# Large Language Model Extraction of Structured Data in Unstructured Sarcoma Pathology Notes: A Pilot Study

Marisa Ulrich, Linjun Yang, Amirali Khosravi, Srikar Namireddy, Miguel M Girod, Sami Saniei, Monty Khela, Cody Wyles, Matthew T Houdek

INTRODUCTION:
Bone and soft-tissue sarcomas are rare, heterogeneous tumors with unique treatment challenges. This heterogeneity has led to difficulty with registry curation, as well as effective outcomes assessment. Pathologic diagnosis guides essential treatment decisions and surveillance. Pathology notes contain unstructured data, requiring manual data abstraction for data tracking or attempted registry creation. Large language models (LLM) have potential to extract structured results from unstructured inputs. Thus, the purpose of this study was to assess the feasibility of LLMs in extracting key information from surgical pathology notes following surgical sarcoma excision.
METHODS:
Surgical pathology notes were retrospectively identified from 353 patients who received surgical intervention for a diagnosis of bone or soft tissue sarcoma at one institution from 2000-2024. Data points of interest included tumor diagnosis (osteosarcoma, Ewing's sarcoma, etc), greatest dimension (cm), closest margin (cm), and grade. Three human observers manually annotated all notes for each data point of interest. All labeled pathology notes were then split into test and training sets in an 85/15 split (303 testing notes, 50 training notes). Data was extracted through customized pipelines built with Llama-3-8B-Instruct (Meta), with the optimal prompting approach (one-shot, few-shot, or chain of thought reasoning) selected for each data point. Prompts were iteratively fine-tuned using the training notes to teach the LLM background information and data extraction methodologies. The accuracy of developed pipelines was evaluated by comparing the LLM predictions with the manual labels (ground truths) on the held-out testing notes.
RESULTS:
The LLM-enabled data extraction pipelines demonstrated excellent accuracy for diagnosis and grade, and moderate accuracy for closest margin and greatest dimension, reflecting data point complexity. The following accuracies were achieved: diagnosis (98%), greatest dimension (88%), closest margin (85%), and grade (94%). The mean time of data extraction for each data point was 4 seconds.
DISCUSSION AND CONCLUSION:
This pilot study demonstrates a critical proof-of-concept for automated registry production utilizing LLM outputs of unstructured pathology note data. The data extraction pipeline utilized an open-source model (critical for medical data safety) to demonstrate satisfactory accuracy in extracting key data elements. This approach holds promise to augment the efficiency and accuracy of human data abstraction for registry production, which can be leveraged for optimal risk stratification and outcomes assessments in future endeavors.