Introduction As artificial intelligence (AI) systems become increasingly integrated into clinical decision-making, the development of transparent, clinician-informed datasets has become paramount. In the field of aphasia rehabilitation, treatment selection is a sophisticated process shaped by a clinician's theoretical orientation and years of hands-on experience. However, this expertise often remains implicit and difficult to quantify for automated systems. Capturing this clinical variation in a systematic format is a necessary step toward building trustworthy AI-based recommendation tools that support, rather than replace, human judgment. This roundtable session explores the challenges of formalizing clinical intuition into objective algorithmic frameworks and seeks to define the future of human-AI collaboration in speech-language pathology.
Aims The primary objective of this session is to foster an academic discourse on reconciling professional clinical variability with automated algorithmic design. We aim to analyze the degree of consensus between experts when selecting treatments for diverse aphasia profiles and evaluate the performance of a rule-based AI engine designed to mimic such clinical logic. Furthermore, we seek to discuss the technical and ethical implications of "clinical divergence" for the future of rehabilitation technology. A critical outcome of this session is to establish a collaborative foundation for an Expert Delphi Panel. This initiative will involve roundtable participants in the longitudinal task of refining automated treatment logic through expert consensus, moving beyond individual variability toward a standardized framework.
Content This discussion is grounded in a pilot study involving six aphasia profiles constructed from anonymized Western Aphasia Battery-Revised (WAB-R) [1] scores representing a range of subtypes and severity levels. These profiles were obtained from the AphasiaBank database [2]. In this study, two expert clinicians independently selected treatment approaches from a predefined checklist and provided rationales for their decisions. These human decisions were compared against a rule-based AI engine developed by mapping WAB-R profile features, such as Aphasia Quotient (AQ), Naming, Comprehension, and Repetition, to expert-informed treatment rules, as detailed in Table 1.

The resulting data reveals significant insights into clinical reasoning patterns. While the percent agreement between the two clinicians was relatively high (mean = 79.49%), Cohen's Kappa was 0.31, reflecting only fair-to-moderate reliability [3]. This indicates that experts often construct different "treatment packages" even when presented with the same patient data. Furthermore, the AI system showed a mean Jaccard similarity of 0.59 with Clinician A, but only 0.25 with Clinician B [4]. As shown in Figure 1, complete divergence (Jaccard = 0) occurred in specific mid-range severity profiles, highlighting a fundamental tension in whose logic should serve as the "standard" for AI training. These findings underscore the complexity of automating expert intuition through rule-based systems [5].
Questions for Discussion The session will address several critical questions derived from these findings. ***First,*** when experts demonstrate only moderate agreement (Kappa 0.31), who should determine the "ground truth" for AI training, and should AI reflect a singular consensus or a diversity of clinical schools of thought? ***Second,*** we will evaluate whether the specific numerical thresholds used in our AI logic, such as a Naming score below 6 triggering Semantic Feature Analysis (SFA), align with the broader clinical intuition of the participants [6,7]. ***Third,*** we will debate the trade-offs

between transparency and complexity: is an explainable rule-based system more trustworthy in clinical practice than a high-accuracy but "black-box" model? ***Finally,*** we will consider the issue of clinical sovereignty, specifically, how a clinician should respond when an AI suggests a treatment path that is diametrically opposed to their own judgment [8,9].

Participant Engagement Methods and Strategies To ensure active engagement, this session will utilize a "Case Mapping" activity. Participants will be provided with sample WAB-R profiles representing different severity levels and will be asked to prioritize treatment interventions. Subsequently, we will reveal the AI's automated recommendations for these profiles, facilitating a real-time comparison between human expertise and algorithmic output. This will lead to a discussion on the rationales behind any observed divergence. The session will conclude with a formal invitation to join an Expert Delphi Panel. This initiative aims to involve the participants in a series of iterative feedback rounds to refine the AI's recommendation logic. By engaging the audience in this process, we treat clinician variability not as "noise," but as a vital data source for building robust, clinician-informed rehabilitation research frameworks.

References

[1] Kertesz, A. (2007). *Western Aphasia Battery–Revised (WAB–R).* Pearson.

[2] Macwhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for Studying Discourse. *Aphasiology*, *25*(11), 1286–1307. https://doi.org/10.1080/02687038.2011.589893

[3] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

[4] Chung, N. C., Miasojedow, B., Startek, M., & Gambin, A. (2019). Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. *BMC Bioinformatics*, 20, 644.

[5] Hayes-Roth, F. (1985). Rule-based systems. *Communications of the ACM*, 28(9), 921–932.

[6] Efstratiadou, E. A., Papathanasiou, I., Holland, R., Archonti, A., & Hilari, K. (2018). A systematic review of semantic feature analysis therapy studies for aphasia. *Journal of Speech, Language, and Hearing Research*, 61(5), 1261–1278.

[7] Kendall, D. L., Oelke, M., Brookshire, C. E., & Nadeau, S. E. (2015). The influence of phonomotor treatment on word retrieval abilities in 26 individuals with chronic aphasia: An open trial. *Journal of Speech, Language, and Hearing Research*, 58(3), 798–812.

