Abstract No:II-3-3

Large Scale Assessment of Consistency in Sleep Stage Scoring Rules among Multiple Sleep Centers Using an Interpretable Machine Learning Algorithm

劉聚仁[1], 林定佑[2], 吳浩榳[3], 許元春[4], 劉景隆[5], 劉文德[6], 楊美貞[7], 倪永倫[8], 周昆達[9], 羅友倫[2]

[1] 國立成功大學數學系, [2] 林口長庚紀念醫院, [3] 杜克大學, [4] 國立交通大學, [5] 台北馬偕紀念醫院, [6] 雙和醫院睡眠中心, [7] 台北慈濟醫院, [8] 台中慈濟醫院, [9] 臺北榮民總醫院

Abstract

STUDY OBJECTIVES: Polysomnography is the gold standard in identifying sleep stages; however, there are discrepancies in how technicians use the standards. Because organizing meetings to evaluate this discrepancy and/or reach a consensus among multiple sleep centers is time consuming, we developed an artificial intelligence (AI) system to efficiently evaluate the reliability and consistency of sleep scoring, and hence the sleep center quality.

METHODS: An interpretable machine learning algorithm was used to evaluate interrater reliability (IRR) of sleep stage annotation among sleep centers. The AI system was trained to learn raters from one hospital, and applied to subjects from the same or other hospitals. The results were compared with the experts' annotation to determine IRR. Intra-center and intercenter assessments were conducted on 679 subjects without sleep apnea from six sleep centers in Taiwan. Centers with potential quality issues were identified by the estimated IRR.

RESULTS: In the intra-center assessment, the median accuracy ranged from 80.3% to 83.3% with the exception of one hospital (designated E) with an accuracy of 72.3%. In the inter-center assessment, the median accuracy ranged from 75.7% to 83.3% when hospital E was excluded from testing and training. The performance of the proposed method was higher for N2, awake, and REM, compared to N1 and N3. The significant IRR discrepancy of hospital E suggested a quality issue. This quality issue is confirmed by the physicians in charge of hospital E.

CONCLUSIONS: The proposed AI system proved effective in assessing IRR and hence the sleep center quality.

Keyword: interrater reliability, intra-center assessments, inter-center assessments, machine learning, sleep stage scoring