Improving Clinical Data Capture From Free-Text Operative Notes for Surgical Approach in Total Hip Arthroplasty Using a Deep Learning Model

Lefko Theo Charalambous[1], Kush Attal[2], Catherine Di Gangi, Daniel J Berry[3], David G Lewallen[3], Joshua Craig Rozell
[1]NYU Langone Orthopedic Hospital, [2]NYU Grossman School of Medicine, [3]Mayo Clinic

INTRODUCTION: Annotating free-text clinical notes into structured data is critical for future large-scale data analysis in healthcare. Toward this goal, machine learning algorithms have been developed but are limited because they generally require (1) specific, context-independent training, (2) large datasets, and (3) do not give clinical justification for their classifications. Novel deep-learning large language models (LLMs) present an opportunity to prompt LLMs using few highly-specific expert-level examples and deploy them over numerous unstructured texts with high-fidelity classification. Annotating surgical approach in total hip arthroplasty (THA) has been a difficult data point to capture and validate using standing reporting from electronic health system reports. In this study, we developed a novel algorithm using OpenAI's GPT4 to capture and justify surgical approach.

METHODS: Using few-shot learning, GPT4 was prompted with 13 examples of "gold-standard" operative notes describing anterior, anterolateral, posterior, and lateral THA. Notes for 120 randomly-selected primary THAs performed at a single institution between April 2012 and February 2024 by 22 surgeons were collected. GPT4 then classified bearing used for each annotation with associated clinical justification.

RESULTS:

GPT4 classified surgical approach with an overall accuracy of 97.5% (Table 1). Regarding anterior and anterolateral THA, precision, recall, and f1 score were 100% for all. Regarding lateral THA, precision, recall, and f1 score were 90.9%, 100%, and 95.2%, respectively. For posterior THA, precision, recall, and f1 score were 100%, 90%, and 94.5%. The model also provided clinical justifications for every classification (Table 2) with an average Flesch-Kincaid Grade Level (FKGL) score of 24.42 (SD: 7.32) and average selfBLEU score, measuring diversity of logic, of $9.76*10^{-232}$ (SD: <0.01) (Table 3). The justifications had a character-level sequence match of 96.47% (SD: 10.64%) with direct citations from the original notes.

DISCUSSION AND CONCLUSION: GPT4 captured the surgical approach documented in THA operative notes with high fidelity while providing clinical justification with text annotation. Clinical justification often highlighted detailed anatomy logic. Additionally, the low SelfBLEU score indicates substantial diversity across clinical reasoning, indicating the model utilized distinct logic to classify approach across a broad array of surgeons. Deep-learning LLMs are a potential alternative to manual chart review in data capture and require fewer examples and offer more clinical context compared to traditional rule-based and machine learning algorithms. The current study validates the use of deep learning to not only extract data but also provide clinical justification in a quick, efficient method. Future applications of this model can be applied to larger datasets to improve clinical data capture.

	precision	recall	f1-score
ANTERIOR	1.000	1.000	1.000
ANTEROLATERAL	1.000	1.000	1.000
LATERAL	0.909	1.000	0.952
POSTERIOR	1.000	0.900	0.945
accuracy			0.975
macro avg	0.977	0.975	0.975
weighted avg	0.977	0.975	0.975

Table 1. Performance of the GPT4 model on the test set for data capture of surgical approach documented in THA operative notes

Classification	Clinical Reasoning
Anterolateral	The input contains the phrase **'Anterior and posterior skin flaps were raised for exposure'** which indicates an anterolateral approach was performed. Additionally, the phrase **'Gluteus medius was split in line with its fibers at the junction of the anterior and middle thirds of the muscle belly'** also suggests an anterolateral approach based on the specific anatomy being manipulated.
Anterior	The input contains the phrase **'incision of approximately 8 cm was carried out over the anterior aspect of the hip'** which directly tells me that the surgery was performed in an anterior approach. In addition, the input contains the phrase **'blunt dissection over the medial aspect of the tensor muscle belly'**, which based on human anatomy, involves cutting into muscles located on the front of the human body.

Table 2. Clinical justification created by GPT4 to justify classification of the surgical approach used in specific THA operative notes. Bolded text represents approximate citations from original THA operative note.

Data Type	Flesch-Kincaid Grade Level		SelfBLEU		Similarity Score	
	Average	S.d.	Average	S.d.	Average	S.d.
GPT4 Clinical Reasoning	24.42	7.32	$9.76e{-232}$	< 0.01	96.47	10.64

Table 3. Readability, diversity, and alignment of the clinical justifications of the GPT4 model