Mapping Aphasia Profiles to AI and Human Treatment

Hyunsoo Yoo, PhD.

Department of Communication Science and Disorders, Baylor University, Waco, TX, US

INTRODUCTION

- As artificial intelligence (AI) systems become increasingly relevant to clinical decision-making, the need for structured, clinician-informed datasets is critical.
- In aphasia rehabilitation, treatment selection often varies among experts, reflecting diverse theoretical orientations and training backgrounds.
- Capturing this variation in a systematic and analyzable format is a necessary step toward transparent and trustworthy AI-based recommendation tools.
- This pilot study investigates **1) agreement between two expert clinicians in aphasia treatment selection** and explores **2) how a rule-based AI system aligns with their decisions**.

METHOD

- Aphasia Profiles:
  - Six aphasia profiles were constructed from anonymized WAB-R (Western Aphasia Battery-Revised) [1] scores representing a range of subtypes and severity levels. WAB-R profile data were obtained from AphasiaBank [2].
- Clinician Input:
  - Two clinicians independently selected treatment approaches from a predefined checklist and provided brief rationales.
- Agreement Metrics:
  - Agreement was assessed using Jaccard similarity [3], percent agreement, and Cohen's Kappa [4].
  - Jaccard and percent agreement were computed using custom Python scripts, while Cohen's Kappa was calculated using the cohen_kappa_score function from scikit-learn library.
- Analyses Pipelines:
  - All analyses were conducted in Python (v3.10).
- A rule-based AI engine was developed by mapping WAB-R profile features (e.g., AQ, Naming, Comprehension, Repetition) to expert-informed treatment rules [5].

METHOD

- These rules were derived from clinician patterns and expanded to address mid-functioning profiles lacking consistent expert agreement.
- AI-generated recommendations were compared to clinician selections using Jaccard similarity.

Table 1. Rule-Based AI Treatment Logic

Condition (based on WAB-R scores)	Recommended 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 treatment strategies to WAB-R profile features, such as AQ, Naming, Comprehension, and Repetition.*

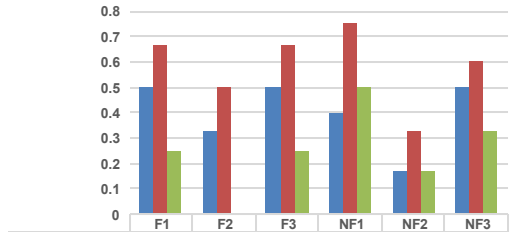
RESULTS

- Clinician agreement varied across participants.
- **Jaccard similarity** between the two clinicians ranged from 0.17 to 0.5, with a mean of 0.40, indicating moderate overlap in treatment selections.

RESULTS

- **Agreement** between AI and Clinician A was generally higher (mean = 0.59), while agreement between AI and Clinician B was lower overall (mean = 0.25).
  - Notably, there was complete divergence (Jaccard = 0) between AI and Clinician B for one participant, underscoring directional differences in treatment selection strategies.
- **Percent agreement** between the two clinicians ranged from 61.54% to 84.62%, with a mean of 79.49%, suggesting relatively high item-level consistency, likely driven by non-selection of shared irrelevant items.
- **Cohen's Kappa** was 0.31, reflecting **moderate reliability** after accounting for chance agreement.
- Greater consistency across measures was observed in profiles with **higher AQ and Repetition scores**.

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



	F1	F2	F3	NF1	NF2	NF3
Jaccard A vs B	0.5	0.33	0.5	0.4	0.17	0.5
Jaccard AI vs A	0.67	0.5	0.67	0.75	0.33	0.6
Jaccard AI vs B	0.25	0	0.25	0.5	0.17	0.33

CONCLUSION

- These findings show that **inter-clinician variability** reflects meaningful differences in clinical reasoning that can inform AI design.
- By capturing these patterns, the proposed framework provides a transparent foundation for AI systems that support, not replace, expert judgment.
- This pilot serves as a proof-of-concept for clinician-informed treatment recommendation models and offers a scalable structure for future learning-based systems.