Clinical Neurophysiology 129 (2018) 2682–2683

Contents lists available at ScienceDirect

Clinical Neurophysiology

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

Letter to the Editor

Machine learning versus human expertise: The case of sleep stage classification in disorders of consciousness. Response to Wislowska et al.

The discussion published in several Letters in *Clinical Neurophysiology* appears to be productive because consensus has been attained regarding most questions except one. This is the question of whether automatic methods of the analysis of sleep stages can, or actually do, outperform scoring by human experts. As regards the former part of the question (i.e., whether they can), we agree that the answer should be positive. As regards the latter part (i.e., whether they already do), some hesitations remain.

Unfortunately, we were unable to find the paper of Anderer et al. (2006) cited by Wislowska et al. (2018). Perhaps this publication is generally difficult to find because, according to Google Scholar, it has only once been cited by somebody outside the same research group. However, we are familiar with other papers of this group published at about the same time (e.g., Anderer et al., 2005, 2010; Saletu et al., 2005) and we are really enthusiastic about the development achieved in these studies as well as the clarity with which their results are presented. Nevertheless, we cannot share the opinion that this system already now can replace human experts.

Firstly, one can read in Anderer et al. (2005) that "the human expert has to decide whether or not, and if so to which extent, the automatic scoring has to be edited and corrected visually" (p. 124–125). Thus, the authors themselves believed that humans can still improve the result of the automatic classification.

Secondly, the reported result of 80% agreement between the automatic system and human experts (as compared with 77% agreement between two experts) is good but not exceptional. In our present work, after highly intensive one-year training of three scorers (Y. G. Pavlov and C. Barner, University of Tübingen; and I. Nopper, Schoen Clinics for Neurological Rehabilitation, Bad Aibling) in the assessment of sleep stages in patients with disorders of consciousness (DoC) and other patients with very severe brain lesions, the agreement in each of the three pairs of scorers was above 80% when using Rechtschaffen & Kales' (1968) 5 stages classification, and above 85% when stages 3 and 4 were not discriminated as recommended by Iber et al. (2007). We emphasize that this level of interrater agreement was attained for scoring highly pathological sleep patterns.

Thirdly and most importantly, we suppose that if human scoring can already be replaced by computer algorithms, it would have already been replaced at least for economic reasons. Given the very high working time costs of high-level experts in Western countries, one can calculate that at least hundreds of millions of dollars have been spent for visual scoring in all sleep labs over the world since the first publications (Anderer et al., 2004, 2005) of the algorithm now suggested by Wislowska et al. (2018). Even if scientists may be so irrational to continuously use less efficient methods while more efficient ones are available (which is unpleasant to think), at least funding agencies such as the National Institutes of Health or the German Research Foundation would have to break off this vast of their money. Moreover, the American Academy of Sleep Medicine (AASM) issued their recommendations for manual sleep scoring a few years after those publications (Iber et al., 2007). To say in mild terms, it would be rather unwise from the AASM to suggest a new set of criteria for visual assessment while there already exists a program making visual assessments obsolete.

We find the proposed approach to train classifiers on a large sample of healthy subjects and then to apply the results on DoC patients questionable for the following reasons. Firstly, as stated above, we do not see clear evidence that automatic classifiers consistently outperform human scorers on healthy sleepers. Secondly, at the empirical level there are big differences between healthy sleepers and DoC patients. Thus, the physiological patterns of *wakefulness* are substantially different and less uniform in patients than in healthy individuals. No healthy subject can ever have delta activity in wakefulness, but in DoC it is not unusual, at least in some EEG leads and from time to time. Human scorers can learn to distinguish between this pathological wakefulness delta and sleep delta and to clearly recognize the transition from one to the other because they are trained on *both normal and pathological* polysomnographies (PSGs). Another example is coming from our data in Pavlov et al. (2017) where we applied an algorithm to calculate sleep spindles density. The algorithm worked well and with comparable precision to a human scorer only after limiting its scope to visually scored NREM stage 2. Otherwise it found numerous spindles in every patient even during wakefulness and REM sleep.

Finally, at a logical level it seems to be a contradiction to claim, on the one hand, that the abnormalities of DoC sleep are too large to be scored even by highly experienced human scorers *trained to score these pathological traces* and, on the other hand, that a classifier can successfully score such severely abnormal traces being trained on normal traces only.

