RWD90

Extracting Patient-Centric Disease Progression Insights From NorstellaLinQ Clinical Notes: MG-ADL Scores in Myasthenia Gravis Using Unstructured Real-World Data

Becky Hollenberg, MPH[1]; Ilan Behm, MPH[1]; Dan Park, PhD[1]; Rahul K. Das, PhD[1]; Allison Perry, PhD[1]
[1]Norstella, One Pennsylvania Plaza, New York, NY 10119 USA

Background

Myasthenia gravis (MG) is a chronic autoimmune neuromuscular disease with fluctuating, fatigable weakness and meaningful variability in functional burden. The MG Activities of Daily Living (MG-ADL) scale is a validated, patient-centric 8-item instrument (total 0–24; higher scores indicate greater impairment),[1] and a ≥2-point change is considered clinically meaningful.[2]

MG-ADL is rarely captured in structured EHR fields, so longitudinal functional status is largely unavailable to claims and structured-data studies. MG-ADL is, however, recorded in physician narratives, neurology consults, and visit summaries within NorstellaLinQ Clinical Notes.

Objective

To demonstrate that LLM-based extraction of MG-ADL scores from NorstellaLinQ Clinical Notes, with human-in-the-loop validation, can systematically characterize longitudinal disease trajectories from routine clinical documentation and enable quantitative real-world outcomes research.

Methods

Data Source
NorstellaLinQ Clinical Notes: de-identified US EHR narrative documentation, including physician narratives, neurology consult notes, and visit summaries. Note date range: August 2015 – October 2026. NorstellaLinQ open claims and structured EHR from January 2017 - October 2026 were used for diagnosis validation.

Population
Patients with MG identified from clinical notes using an LLM with human-in-the-loop validation (n=5,477); 99.5% (n=5,451) also had a structured ICD-10-CM G70.0x diagnosis.

MG-ADL Extraction
- Analytic records: numeric values only (268 non-numeric strings excluded) and totals restricted to 0–24 (289 records >24 excluded); when several totals fell on the same date, the most recent extraction was retained.

Longitudinal Definitions
Baseline: first documented MG-ADL total. **Timepoints:** distinct note dates with a total score. **Follow-up:** first to last total-score date. **12-month score:** latest total 270–450 days after baseline.

Trajectory Classification
Each patient's totals were ordered chronologically and an ordinary least-squares slope of score on days from baseline was fitted across all assessments, scaled to points per year. Rules (applied in order):

Class		Rule	Interpretation
Single timepoint		<2 dated totals	Not classifiable
Sustained improvement		Slope ≤ -2 points/year	Consistent decrease in MG-ADL
Progressive worsening		Slope ≥ +2 points/year	Consistent increase in MG-ADL
Relapsing-remitting pattern		|Slope| <2/year and score range ≥4	Fluctuation without net trend
Stable / minimal change		|Slope| <2/year and range <4	No meaningful change

Analysis
Descriptive statistics: counts and percentages; mean (SD) and median for baseline MG-ADL. The 2-point slope threshold is anchored on the MG-ADL clinically meaningful change.[2] No adjustment for treatment.

LLM Extraction & Validation
Data were extracted from clinical notes by an open-weight medical large language model (LLM) with human-in-the-loop prompt development: prompts were refined through iterative review of sampled outputs against the source notes before application to the full corpus, and extracted values were harmonized to standardized categories. Extraction quality was assessed on 300 clinical notes by an artificial intelligence (AI) evaluation panel of three LLMs from different model families. Each model independently assessed each extracted value against the corresponding note text and the extraction specification. Because no single reviewer, whether human or AI, is error-free, the majority verdict of the panel served as the reference. In prior evaluations, this AI evaluation approach showed approximately 85–95% agreement with clinician reviewers. Precision, recall, and F1 score were calculated.

Results

5,477 Patients with LLM-identified MG

91.5% Had a documented MG-ADL total score (n=5,011)

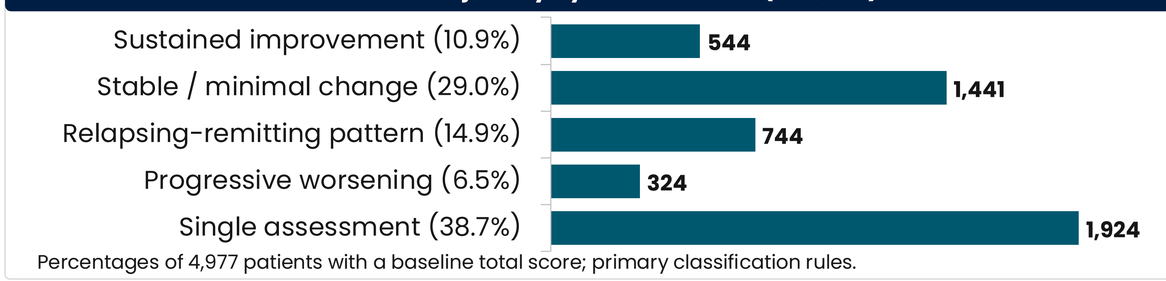
16,937 Unique clinical notes with MG-ADL scores

61.3% Of patients with totals had ≥2 timepoints (n=3,073)

Patient Flow

5,477	Patients with LLM-identified MG
5,448	≥1 MG-ADL score (total or domain) extracted (99.5%)
5,011	≥1 documented MG-ADL total score (91.5%)
3,073	≥2 longitudinal timepoints (61.3% of 5,011); mean follow-up 2.3 years

Disease Trajectory by MG-ADL Total (n=4,977)

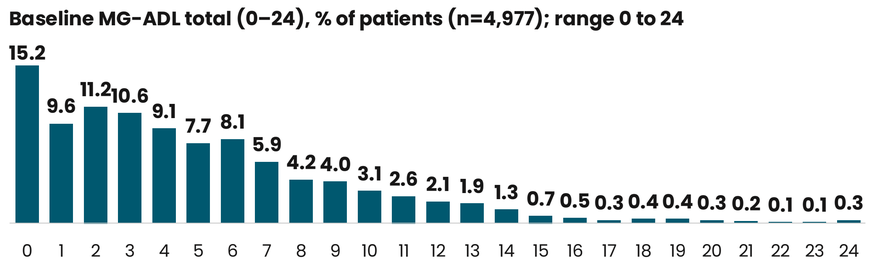


Percentages of 4,977 patients with a baseline total score; primary classification rules.

