Utilizing NLP to Enhance EHR Data Abstraction Accuracy in Waldenström Macroglobulinemia

Alisha Monnette Kimble, Debra Rembert, Avi Raju, Gayathri Namasivayam, Junxin Shi, Wanmei Ou, Robert Reid

Author Affiliations: Ontada, Boston, MA, USA

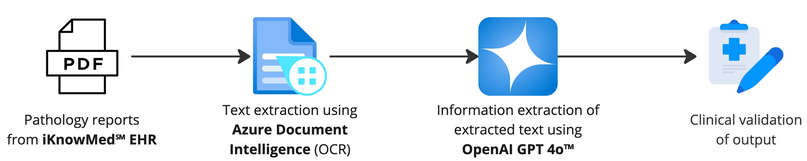
Background

- Accurate cancer diagnosis data is critical for research but may be incomplete when diagnoses occur outside the healthcare system or are incompletely referenced in the EHR.
- To address these gaps, natural language processing (NLP) was applied to extract confirmed Waldenström macroglobulinemia (WM) diagnoses and diagnosis dates from unstructured pathology reports
- This approach aimed to improve data completeness for WM—a rare B-cell neoplasm—and help distinguish true WM cases from other hematologic malignancies or second primaries, which often share overlapping terminology.

Methods

- We developed an NLP pipeline to extract diagnosis names and dates from pathology reports in iKnowMed, an EHR used by US Oncology clinics.
- Azure Document Intelligence was used for optical character recognition for text extraction and OpenAI's GPT-4 for entity recognition and information retrieval.
- On a development set, the model achieved F1 scores of 0.87 for diagnosis name and 0.86 for diagnosis date, indicating strong performance in entity recognition and classification.
- Patients with related but non-specific terms (e.g., lymphoma, myeloma) were reviewed to confirm WM as the primary diagnosis.
- NLP output was validated by clinicians against structured data to ensure accuracy.

Figure 1: NLP pipeline workflow



Acknowledgements

The authors thank the investigators, study team, and the patients and their families at each of the practices.

Correspondence: Alisha Monnette Kimble
alisha.monnette@mckesson.com

Results

- 3,498 patients with structured WM diagnosis identified (2014–2022)
  - 880 (25%) patients had missing structured diagnosis dates
  - 514 (15%) patients had related terms or a possible second primary

Patients with missing diagnosis dates (n=880) (Figure 2)

- 485 (55%) excluded by NLP due to no WM-specific pathology report found
  - *Eliminated need for abstraction and clinician validation*
- 395 had WM-specific pathology terms and were reviewed
  - 284 (72%) confirmed by clinician as valid WM diagnosis and date
  - 111 (28%) excluded: (89 diagnosis date outside study period; 22 vague B-cell lymphoma term (no WM confirmation))

Figure 2: Attrition table for WM diagnosis date verification

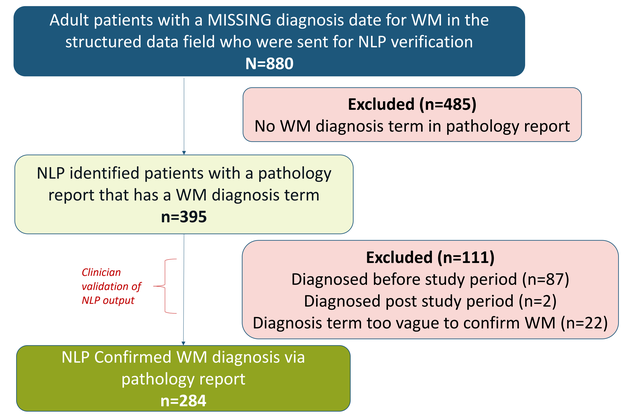
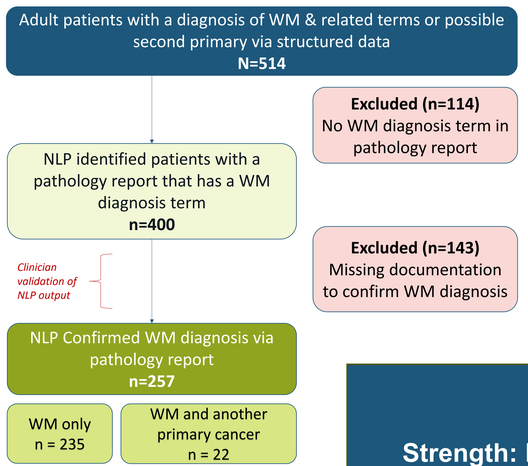


Figure 3: Attrition table for NLP WM and another primary confirmation



Patients with related terms or possible second primary (n=514) (Figure 3)

- 114 (22%) excluded due to absence of WM-specific terms in pathology reports
- 400 proceeded to clinician review
  - 257 (64%) confirmed WM: 235 (91%) had WM only and 22 (9%) had WM and second primary
  - 143 (36%) excluded due to lack of confirmed WM

Efficiency Gains

- NLP excluded 599 patients reducing clinician abstraction workload
- NLP reduced total abstraction timeline from ~6 months (manual-only) to ~1 month
- NLP allowed for parallel review and exclusion, achieving a 5× acceleration in data processing and cohort readiness

Strengths and Limitations

Strength: NLP efficiently extracted diagnostic information from unstructured pathology reports, reducing manual abstraction time and enabling scalable data capture.

Limitation: Clinician review was still necessary to validate NLP output, as the model could not always definitively confirm diagnosis or date. NLP serves as a complement—not a replacement—for manual review in complex oncology data.

Conclusions

- NLP improved capture of WM diagnoses and dates by addressing gaps in structured EHR data and reducing reliance on manual abstraction. This approach streamlined data collection, shortened timelines, and improved overall data quality.
- Findings demonstrate NLP's potential to strengthen real-world data infrastructure and accelerate large-scale oncology research.