A Cross-Validation Approach to Inter-scorer Reliability (ISR) Assessment

An application of novel Machine Learning methods to improve ISR evaluation

Chris Fernandez, MS[1,2], Samuel Rusk[1,2], Nick Glattard, MS[1,2], Broderick Hensen[2],
Mehdi Shokoueinejad, Ph.D.[3], Shane Creado, M.D.[2], James Hungerford, M.D.[4].
[1]Department of Population Health Sciences, University of Wisconsin-Madison, [2]EnsoData Research, EnsoData,
[3]Department of Biomedical Engineering, University of Wisconsin-Madison,
[4]Department of Pediatrics Sleep Medicine, University of Arkansas for Medical Sciences

Introduction & Motivation

- The aim of ISR assessment is to estimate the *generalized* scoring performance and competency of clinicians relative to a designated scoring reference measured by epoch-by-epoch sleep stage and event scoring agreement.
- ISR assessment plays in important role in AASM Accreditation, sleep technologist certification, automated scoring validation, analysis of variability, and overall clinical quality management
- ISR assessments may be conducted:
  - Locally, where clinicians are compared to an individual reference scorer [1]
  - Or via the AASM ISR assessment, which compares to a consensus of multiple reference scorers
- The AASM Digital Task Force evidence grading framework for clinical validation studies recommends blind independent scoring by multiple scorers to set a reference standard [2]
- In this work, we aim to explore the differences that result from comparing clinicians to *individual* reference scorers versus a *consensus* multiple reference scorers:
  - To what extend can ISR evaluation be biased by scoring attributes of an individual reference scorer?
  - Does the inclusion requirement of establishing consensus among multiple reference scores effect the total number of events or diagnostic indices?
  - Do these differences impact the generalizability and utility of individual vs consensus ISR assessments?

Leave-one-out Cross-Validation (LOOCV) for ISR

Leave-one-out-cross-validation (LOOCV) is an powerful technique for evaluating how the results of a statistical analysis will *generalize* to an independent dataset.

In the present study, we adapt the LOOCV approach from machine learning, proposing a novel application of the methodology to analyze ISR assessments, whereby we elucidate statistical relationships of clinical relevance and introduce the concept of *overfitting* in the sleep scoring context.

Study Sample and Experimental Controls

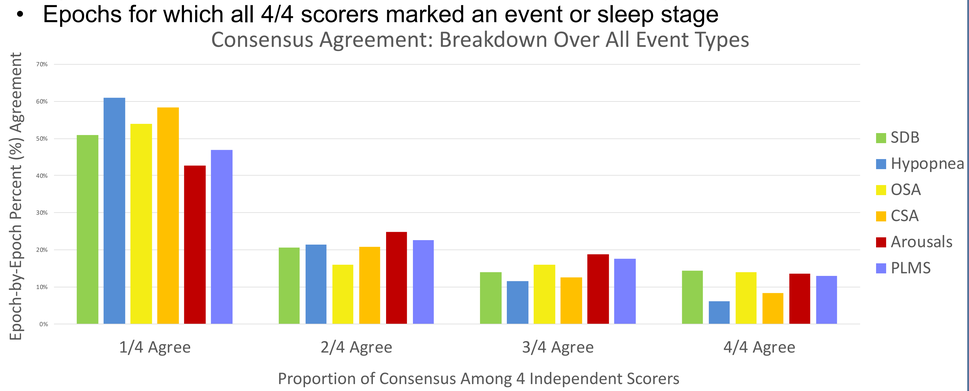
- A diagnostic cohort (N=72) was selected using stratified sampling with proportionate allocation to control for sleep apnea severity, medical conditions, medications, and demographic factors.
- The cohort was scored by four independent sleep technologists (RPSGT). ISR was assessed by epoch-by-epoch agreement for sleep stages, respiratory, arousal, and movement events under two LOOCV settings;
- First, average agreement was calculated with each clinician as the compared to each of the three “held out” clinicians.
- Second, average agreement was calculated by constructing a reference based on events that a 2/3 majority of clinicians agree with compared to the fourth “held out” clinician.

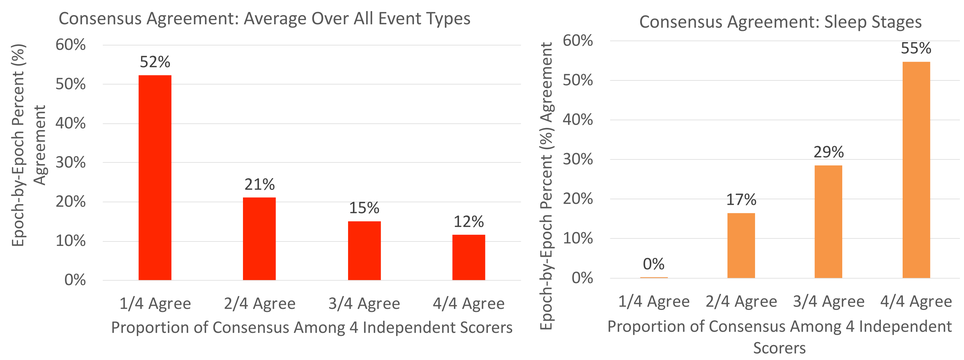
Statistical Analysis of Individual Versus Consensus Scoring Agreement

