Assessing the Viability of Automated Sleep Scoring in an ICU Environment Using a Compact EEG Configuration

Benjamin Wittevrongel[1], Caroline Neuray[1], Alex Grant[2], Baharan Kamousi[2], Raymond Woo[2], Pieter van Mierlo[1]*

*[1] Clouds of Care, Ghent (Belgium) [2] Ceribell, Sunnyvale (US) *Corresponding author*

clouds of care

INTRODUCTION

The presence of sleep elements in the EEG of an ICU patient is correlated with a favorable clinical outcome [1,2], therefore monitoring sleep elements can add value to clinical evaluation and decision-making in these patients. However, in today´s reality, subjective measurements of sleep (by bedside nurse) are unreliable, [3] and a full polysomnography setup is not desirable in the ICU [4]. **We, therefore, aim to evaluate the validity of automated scoring of sleep elements obtained from a minimal EEG setup using the Ceribell point-of-care**

STUDY DESIGN

- Retrospectively collected EEG recordings from 100 ICU patients were selected in this study based on visual assessment of the spectrogram, data quality, and occurrence of sleep graphoelements.
- EEGs were recorded continuously (approx. 6 to 10 hours long) from 8 bipolar EEG channels using the Ceribell point-of-care device.
- Patients with epilepsy diagnosis were excluded, no additional metadata was available.

METHODS

- **2 independent experts** manually scored the data:
  - sleep staging (Wake, NREM, REM)
  - spindles
  - k-complexes
- **Tailored algorithms**, were retrained for the automated sleep scoring in ICU patients:
  - sleep staging: feature extraction & gradient boosting model
  - spindles: feature extraction & rule-based model
  - k-complexes: deep learning model
- **Performance evaluation** based on accuracy for sleep staging and F1-score with 5-fold cross-validation for spindles and K-complexes. The inter-expert agreement was used as the baseline for assessing whether the models reached acceptable performance. The Pearson correlation coefficient was used to investigate trends between the expert and model scoring.

RESULTS

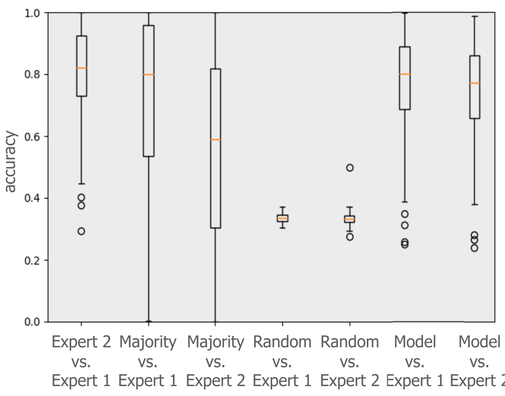
Sleep staging

Expert agreement: accuracy 80.4%

NREM majority class: - Expert 1: 73.3%
- Expert 2: 53.6%

Model performance:

- 76.7% compared to scoring of expert 1 (p=0.11)
- 74.9% compared to scoring of expert 2 (p<0.05)



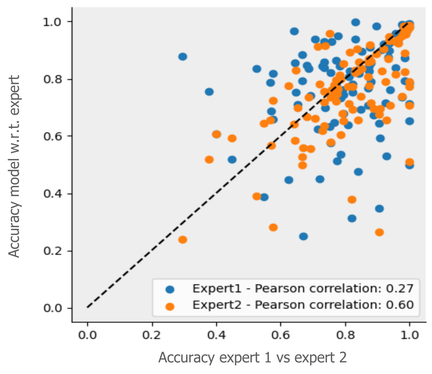
*The **automated scoring has similar accuracies compared to the inter-expert agreement** and exceeds the random and majority models.*

The random model randomly returns one of the three possible sleep stages and the majority model always returns the majority class.

Model performance vs. inter-expert agreement:

ρ = 0.27; **p<0.01**
model vs expert 1

ρ = 0.60, **p<0.01**
model vs expert 2



The accuracy of the model reaches a higher level for patients for whom also the readers have a higher agreement.

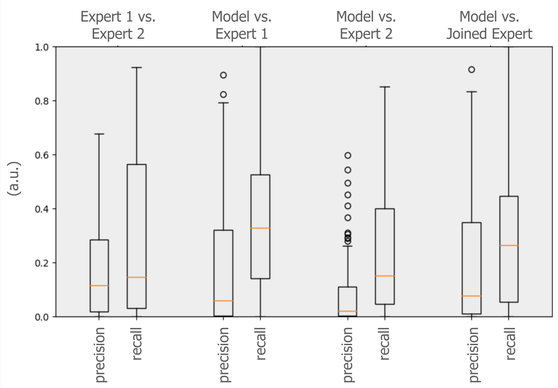
Spindles

Expert agreement: F1 0.22

Experts do not agree strongly on the spindle events.

