RWD102

Improving Real-World Identification and Molecular Characterization of Diffuse Midline Glioma Using Human-in-the-Loop LLM Extraction of NorstellaLinQ Unstructured Electronic Health Record (EHR) Clinical Notes

Bharath P[1], Eric Mitchell[1], Ilan Behm, MPH[1], Allison Perry, PhD, MHS[1]
[1]Norstella, One Pennsylvania Plaza, New York, NY 10119 USA

Background

- Diffuse midline glioma (DMG) is an H3 K27-altered, WHO Grade IV central nervous system tumor.
- It is an ultra-rare malignancy with a median overall survival of approximately 9–13 months.
- Structured diagnosis coding incompletely captures DMG in real-world data because of:

Evolving molecular nomenclature
e.g., H3K27M, H3K27-altered

Legacy terminology
e.g., diffuse intrinsic pontine glioma (DIPG)

Anatomical descriptors
Disease-specific site descriptions in free text

Result
Patients missed or misclassified in claims and structured EHR data

Objective

To evaluate whether human-in-the-loop (HITL) large language model (LLM) extraction of unstructured clinical notes can improve real-world identification and molecular characterization of patients with DMG.

Methods

Study Design	Retrospective observational study.
Data Source	NorstellaLinQ US unstructured electronic health record (EHR) clinical notes, 2016–2026.
Inclusion	Patients with confirmed DMG-related documentation.
Molecular Status	Extracted from free text: H3K27M-positive, H3K27M-negative, indeterminate, wildtype, and tested only.

Human-in-the-Loop LLM Extraction Workflow

1 **Unstructured notes** — NorstellaLinQ US EHR, 2016–2026
2 **Terminology framework** — Molecular, historical, WHO, anatomical
3 **HITL LLM extraction** — Concept normalization
4 **DMG cohort** — Identification and molecular status

Terminology Framework

A comprehensive framework developed for HITL LLM extraction and concept normalization:

Molecular markers
H3K27M, H3K27-altered, H3-wildtype

Historical disease terminology
e.g., DIPG

WHO classification
CNS tumor classification terms

Anatomical descriptors
Midline and disease-specific sites

Results

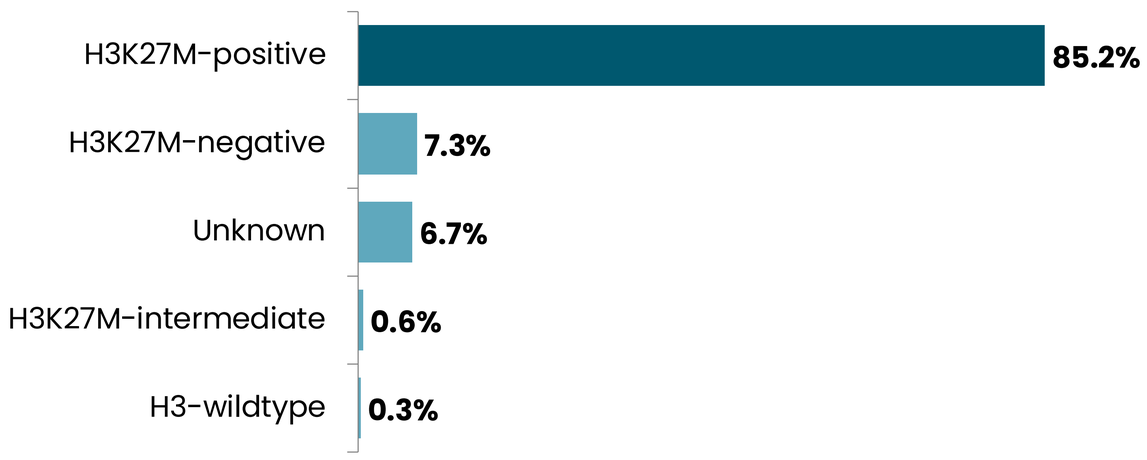
2,513
unique patients with documented DMG

85.2%
documented H3K27M-positive (n = 2,141)

7.3%
documented H3K27M-negative (n = 183)

Documented Molecular Status Among Patients with DMG

Percent of the 2,513 identified patients



Patient Counts by Molecular Status

Percentages are of the 2,513 identified patients

Molecular status	n	%
H3K27M-positive	**2,141**	**85.2%**
H3K27M-negative	183	7.3%
H3K27M-intermediate	14	0.6%
H3-wildtype	7	0.3%
Unknown	168	6.7%
Tested, no definitive classification	***102***	***4.1%***

Key Findings

- HITL LLM extraction identified 2,513 unique patients with documented DMG, representing substantial capture of an ultra-rare malignancy within a national real-world database.
- Most patients (85.2%) had documentation consistent with H3K27M-positive disease.
- An additional 102 patients had documented H3K27M testing without a definitive molecular classification.
- Unstructured notes contain clinically relevant molecular, historical, and anatomical information not routinely available through structured diagnosis coding alone.

Conclusions

1. HITL LLM extraction of unstructured clinical notes enables scalable identification and molecular characterization of patients with DMG beyond what is typically available in structured EHR data.
2. Incorporating unstructured documentation into real-world evidence workflows may improve:

Rare disease cohort construction | **Clinical trial feasibility**
Epidemiologic studies | **Treatment pattern research**

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.