AI-Augmented Narrative Screening

Using a Multi-Modal Processing Engine to Reduce Failure Rates in Central Nervous System Trials

Authors

Benjamin Israel MD, Mathew Robinson PhD, Scott Aaronson MD

Affiliations

Featherglass Health, Inc. (1, 2), Mass General Brigham (2), Sheppard Pratt Health System (3)

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 heterogeneous narrative sources vary in format, completeness, and objectivity. This creates inconsistency in determining eligibility and contributes to screening failures.

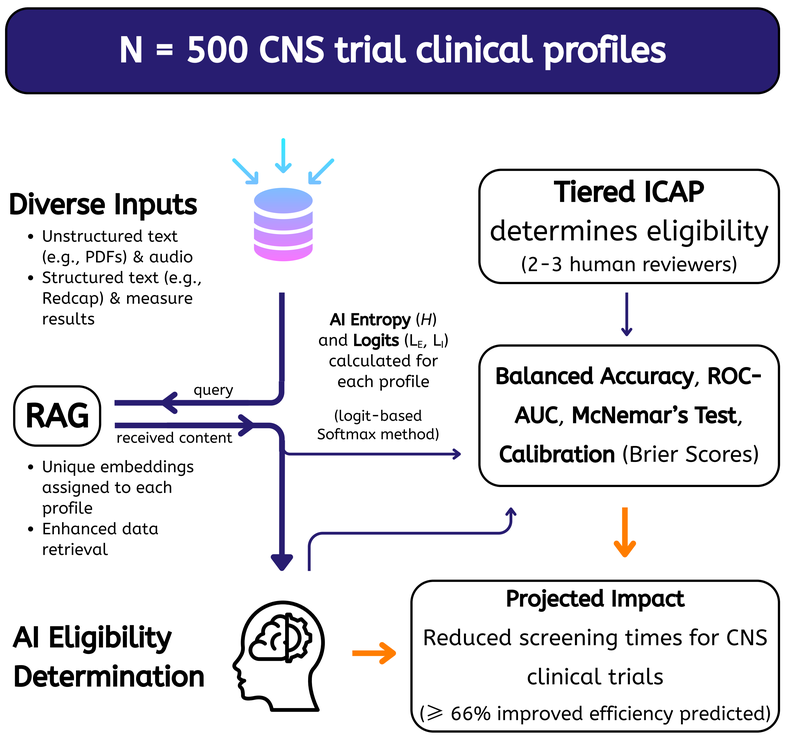
A clinical narrative processing engine – utilizing Retrieval-Augmented Generation (RAG) to interpret unstructured text, audio, and structured scores – offers a potential solution. However, validating such a system requires rigorous statistical demonstration of non-inferiority to expert consensus and accurate uncertainty calibration.

Methods

A prospective, blinded validation study will evaluate a RAG-specialized clinical narrative 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).

1. **Gold Standard Establishment**: Eligibility will be determined by a Tiered Independent Clinical Adjudication Panel (ICAP).
2. **Performance Metrics**: We will calculate Sensitivity, Specificity, Positive/Negative Predictive Values, 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.
3. **Statistical Comparison**: Differences in classification performance between the AI and the human baseline will be evaluated using McNemar's test for paired nominal data ($\alpha = 0.05$).
4. **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.
5. **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 (Projected)

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.

1. **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.
2. **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.
3. **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 to enable 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.

Softmax function

$$P(\text{Eligible}) = \frac{e^{L_E}}{e^{L_E} + e^{L_I}} \qquad P(\text{Ineligible}) = \frac{e^{L_E}}{e^{L_E} + e^{L_I}}$$

AI Entropy

$$H = -[P(\text{Eligible}) \times \log_2(P(\text{Eligible})) + P(\text{Ineligible}) \times \log_2(P(\text{Ineligible}))]$$

References

- Hendrycks, D., & Gimpel, K. (2017). "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks."Proceedings of the International Conference on Learning Representations (ICLR).
- Shannon, C. E. (1948). "A Mathematical Theory of Communication." The Bell System Technical Journal, 27(3), 379–423.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. (Chapter 6.2.2.3, "Softmax Units for Multinoulli Output Distributions").
- Kompa, B., Snoek, J., & Beam, A. L. (2021). "Second opinion needed: communicating uncertainty in medical machine learning."NPJ Digital Medicine, 4(1)
- Askin, S., Burkhalter, D., Calado, G., & El Dakrouni, S. (2023). Artificial Intelligence Applied to clinical trials: Opportunities and challenges. Health and Technology, 13(2), 203–213.
- Bernasconi, L., Avakyan, G., Hovaguimian, F., & Grossmann, R. (2025). Natural Language Processing in Clinical Research Recruitment: A Scoping Review Enriched with Stakeholder Insights. Ethics & Human Research, 47(5), 13–23.
- Jin, Q., Wang, Z., Floudas, C. S., Chen, F., Gong, C., Bracken-Clarke, D., Xue, E., Yang, Y., Sun, J., & Lu, Z. (2024). Matching patients to clinical trials with large language models. Nature Communications, 15(1), 9074.

Featherglass HEALTH