Model performance:

- F1 0.22 compared to scoring of expert 1 (p=0.55)
- F1 0.13 compared to scoring of expert 2 (p<0.05)

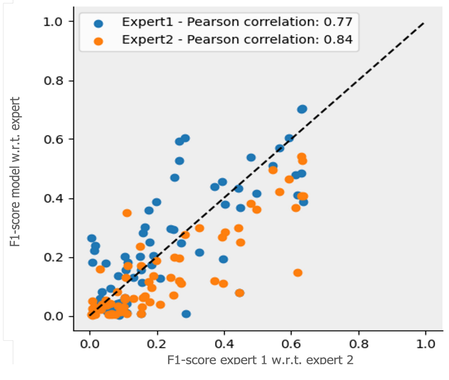


***The automated scoring has similar F1 scores compared to the inter-expert agreement.** The model overestimates the number of spindles (higher recall) compared to the experts, who have more balanced scorings in terms of precision and recall. The higher performance vs Expert 1 is because this expert indicated more spindles than Expert 2.*

Model performance vs. inter-expert agreement:

ρ = 0.77; **p<0.01**
model vs expert 1

ρ = 0.84, **p<0.01**
model vs expert 2



The F1-score of the model reaches a higher level for patients for whom also the readers have a higher agreement. The performance level of the model is in line with the inter-expert agreement.

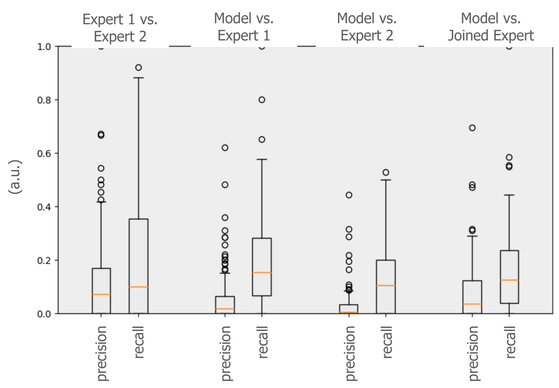
K-complexes

Expert agreement: F1 0.19

Experts do not agree strongly on the k-complex events.

Model performance:

- F1 0.10 compared to scoring of expert 1 (p<0.01)
- F1 0.06 compared to scoring of expert 2 (p<0.01)



***The automated scoring has considerably lower F1 scores compared to the inter-expert agreement.** The model seems unable to capture K-complexes effectively, highlighted by the low precision values.*

Model performance vs. inter-expert agreement:

ρ = 0.52; **p<0.01**
model vs expert 1

ρ = 0.83, **p<0.01**
model vs expert 2

The F1-score of the model reaches a higher level for patients for whom also the readers have a higher agreement. However, the performance level of the model remains considerably below the inter-expert agreement.

CONCLUSION

- Automated sleep staging and detection of spindles and K-complexes in an ICU environment is feasible from a reduced EEG set-up such as the Ceribell point-of-care device.
- Model performance for sleep staging and spindle detection approximate inter-expert agreement, while automated k-complex detection is challenging. Higher model performance is achieved in patients where inter-rater agreement is higher.
- This holds promise toward the clinical utility of automated sleep assessment in an ICU environment derived from an easy-to-use wearable recording device.

LIMITATIONS

- No clinical metadata was available, therefore no subgroup analysis could be performed.
- Guidelines for sleep-scoring provided by the American Academy of Sleep Medicine are difficult to follow in this set-up (reduced EEG montage + recorded on ICU data with many potential confounders).

[1] Sutter, R., Barnes, B., Leyva, A., Kaplan, P.W. and Geocadin, R.G. (2014), Electroencephalographic sleep elements and outcome in acute encephalopathic patients: a 4-year cohort study. Eur J Neurol, 21: 1268-1275. [2] Sandsmark DK, Kumar MA, Woodward CS, Schmitt SE, Park S, Lim MM. Sleep Features on Continuous Electroencephalography Predict Rehabilitation Outcomes After Severe Traumatic Brain Injury. J Head Trauma Rehabil. 2016 Mar-Apr;31(2):101-7. [3] Darbyshire JL, Borthwick M, Edmonds P, Vollam S, Hinton L, Young JD. Measuring sleep in the intensive care unit: Electroencephalogram, actigraphy, or questionnaire? J Intensive Care Soc. 2020 Feb;21(1):22-27. [4] Richards KC, Wang YY, Jun J, Ye L. A Systematic Review of Sleep Measurement in Critically Ill Patients. Front Neurol. 2020 Nov 6;11:542529.