SL, et al. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. *J Sleep Res* 2004; 13 (1): 63–69.