**Background:** Common quantitative scalar evaluation metrics such as sensitivity and specificity can often be misleading for sequential data applications. There is a lack of standardization of scoring metrics in biomedical applications involving sequential data.

**Methods:** We analyze three popular scoring metrics and introduce two new metrics for a seizure detection task. We compare and contrast their results collected from machine learning models developed on the TUH EEG Corpus. Performance evaluated in terms of sensitivity and specificity does not address the time scales over which the scoring must occur. This is critical for sequential data applications. To address these issues, our proposed metrics introduce two classes of scoring metrics: term-based and time-aligned. We also compare these metrics using a more holistic view based on a Detection Error Trade-off curve.

**Results:** We show that time-aligned scoring is consistent with popular scoring approaches but provides more accurate assessments and diagnostics by comparing the degree of match. We also evaluate existing metrics adapted from the speech recognition community, where sequential scoring techniques are very mature.

**Conclusion:** The metrics proposed in the study are excellent candidates for standardizing scoring across the industry. These metrics will be used in several upcoming open-source seizure detection challenges.