Baseline MG-ADL Total Score

4 Median
4.9 Mean
4.4 SD

Baseline MG-ADL total (0–24), % of patients (n=4,977); range 0 to 24



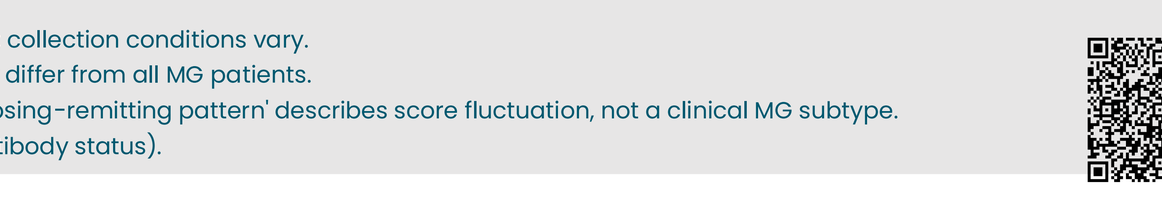
12-Month Change and Domain Scores

Measure	Result
Patients with 12-month score (270–450 days)	1,102
Mean (SD) MG-ADL at 12 months	3.9 (3.8)
Mean change from baseline	-0.9
Patients with ≥2-point improvement	367 (33.3%)

Domain scores (SQL G4): 729 distinct item labels extracted (0–3 scale). Six most frequent, mean item score (n patients): eyelid droop 1.02 (2,705); double vision 0.82 (3,002); breathing 0.54 (3,166); swallowing 0.49 (3,271); talking 0.43 (3,214); chewing 0.41 (3,271).

MG-ADL Over Time by Trajectory Class

Mean MG-ADL total by 6-month interval from baseline (months), by trajectory class



n patients per interval (0–6 to 30–36 months): Improvement 544, 241, 133, 64, 41, 27; Stable 1,441, 522, 438, 367, 290, 203; Relapsing-remitting 744, 347, 342, 305, 296, 276; Progressive worsening 324, 133, 55, 38, 32, 11.

Extraction Validation and Documentation Density

HITL accuracy vs clinician review
Precision 99.4% | Recall 96.2%

MG-ADL notes per patient
Mean (SD) 3.1 (2.9) | Median 2 | Range 1–31

Sensitivity rules (SQL G3)
≥3 timepoints for fluctuation; ≥90-day span: sustained improvement 9.1%, stable 26.9%, relapsing-remitting 14.5%, progressive worsening 5.4% (of 4,977)

LLM Extraction Quality vs AI Evaluation Panel (300 Notes)

Relative to the panel reference, MG-ADL score extraction achieved a precision of 99.7% (99.4%–99.9%), a recall of 97.5% (96.6%–98.2%), and an F1 score of 98.6% (98.1%–99.0%).

Table 1. Agreement (%, 95% CI) between LLM Extraction and a three-model AI Evaluation Panel (Majority Vote).

Field	n (Notes)	Precision	Recall	F1
MG-ADL Score	300	99.7 (99.4-99.9)	97.5 (96.6-98.2)	98.6 (98.1-99.0)

Conclusion

1. LLM-assisted extraction with human-in-the-loop validation captured MG-ADL for 99.5% of LLM-identified MG patients, with a documented total score for 91.5%, across 16,937 unique notes.
2. Longitudinal assessment was available for 61.3% of patients with total scores; based on the trend across all documented assessments, 10.9% showed sustained improvement, 29.0% remained stable, 14.9% showed a relapsing-remitting pattern, and 6.5% demonstrated progressive worsening.
3. These findings demonstrate the feasibility of generating patient-reported outcome measures from routine clinical care, expanding opportunities for longitudinal real-world MG studies; similar approaches may apply to other neurological diseases with clinically documented functional assessments.

Key Limitations

1. MG and MG-ADL are identified from notes; patients whose MG-ADL is not documented, or who receive care outside contributing health systems, are not captured, enriching the cohort for neurology-followed patients.
2. MG-ADL recorded in notes may be clinician-administered or transcribed rather than patient-completed; collection conditions vary.
3. 38.7% had a single assessment and could not be classified; trajectories among classifiable patients may differ from all MG patients.
4. Trajectory rules are heuristic; slopes from few or closely spaced visits are sensitive to single values. 'Relapsing-remitting pattern' describes score fluctuation, not a clinical MG subtype.
5. Descriptive analysis without adjustment for treatment or disease subtype (e.g., ocular vs generalized, antibody status).

References
1. Wolfe GI, Herbelin L, Nations SP, et al. Myasthenia gravis activities of daily living profile. Neurology. 1999;52(7):1487-1489.
2. Muppidi S, Wolfe GI, Conaway M, Burns TM; MG Composite and MG-QOL15 Study Group. MG-ADL: still a relevant outcome measure. Muscle Nerve. 2011;44(5):727-731.

This research was sponsored by Norstella, which all authors were employed by at the time of conduct. The authors declare no conflict of interest in connection with the presented work.