Poster MSR213

Use of a Small Language Model to Identify MG-ADL Scores from Encounter Notes in an EMR System

Scan here for e-poster

Ravindra Telidevara[1], Neisha Opper[2], Ishtiyaque Ahmad[1], Vivek Rudrapatna[3], Trinabh Gupta[1], Shivani Aggarwal[2]

[1]DataUnite, Inc., Cupertino, CA, USA, [2]Landmark Science, Inc., Los Angeles, CA, USA [3]University of California San Francisco, CA, USA

Background & Rationale

- Generalized myasthenia gravis (gMG) is a chronic autoimmune neuromuscular disorder characterized by fluctuating muscle weakness impacting daily functioning.
- Myasthenia Gravis Activities of Daily Living (MG-ADL) scores are frequently used in clinical trials as a key indicator of functional status for gMG patients.
- While electronic medical records (EMRs) represent a rich source of real-world data for ADL-related information, the documentation of such information is often limited to unstructured formats which can be time consuming and resource intensive to extract.
- Large language models show promise for extracting unstructured EMR data, but their infrastructure requirements limit feasibility in many healthcare environments. Small language models (SLMs) offer more computational efficiency, but performance evaluation is needed.

Objectives:

- **To characterize the performance of an SLM in extracting available MG-ADL scores from unstructured EMR data**
- **To describe the baseline characteristics of gMG patients with and without extractable MG-ADL scores**
- **To characterize the longitudinality of MG-ADL scores among the cohort of gMG patients with extractable scores**

Methods

Study Design

- We conducted a descriptive analysis to characterize data capture of MG-ADL scores using a retrospective cohort of prevalent gMG patients.
- **Data Source**: Structured and unstructured data from within the University of California San Francisco EMR system, accessed through DataUnite's privacy-preserving platform.
- A computationally-lightweight, CPU-only SLM was developed and deployed to extract available MG-ADL scores.

Study Population & Performance Assessment

Study Population

- Adult patients (≥18 years) with prevalent gMG (ICD-10-CM diagnosis code: G70.x, with earliest as index date) indexed between January 01, 2016 – September 30, 2025.

Performance Assessment

- MG-ADL scores were extracted from encounter notes using the SLM (Figure 2).
- Unstructured MG-ADL data were converted to MG-ADL total and domain-level scores.
- Extracted scores were evaluated.
- A sample of randomly selected notes classified by the SLM as present or absent (n=200) was clinically reviewed.

Analysis

- Demographic characteristics were described among all gMG patients and those with ≥1 MG-ADL score(s) using descriptive statistics.
- Availability and longitudinality of MG-ADL scores were described.
- Performance metrics including sensitivity, specificity, positive and negative predictive value, and F1 score were evaluated for:
  - presence/absence classification;
  - total score value; and
  - domain-level values.

Key Findings

- MG-ADL scores were infrequently documented within the EMR. Among 1,962 gMG patients, none had MG-ADL scores documented in commonly available structured tables. A total of 7.8% (n=153) had ≥1 and 5.9% (n=116) had 2+ documented MG-ADL scores in unstructured data (Figure 1).
- MG-ADL score documentation does not appear to differ by demographic characteristics (Table 1).
- Among those with ≥1 MG-ADL score, a median of 3 MG-ADL scores were documented per patient (Figure 3a) over a median of 21.7 months of follow-up (Table 1). Among those with 2+ MG-ADL scores, a median of 4 MG-ADL scores were documented per patient (Figure 3a).
- Among those with ≥1 MG-ADL score, the median MG-ADL score value was 3 (Figure 3b). 68% of patients had any scores in the mild range (0-4); while 43% had any moderate scores (5-9) and 20% had any severe scores (10+). 42% experienced a clinically meaningful change (2+ points) during follow-up (data not shown).
- The lightweight SLM demonstrated highly discriminative performance for identifying MG-ADL scores and strong concordance for total and domain-level values.
  - Of encounter notes manually reviewed (n=200), the SLM correctly classified 195/200 (Figure 4).
  - Sensitivity, specificity, negative predictive, and positive predictive values ranged from 97-98% (Figure 4).
  - Classification accuracy and F1-score for absence/presence were 97.50% (Figure 4).
  - Among true positives (n=97), the SLM extracted total scores with 100% accuracy and concordance of domain score values ranged from 94.85% for 'brushing teeth/hair' to 98.97% for 'arising from chair' and 'diplopia' (Table 4).
  - Among the subset of encounter notes containing both total and domain-specific scores (N=565), there was 92% concordance between the extracted score and calculated score (data not shown). Reasons for discordance (e.g., accidental extraction of general ADL score) have implications for SML model tuning.