[8] Wallace, S. E., Patterson, J., Purdy, M., Knollman-Porter, K., & Coppens, P. (2022). Auditory comprehension interventions for people with aphasia: A scoping review. *American Journal of Speech-Language Pathology*, 31(5S), 2404–2420.

[9] Schuell, H., Jenkins, J. J., & Jimenez-Pabon, E. (1964). *Aphasia in adults: Diagnosis, prognosis, and treatment*. New York: Harper & Row.

Figure 1. Jaccard Similarity Between Clinician Treatment Recommendations and AI Outputs Across Aphasia Profiles

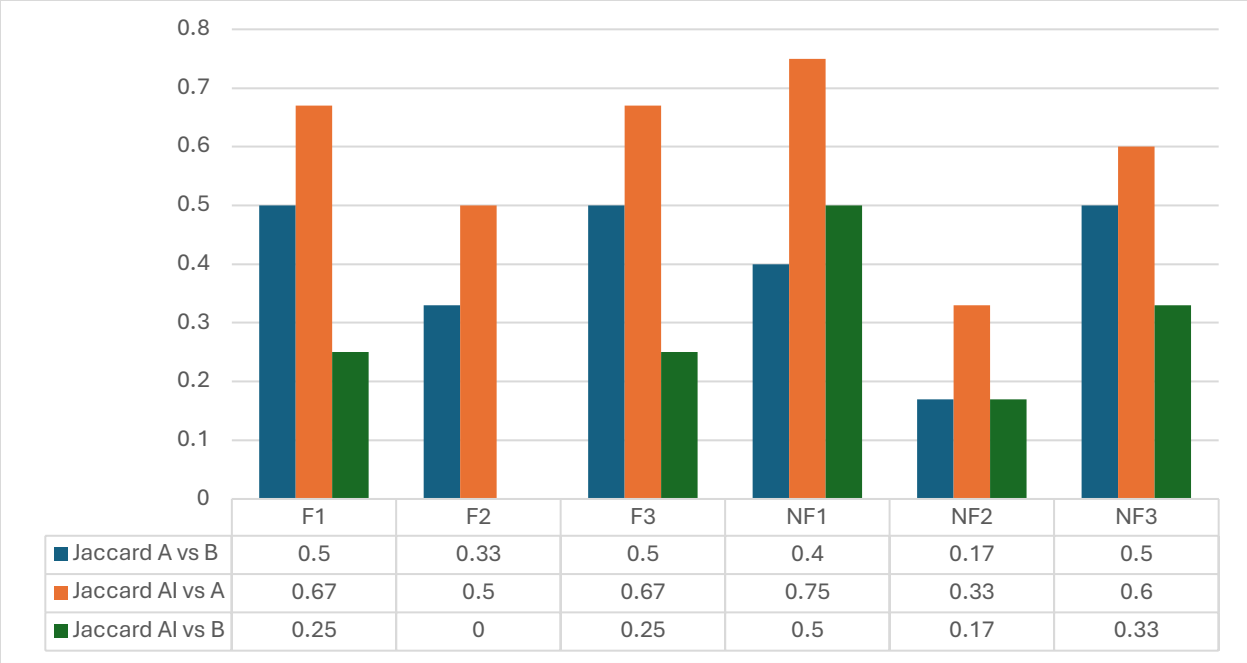

Figure 1. Jaccard similarity between clinician treatment recommendations and AI outputs across aphasia profiles. The bar chart illustrates the degree of overlap in treatment selection among Clinician A, Clinician B, and the rule-based AI engine for six distinct profiles (F1-F3, NF1-NF3). While Clinician A and the AI engine show relatively high similarity across most profiles, reaching a peak of 0.75 for NF1, significant divergence is observed with Clinician B. Notably, Case F2 demonstrates a Jaccard similarity of 0.0 between the AI and Clinician B, highlighting the fundamental challenge of establishing a singular algorithmic standard amidst professional clinical variability.

Table 1. Rule-Based AI Treatment Logic

Condition (based on WAB-R scores)	Recommended Approaches & Treatment(s)	Rationale
Naming < 6	SFA, Phonomotor Treatment	Frequently selected by clinicians
Comprehension < 6	Auditory Comprehension Training	Triggered consistently by clinicians
AQ ≥ 85 or both Naming and Comprehension preserved	Story Retelling, Group Discourse Therapy, SFA	Consensus among clinicians for high-functioning profiles
AQ < 30 or Naming = 0	AAC, Visual Action Therapy, Schuell's Stimulation Approach (SSA)	Global aphasia protocols; expert judgment
AQ 40–85 AND Naming ≥ 5 AND Comprehension ≥ 6	Script Training, Story Retelling, Group Discourse Therapy	Added by AI to address under-represented mid-range profiles
+ Repetition ≥ 4	MIT, SPP	Added for modeling-based training support

Note: This rule-based AI recommendation logic was derived from observed clinician patterns across six aphasia profiles. It maps intervention strategies to WAB-R profile features, such as AQ, Naming, Comprehension, and Repetition. Abbreviations: SFA: Semantic Feature Analysis; AAC: Augmentative and Alternative Communication; SSA: Schuell's Stimulation Approach; MIT: Melodic Intonation Therapy; SPP: Sentence Production Program for Aphasia.