Utilisation of Artificial Intelligence in the Scoring of Nocturnal Polysomnography at the Queensland Children's Hospital

Rylan J. Steinkey[1], Andrew Collaro[1,2], Ajay Kevat[1,2]

[1]Child Health Research Centre, Faculty of Medicine, University of Queensland, Brisbane, Australia
[2]Department of Respiratory and Sleep Medicine, Queensland Children's Hospital, Brisbane, Australia

Introduction

Children suffering from sleep disorders are diagnosed using nocturnal polysomnography (PSG). PSG takes place overnight in the sleep unit and involves highly-trained staff manually identifying the patient's sleep stage on every 30-second epoch of data by examining electroencephalography (EEG), electrooculography (EOG) and electromyography (EMG) measurements, amongst others.

Promising artificial intelligence (AI) approaches to automatic sleep staging are emerging but their accuracy for paediatric patients is unverified. As a quality assurance exercise, concordance between sleep laboratory staff is periodically evaluated by comparing their sleep staging of a PSG extract against that of a "gold standard" scoring determined by expert consensus.

We used this gold standard scoring to evaluate the performance of an existing, freely-available AI automated sleep staging program called U-Sleep.

Questions:

1. How accurate are artificial intelligence models at predicting sleep stages in children?
2. If a child knocks off one of the EEG leads in their sleep, can the AI model use a different lead instead?
3. Which sleep-stages are the AI models good at predicting?

Methods

For six concordance studies, three experts (known as gold contributors) determined what sleep stage the patient was in for every 30-second epoch of roughly 2 hours of nocturnal polysomnography. Staff including nurses, scientists and doctors were then tasked with going over the same polysomnograph, also assigning sleep stages to each 30-second epoch.

We gave each of 3 EEG leads (F4, C4, O2) along with a single EOG lead to the AI model U-Sleep for sleep staging. The human and AI sleep stage classifications were compared to the gold standard for similarity and Cohen's-kappa using Python v3.10.4 and displayed graphically using Seaborn v0.11.2.

Cohen's kappa is used for inter-rater reliability but unlike percent similarity, Cohen's kappa accounts for agreement due to random chance.

$$\text{Cohen's Kappa} = \frac{p(actual) - p(chance)}{1 - p(chance)}$$

Where p(actual) is the observed similarity and p(chance) is the similarity expected due to random chance

Code at: github.com/RylanSteinkey/QCH_sleep

Results

Each concordance has a different patient and as a result, some will be inherently harder to stage than others. If we look at each concordance study separately, the U-Sleep AI exceeds the performance of the nurses (the most abundant group) in 4 of the 6 studies (figure 1). Only one lead is required by the U-Sleep model, and in the concordance study of February 7th, 2022, the best lead is comparable to the human sleep-stagers.

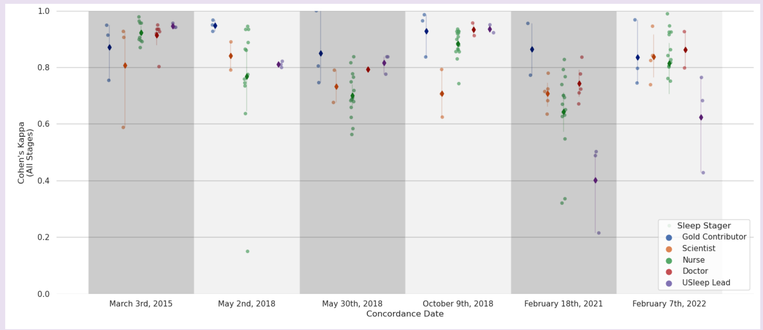

Figure 1. Cohen's Kappa for AI and human labelled polysomnography compared to the gold standard when predicting all sleep stages for each of the six concordance studies. The mean of all staff of the same role is represented with a diamond (error bars represent one standard deviation).

If we are only evaluating REM vs non-REM vs being awake instead of every stage, we see a similar trend as before (figure 2). In the study in which the AI performed most poorly, on manual review the key error the AI made was not correctly identifying REM epochs. Notably, the performance of the top 2 U-Sleep leads is roughly the same as each other, suggesting interchangeability should one lead be removed.

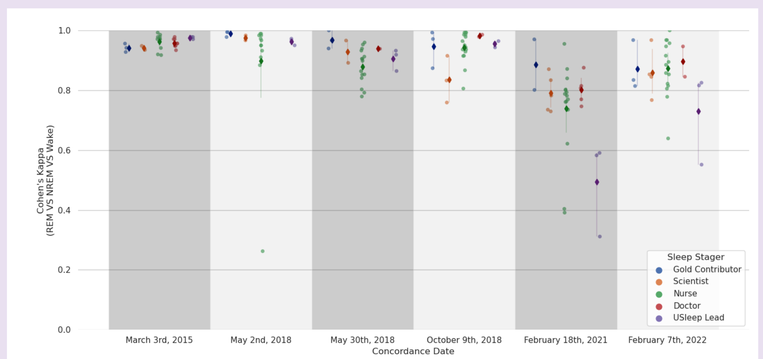

Figure 2. Cohen's Kappa for AI and human labelled polysomnography compared to the gold standard when predicting only REM, non-REM, and awake for each of the six concordance studies. . The mean of all staff of the same role is represented with a diamond (error bars represent one standard deviation).

When looking at the mean accuracies across all 6 concordance studies, the performance of the U-Sleep AI is roughly equivalent to the human stages (figure 3).

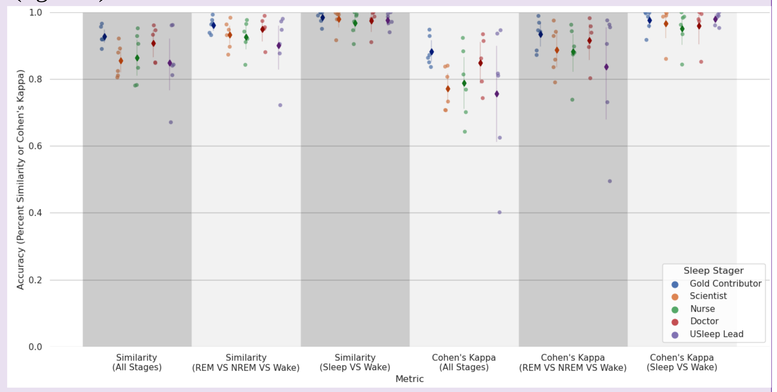

Figure 3. Accuracies of AI and human labelled polysomnography compared to the gold standard for each of six concordance studies. For both percent similarity and Cohen's Kappa, the sleep-stagers are evaluated when they have to predict all stages, when they have to predict if the patient is in REM, non-REM, or awake, or predict if the patient is asleep or awake. The mean of the six concordance studies is represented with a diamond (error bars represent one standard deviation).

Answers:

1. The U-Sleep AI has comparable accuracies to humans.
2. If a child removes an EEG lead in their sleep, another U-Sleep EEG lead can be used in its place.
3. AI sleep staging algorithms warrant validation, edge case testing, and improvement in order to be suitable for routine paediatric use.

Discussion

AI-driven automated sleep staging is a promising means of optimising staff workloads, which may reduce costs and improve sleep laboratory functioning, especially during pandemic-driven staff shortages. AI sleep-stagers use less information than is required by humans, so if a child removes an EEG lead in their sleep, the model can continue functioning using a different lead.

Further research and development is needed to identify and resolve areas of performance weakness in AI models before widespread use in paediatric sleep units can be implemented.