Limitations

- Findings on the availability of MG-ADL scores within this AMC may not be generalizable to other academic or community centers within the US.
- A comprehensive evaluation of existing EPIC flowsheets was not conducted as part of this work.
- This is a descriptive analysis, and no causal conclusions may be drawn from this work.

Why is this Research Important?

- These findings highlight the potential of SLM-assisted approaches for MG-ADL extraction to enable real-world outcomes research in gMG and other therapeutic areas reliant on outcome information stored in unstructured data.
- Increased access to such outcomes, whether facilitated by better documentation in structured data, manual extraction, or NLP-assisted approaches:
  - Supports real-world evidence generation for post-marketing studies, label expansions, and health economics analyses in diverse patient populations.
  - Aligns with regulatory interest in reliable RWE for supplemental indications and comparative effectiveness research.
  - Enhances research relevance by facilitating the incorporation of meaningful patient-centered outcomes related to functional status and quality of life.

LANDMARK SCIENCE

Presented at ISPOR 2026 | May 17–20, 2026 | Philadelphia, Pennsylvania.

DataUnite

Results

Figure 1. Attrition Diagram for Patients Eligible for Inclusion

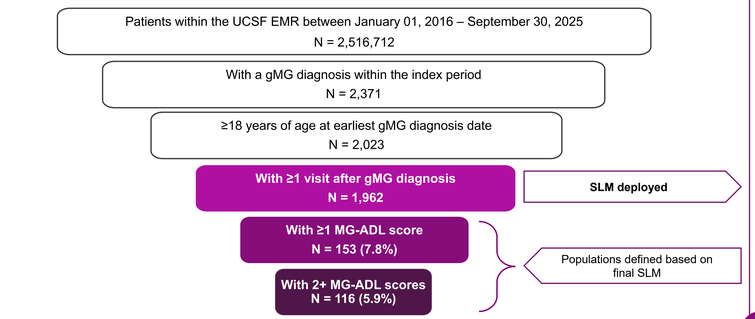


Table 1. Baseline Demographic Characteristics

	gMG patients active between January 01, 2016 – September 30, 2025 N = 1,962	gMG patients with ≥1 MG-ADL Score N = 153
Age at index date, years		
Mean (STD)	59.9 (17.5)	60.1 (16.8)
Median (IQR)	64 (25.8)	63 (25)
Min, Max	18, 89	18, 88
Sex, N (%)		
Male	921 (46.9%)	66 (43.1%)
Female	1,036 (52.8%)	86 (56.2%)
Other/Unknown	5 (0.3%)	1 (0.7%)
Race, N (%)		
White	1,176 (59.9%)	94 (61.4%)
Black	99 (5%)	8 (5.2%)
Asian	256 (13%)	17 (11.1%)
Other/Unknown	431 (22%)	34 (22.2%)
Ethnicity, N (%)		
Hispanic	231 (11.8%)	17 (11.1%)
Non-Hispanic	1,582 (80.6%)	126 (82.4%)
Unknown/Declined	149 (7.6%)	10 (6.6%)
Region, N (%)		
Midwest	3 (0.2%)	1 (0.7%)
Northeast	5 (0.3%)	0
South	13 (0.7%)	0
West	1,937 (98.7%)	151 (98.7%)
Unknown	4 (0.2%)	1 (0.7%)
Follow-Up Time, months		
Mean (STD)	34.6 (30.0)	33.3 (27.8)
Median (IQR)	25.1 (47.9)	21.7 (34.3)
Min, Max	0.03, 109.3	1.16, 108.0

Figure 2. Distribution of the functional burden of myasthenia gravis across the follow-up period among patients with ≥1 MG-ADL Score (N=153)