Distribution of Individual and Consensus Scoring Agreement

First, the epochs for all N=72 epochs were pooled, leading to a sample of 59,719 total sleep study epochs with each scored by 4 independent scorers. Sleep scoring events including Sleep Disordered Breathing (SDB), Obstructive Sleep Apnea (OSA), Central Sleep Apnea (CSA), Hypopnea, Sleep Stages, Arousals, and Periodic Limb Movements in Sleep (PLMS) were segmented into four buckets:

- Epochs for which only 1/4 scorers marked an event or sleep stage
- Epochs for which 2/4 scorers marked an event or sleep stage
- Epochs for which 3/4 scorers marked an event or sleep stage
- Epochs for which all 4/4 scorers marked an event or sleep stage





Statistical Significance of Differences in Individual and Consensus Scoring

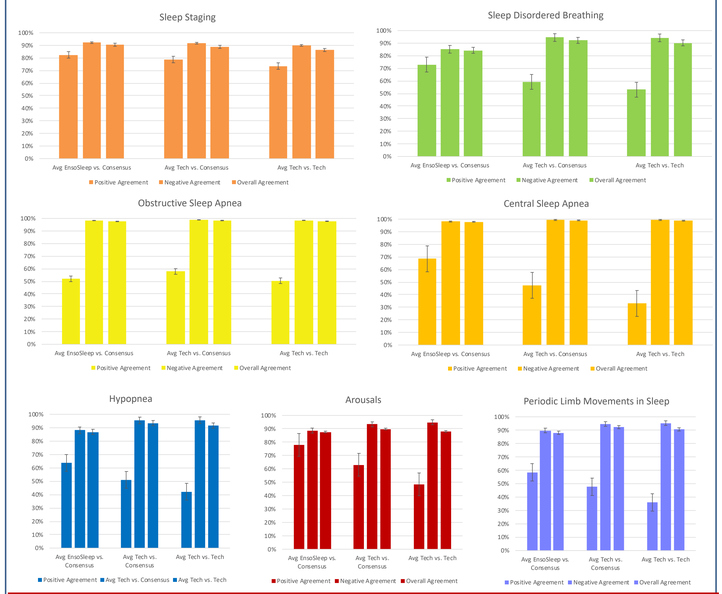
The average and standard deviation of the number of epochs marked by each of the 4 scorers was calculated for each event type for each subject. The average and standard deviation were then calculated based on 4 unique combinations of 2/3 majority scoring. A two-sample t-test assuming unequal variance was applied to analyze the statistical significance of any differences in mean and variance estimates for the number of events.

Event Type	p-value of Mean # Events	p-value of Standard Deviation	p-value of Mean # Events < 0.05?	p-value of Standard Deviation < 0.05?
SDB	0.17	6.16E-10	No	Yes
Hypopnea	0.01	1.92E-11	Yes	Yes
OSA	0.48	2.09E-07	No	Yes
CSA	0.42	4.44E-03	No	Yes
Arousals	0.15	2.70E-15	No	Yes
PLMS	0.44	2.00E-09	No	Yes

LOOCV for Individual and Consensus ISR Assessments

LOOCV Comparison of Consensus vs. Individual References

- First, LOOCV was used to compare each individual scorer to all other individual scorers. Then, each individual scorer was compared to a 2/3 Majority of the “held out” scorers. Lastly, an AI based automated scoring system was compared to a 2/3 Majority of ”held out scorers”. The average was Positive Agreement (PA), Negative Agreement (NA), and Overall Agreement (OA), with two-sided 95% bootstrap percentile confidence intervals were computed to provide an estimate of scoring performance in each ISR assessment configuration.



Discussion & Conclusions

- On average, 52% of all epochs marked as having an SDB, OSA, CSA, Hypopnea, Arousal, or PLMS event were marked as such by only 1 of 4 scorers, suggesting significant variability in scoring attributes, while only 12% of all epochs were marked as such by a unanimous consensus of all 4 of 4 scorers
- With the exception of Hypopnea events, no statistically significant differences were observed in the mean number of epochs marked with events by individual vs consensus scoring
- Statistically significant differences were observed in the standard deviation of individual vs consensus scoring for all event types analyzed, with greater variability among individuals
- Despite on average marking the same number of epochs with events, ISR agreement was observed to be greater for all event types when compared to consensus vs individual scoring, including statistically significant differences in some event types
- LOOCV enables measurement of ISR agreement with greater generalizability, by reducing potential overfitting of ISR assessments relative to an individual’s scoring attributes