Good work on this!!

Systematic Review and Meta-Analysis Protocol

Title

Classification Sensitivity and Specificity of Machine Learning Algorithms for Obstructive Sleep Apnea Classification utilizing: A Systematic Review and Meta-Analysis of Sleep Questionnaire-Based Models Compared to In-Lab Polysomnography.

1. Background and Rationale

Obstructive sleep apnea (OSA) is a highly prevalent sleep disorder that can lead to severe cardiovascular and metabolic consequences if left untreated. Polysomnography (PSG) is the gold standard for diagnosing OSA, but it is costly, time-consuming, and not widely accessible. Machine learning (ML) models have emerged as promising alternatives for classifying OSA severity using sleep questionnaire data. However, the performance of these models varies, and a systematic synthesis of their classification accuracy is needed to assess their clinical utility.

This systematic review and meta-analysis aim to evaluate the diagnostic accuracy of ML-based models for detecting OSA severity (no OSA, mild, moderate, severe) using sleep questionnaire data compared to PSG.

2. Objectives

- To assess the classification accuracy of ML models in predicting OSA severity using sleep questionnaire data.
- To compare the sensitivity, specificity, AUC-ROC, and other performance metrics of different ML models.
- To evaluate the methodological quality and risk of bias in included studies.
- To identify factors contributing to heterogeneity in model performance (e.g., dataset diversity, model type, sample size).

3. PICO Framework and Research Question

PICO Framework

- **Population (P):** Adults undergoing sleep assessments for suspected OSA.
- **Index Test (I):** Machine learning models trained on sleep questionnaire data.
- **Comparator (C):** Standard PSG-based OSA classification.

- **Outcomes (O):** sensitivity, specificity, and AROC.

Research Question

This study aims to evaluate the classification accuracy of machine learning algorithms for detecting obstructive sleep apnea (OSA) severity using sleep questionnaire data, in comparison to polysomnography (PSG) as the gold standard. The research will investigate different machine learning models, their performance metrics (sensitivity, specificity, AUC-ROC, etc.), and the impact of dataset characteristics on predictive accuracy. Additionally, the study will assess the applicability of these models in clinical practice, identifying potential limitations and areas for improvement**.**

4. Methods

In the following section, we present the methods intended for the selection of studies in the meta-research.

4.1 Eligibility Criteria

Inclusion Criteria:

- Studies that apply ML models for OSA classification based on sleep questionnaire data. Are sleep questionnaires standardized or are they widely variable? Which questionnaires are commonly used?
- Use of PSG as the reference standard (I assume this is sleep lab PSG, not portable versions?)
- Reporting of classification accuracy metrics (e.g., sensitivity, specificity, AUC-ROC, F1-score). I would not recommend these terms be in the search, I think this needs to be determined by human reviewers. Thoughts?)
- Studies published in peer-reviewed journals or reputable conference proceedings.
- Adult population (18+ years).

Exclusion Criteria:

- Studies using only physiological signals (e.g., ECG, SpO2) without questionnaire data.
- Non-ML statistical models (e.g., logistic regression alone).
- ~~Studies without sufficient classification performance reporting. (Please, explain more so I can understand)~~

- ~~Reviews~~, commentaries, or case reports. Note: Reviews of all sorts can be great sources through the references they cite.

4.2 Search Strategy

A comprehensive search will be conducted in the following databases:

- **PubMed**
- **Embase**
- **IEEE Xplore**
- **Cochrane Library** this is really a database of RCTs and the systematic reviews of interventions based on the RCTs. Unlikely to have the kind of material you want. The material in Cochrane appears in Embase and PubMed.
- **Google Scholar**
- **Consider Web of Science, Scopus, Biosis.**

Search Terms:

("machine learning" OR "artificial intelligence" OR "deep learning") AND ("obstructive sleep apnea" OR "OSA") AND ("questionnaire" OR "survey" OR "screening") AND ("classification" OR "diagnosis").

To show the effect of phrase searching (using quotation marks) vs not: See tab 2

https://docs.google.com/document/d/1HG170BHUj2TZrcnDKMeuxhkmrLf7RIEG1gtYTPOqx-I/edit?tab=t.ca27xim7sgjo

- Following your inputs (thank you again), let's modify the search to include only machine learning, that is what is used for questionnaires. Deep learning and AI is used more for signals. This reduces to ~300. For the 300, do I need to read each abstract or select the N first?

(machine learning) AND (obstructive sleep apnea OR OSA) AND (questionnaire OR survey OR screening) AND (classification OR diagnosis)

Hand-searching of relevant systematic reviews and reference lists of included studies will also be performed.

4.3 Study Selection Process (You might use Covidence to help organize the work)

- ~~**Screening:** Two independent reviewers will screen titles and abstracts based on eligibility criteria.~~ I would keep this in
- **Full-text review:** Articles passing the initial screening will undergo full-text evaluation.
- **Disagreements:** Resolved through discussion or consultation with a third reviewer.
- **PRISMA Flow Diagram** will be used to document the study selection process.

4.4 Data Extraction (You might use Covidence to help organize the work)

A standardized data extraction form will be used to collect the following:

- Study details (author, year, sample size, country, study design).
- ML model type (e.g., SVM, random forest, neural networks).
- Threshold values used for the definition of apnea event of not e.g., AHI >5, or AHI >15, or multiple thresholds for the different severity values.
- Sleep questionnaire used (e.g., STOP-BANG, Berlin, ESS).
- Performance metrics (accuracy, sensitivity, specificity, AUC-ROC, precision, recall, F1-score).
- Training/validation sample sizes and dataset diversity.

4.5 Risk of Bias and Quality Assessment (You might use Covidence to help organize the work)

- **RoB 2** tool will be used for randomized controlled trials (RCTs).
- **QUADAS-2** for diagnostic accuracy studies.
- **GRADE framework** to assess the certainty of evidence.

4.6 Data Synthesis and Meta-Analysis (You might use Covidence to help organize the work)

- **Descriptive Synthesis:** If meta-analysis is not feasible due to heterogeneity, findings will be narratively synthesized.
- **Meta-Analysis (if applicable):**

- Random-effects model using inverse-variance weighting.
- Sensitivity and specificity will be pooled using hierarchical summary ROC curves.
- Subgroup analyses based on ML model type, dataset characteristics, and questionnaire used.
- Heterogeneity assessment using I^2 statistics, τ^2 estimation, and meta-regression.
- Publication bias assessment via funnel plots and Egger's test.

5. Discussion and Impact

This study will provide a comprehensive synthesis of ML-based screening models for OSA. Findings will inform clinicians and researchers about the reliability and applicability of ML models for classifying OSA severity using questionnaire data, potentially improving early screening and reducing reliance on PSG.