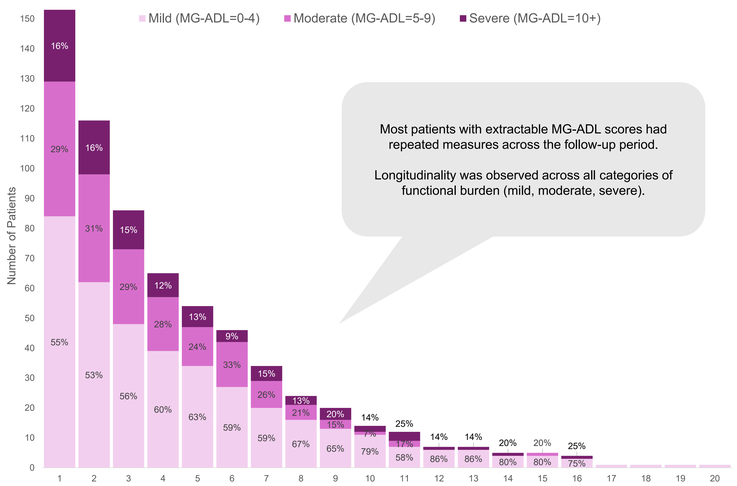
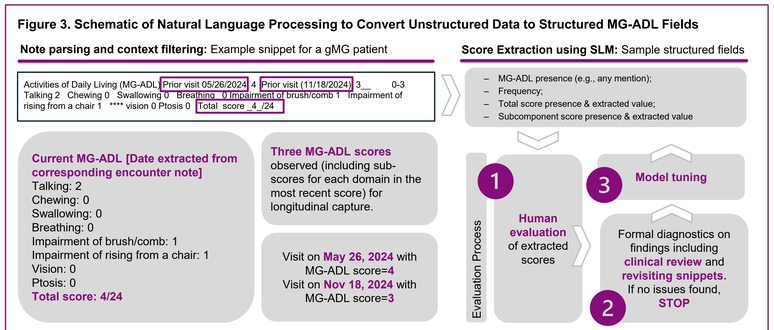


Figure 3. Schematic of Natural Language Processing to Convert Unstructured Data to Structured MG-ADL Fields



1 Figure 4. Evaluation: SLM Performance on Presence of Total MG-ADL Score (N = 200 Notes*)

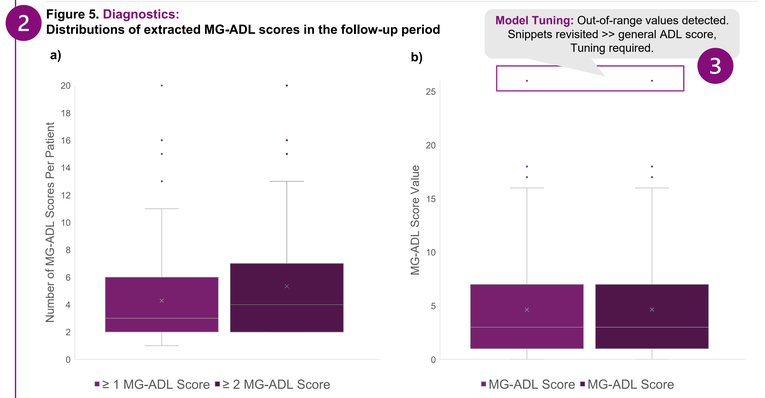
		Manual Review: Present	Manual Review: Absent	
				F1 Score = 97.49%
SLM Review	Present	97	3	PPV/Precision = 97.00%
SLM Review	Absent	2	98	NPV = 97.98%
		Sensitivity/Recall = 97.98%	Specificity = 97.03%	Accuracy = 97.50%

*200 SLM-classified notes were randomly selected for manual review; 100 classified as MG-ADL present, 100 classified as MG-ADL absent.

Table 2. SLM Performance on Total MG-ADL & Sub-score Value (N = 97 True Positives)

	Accuracy
Total score	100%
Domain-level sub-score	
Talking	96.91%
Chewing	96.91%
Swallowing	96.91%
Breathing	96.91%
Brushing teeth/hair	94.85%
Rising from chair	98.97%
Diplopia	98.97%
Ptosis	97.94%

2 Figure 5. Diagnostics: Distributions of extracted MG-ADL scores in the follow-up period



Total Number of MG-ADL Scores Extracted

Patients with ≥1 MG-ADL Score (N = 153)		
Scores per Patient	Mean (STD)	4.29 (3.73)
	Median (IQR)	3 (4)
	Min, Max	1, 20
Patients with ≥2 MG-ADL Score (N = 116)		
Scores per Patient	Mean (STD)	5.34 (3.71)
	Median (IQR)	4 (5)
	Min, Max	2, 20

Total MG-ADL Score Value (As Extracted)

Patients with ≥1 MG-ADL Score (N = 153)		
Scores Across All Patients N = 656	Mean (STD)	4.63 (4.65)
	Median (IQR)	3 (6)
	Min, Max	0, 26
Patients with ≥2 MG-ADL Score (N = 116)		
Scores Across All Patients N = 619	Mean (STD)	4.65 (4.67)
	Median (IQR)	3 (6)
	Min, Max	0, 26

References: References are available upon request to the corresponding author at Shivani@landmarkscience.com;
Funding: No external funding was received for this study; **Disclosures:** NO and SA are employees or contractors of Landmark Science, Inc.; RT, IA, and TG are employees or contractors of DataUnite, Inc.