We apologize for potentially imprecise formulations in our previous letter (Kotchoubey and Pavlov, 2018) that gave rise to the idea that we find the process of building classifiers "circular". Of course, we do not think so because if it were circular, classifiers could not outperform human experts simply by definition, which is obviously false. We just mean that creating an automatic assessment program necessarily begins with human expertise. Now chess machines can outperform a grand master and even a world champion, but originally there would be no efficient chess machine without the expertise of grand masters. If, as Wislowska et al.

https://doi.org/10.1016/j.clinph.2018.09.020

(2018) state, human experts do not possess even above-average (e.g., comparable with chess masters, not grand masters) expertise to score sleep patterns in DoC, then we cannot even hope to develop automatic scoring algorithms for this purpose.

To conclude, we agree with the idea of Wislowska et al. (2018) about two ways to cope with the difficulties of the analysis of sleep in DoC patients: (i) the enhancement of human scoring expertise and the respective adjustment of scoring standards to the population of patients, and (ii) the development of machine algorithms. However, we believe that they are not alternatives but rather, two stages of the same process of investigation. First, special training programs should be created and human sleep experts should be trained on large samples of normal *and pathological* PSGs to increase reliability and validity of their scoring. Second, and *only on the basis of this expertise*, computer systems can be developed that would finally beat human scorers.

Conflict of interest statement

Supported by Deutsche Forschungsgemeinschaft, Grant KO/1753-13. The authors thank S. Diekelmann, I. Nopper, C. Barner, and F. Müller.

References

Anderer P, Saletu B, Saletu-Zyhlarz GM, Gruber G, Parapatics S, Miazhynskaia T, et al. Recent advances in the electrophysiological evaluation of sleep. In: Drinkenburg W, Ruigt G, Jobert M, editors. Essentials and applications of EEG research in preclinical and clinical pharmacology. Berlin: Verlag für Studium & Praxis OHG; 2004. p. 307–39.

Anderer P, Saletu B, Saletu-Zyhlarz G, Gruber G, Parapatics S, Miazhynskaia T, et al. Electrophysiological evaluation of sleep. In: Textbook for the training course of the International Pharmaco-EEG Society, September 7 & 8. p. 107–29.

Anderer P, Gruber G, Parapatics S, Woertz M, Miazhynskaia T, Klösch G, et al. An E-Health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24 x 7 utilizing the Siesta database. Neuropsychobiology 2005;51:115–33.

Anderer P, Moreau A, Woertz M, Ross M, Gruber G, Parapatics S, et al. Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: validation study of the AASM version of the Somnolyzer 24 x 7. Neuropsychobiology 2010;62:250–64.

Iber C, Ancoli-Israel S, Chesson Jr AL, Quan SF. The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications. 1st ed. Westchester, IL: American Academy of Sleep Medicine; 2007.

Kotchoubey B, Pavlov YG. Approaches to sleep in severely brain damaged patients: opposite or complementary? Reply to "Sleep and Circadian Rhythms in Severely Brain-Injured Patients – A Comment". Clin Neurophysiol 2018;129:1785–7.

Pavlov YG, Gais S, Müller F, Schönauer M, Schäpers B, Born J, Kotchoubey B. Night sleep in patients with vegetative state. J Sleep Res 2017;26(5):629–40. https://doi.org/10.1111/jsr.12524.

Saletu B, Prause W, Anderer P, Mandl M, Aigner M, Mikova O, Saletu-Zyhlarz GM. Insomnia in somatoform pain disorder: Sleep laboratory studies on differences to controls and acute effects of trazodone, evaluated by the Somnolyzer 24 x 7 and the Siesta database. Neuropsychobiology 2005;51:148–63.

Wislowska M, Blume C, Angerer M, Wielek T, Schabus M. Approaches to sleep in severely brain damaged patients – further comments and replies to Kotchoubey & Pavlov. Clin Neurophysiol 2018;129(12):2680–1.

Boris Kotchoubey *
Institute of Medical Psychology, University of Tübingen, Germany
* Corresponding author at: Institute of Medical Psychology and Behavioral Neurobiology, University of Tübingen, Silcherstr. 5, 72076 Tübingen, Germany.
E-mail address: boris.kotchoubey@uni-tuebingen.de

Yuri G. Pavlov
Institute of Medical Psychology, University of Tübingen, Germany
Department of Psychology, Ural Federal University, Russian Federation

Available online 25 October 2018