AI-Augmented Narrative Screening: Using a Multi-Modal Processing Engine to Reduce Failure Rates in Central Nervous System Trials

Submitter Benjamin Israel

Affiliation Featherglass Health, Inc.

SUBMISSION DETAILS

I agree to provide poster pdf for attendee download. Yes

I have used the poster abstract template to develop my abstract. Yes

Methodological Issue Being Addressed Central nervous system (CNS) clinical trials experience high screen-failure rates—averaging 57% in psychiatry and up to 88% in preclinical Alzheimer's disease. This inefficiency is driven largely by the heterogeneity of eligibility data: workflows depend on narrative clinical notes, prior records, psychometric evaluations, and semi-structured interviews, which vary significantly in format and subjectivity.

Current screening methods also often lack reproducibility, as individual raters may interpret ambiguous clinical histories differently. A clinical narrative processing engine - utilizing Retrieval-Augmented Generation (RAG) to interpret unstructured text, audio, and structured scores - offers a potential solution. However, validating such a system requires rigorous statistical demonstration of non-inferiority to expert consensus and accurate uncertainty calibration.

Introduction Central nervous system (CNS) clinical trials continue to experience high screen-failure rates – 57% on average, and up to 88% in preclinical Alzheimer's disease – largely because eligibility criteria must be applied to heterogeneous data sources that do not share a common structure. Existing workflows depend on narrative clinical notes, prior records, psychometric evaluations, and clinician-administered screening interviews, yet these sources vary significantly in format, completeness, and objectivity. This creates inconsistency in determining eligibility and contributes to screening failures.

Methods A prospective, blinded validation study will evaluate the Featherglass AI engine on a stratified random sample of N=500 diverse CNS trial profiles drawn from multiple health systems to ensure generalizability. The AI model (including prompt logic and retrieval parameters) will be frozen prior to validation to prevent data leakage; the validation set is strictly prospective and distinct from the training corpus. Participant data will include structured inputs (e.g., psychometric scores, lab results) and unstructured natural language (e.g., clinical notes, screening interview transcripts).

● Gold Standard Establishment: Eligibility will be determined by a Tiered Independent Clinical Adjudication Panel (ICAP). Two blinded senior clinicians will independently review full patient dossiers. Discordant cases will be resolved by a third senior adjudicator to establish the consensus ground truth.

● Performance Metrics: We will calculate Sensitivity, Specificity, Positive/Negative Predictive Values (PPV/NPV), and Balanced Accuracy with 95% Wilson Score confidence intervals. We will generate the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) to evaluate discrimination thresholds.
● Statistical Comparison: Differences in classification performance between the AI and the human baseline will be evaluated using McNemar's test for paired nominal data ($\alpha = 0.05$).
● Uncertainty Calibration: To validate the system's risk stratification, human raters will assign a Diagnostic Confidence Score (Likert scale 1-5). We will assess calibration using Reliability Diagrams and the Brier Score to quantify the concordance between AI uncertainty signaling and human-rated ambiguity.
● Efficiency Analysis: Workflow impact will be evaluated using a paired Wilcoxon Signed-Rank Test to compare median time-to-determination between manual and AI-assisted review, controlling for case complexity.

Results We project the analysis of 500 profiles will demonstrate sufficient power (>99%) to detect significant differences in workflow efficiency and provide high precision (±4.5%) for sensitivity estimates.

● Diagnostic Accuracy: We project the AI engine will achieve a Balanced Accuracy ≥0.87 and substantial inter-rater agreement (Cohen's $k > 0.80$) with the ICAP gold standard. We anticipate no statistically significant difference in error rates (discordant pairs) between the AI and human consensus as measured by McNemar's test.
● Calibration: We hypothesize a strong positive correlation between AI entropy scores and human low-confidence ratings, with a low Brier Score indicating that the system accurately identifies ambiguous cases.
● Efficiency: We hypothesize a statistically significant reduction in screening time ($p < .001$), reducing the median review time from approximately 45 minutes to under 10 minutes.

Conclusion An AI-enabled narrative screening engine capable of integrating multimodal data may significantly reduce screen-failure rates in CNS trials. By validating this technology against a tiered consensus gold standard and rigorously measuring uncertainty calibration, this study aims to demonstrate that AI-augmented screening is not only more efficient but scientifically robust enough for regulatory-grade clinical research. This approach offers a scalable method to unify eligibility assessments, reducing the "noise" of ineligible enrollment and improving signal detection in complex neuropsychiatric trials.

Co-Authors

Benjamin Israel[1], Matthew Robinson[2], Scott Aaronson[3]

[1] Featherglass Health, Inc.

[2] Featherglass Health, Inc.; Mass General Brigham

[3] Sheppard Pratt Health System

Keywords

Keywords
Artificial intelligence
Large language model
Screening failure

Guidelines I have read and understand the Poster Guidelines

Disclosures Benjamin Israel, M.D., and Matthew Robinson, PhD, have a relevant financial relationship with an ACCME-defined commercial interest Featherglass Health, Inc. (CEO) All relevant financial relationships listed have been mitigated to ensure that the poster presentation will be evidence-based and unbiased.