P-56

A Hybrid Rule–LLM AI System for Electrodiagnostic Reporting: a Two-site Retrospective Evaluation

Yesung Jung[1], Min Kyun Sohn[1,2], Ja Young Choi[1], Kang Hee Cho[1,2*]

1 Department of Rehabilitation Medicine, Chungnam National University College of Medicine, Daejeon, Korea

2 Department of Biomedical Institute, Chungnam National University, Daejeon, Korea

Objective

❖ Evaluate a hybrid rule-based extractor + label-constrained LLM to draft standardized NCS/EMG reports from structured measurements across primary-care and tertiary referral settings.

Methods

❖ **Retrospective cohorts: Primary-care (Apr–Jun 2025) and Tertiary referral (Sep 2023–Feb 2024)**

❖ **Exclusion criteria**

1) CNS lesion (60), 2) normal study (82), 3) non-diagnostic (21), 4) out-of-scope (5)

❖ **AI system pipeline**

1. Parse to UI
Tables to unified schema; clinical note entered separately

→

2. NCS Autofill
Compare with server references; flag segment-level abnormalities

↓

3. Generate Summary
Integrate NCS & EMG abnormalities; apply rule-based criteria

←

4. Generate Diagnostic Interpretation
Prompted LLM produces Diagnostic interpretation

❖ **Metrics**

1) Summary recall (human vs rule)

2) CLI (Coverage/Localization) agreement

Results 1

Cohort flow

	Initial	Included	Excluded	CNS	Normal	Non-diagnostic	Out-of-scope
P	295	241	54	2	39	13	0
T	205	91	114	58	43	8	5
Total	500	332	168	60	82	21	5

Top diagnostic labels

Label	Cases	Share of all labels (%)	Prevalence per case (%)
Median-Wrist	134	33.7	40.4
L-radiculopathy	107	26.9	32.2
PPN	42	10.6	12.7
Ulnar-Elbow	39	9.8	11.7
C-radiculopathy	23	5.8	6.9
BPI	6	1.5	1.8
LFCN neuropathy	4	1.0	1.2
Superficial peroneal	4	1.0	1.2
Common peroneal	3	0.8	0.9
Median-Elbow	3	0.8	0.9

Results 2

Summary recall (%)

	P (n=241)	T (n=91)	Overall (n=332)
NCS recall (Human)	99.5	91.3	96.9
NCS recall (Rule Engine)	100.0	100.0	100.0
EMG recall (Human)	97.6	98.8	97.9
EMG recall (Rule Engine)	100.0	100.0	100.0

Diagn

	Coverage	Localization	Total
P (n=241)	98.7 (97.4–99.8)	97.4 (95.7–98.8)	98.3 (96.9–99.4)
T (n=91)	93.8 (89.0–97.6)	90.0 (85.2–94.3)	92.6 (88.1–96.5)
Overall (n=332)	97.3 (95.7–98.7)	95.4 (93.5–97.0)	96.7 (95.1–98.1)

Performance by examination complexity

Complexity tertiles by total abnormal count
Cutoffs: Low ≤4, Mid 5–8, High ≥9.

Complexity	n	Total abnormal count, mean (min-max)	Human omission rate (%)	Mean Total agreement (%)
Low	141	2.74 (1–4)	5.7	98.8
Mid	100	6.47 (5–8)	6.0	97.4
High	91	16.23 (9–44)	31.9	92.8

Conclusions

❖ Rule-based extraction achieved 100% recall for rule-defined abnormalities; human summaries omitted findings more in high-complexity exams.

❖ Label-constrained LLM produced high agreement with clinicians (Overall Total CLI 96.7%; Primary 98.3%, Tertiary 92.6%) with no safety-critical errors; Agreement decreased in complex / multi-label cases.

❖ Prospective validation on consecutive routine reports and workflow impact is needed.