Artificial Intelligence Assisted Procedural Coding for Knee and Femur Surgery: Evaluation of the Compact Mistral-Nemo Large Language Model

Michael Li, Sri Guttikonda, Yash Lahoti, Ula N Isleem, Samuel Kang-Wook Cho, Jun Sup Kim

INTRODUCTION:

This study evaluated the performance of the compact-sized Mistral-Nemo large language model in the task of automating the coding process for Current Procedural Terminology (CPT) codes, specifically for femur and knee related procedures. Mistral-Nemo has the potential to reduce administrative burden and improve workflow, while also safeguarding sensitive patient data.

METHODS:

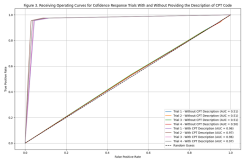
The Mistral-Nemo model was used to evaluate 1,000 operative notes labeled with CPT codes, drawn from 177 orthopedic providers performing femur and knee procedures. The dataset included 46 unique CPT codes, and the most common were knee arthroscopy with medial or lateral meniscectomy (29881), medial and lateral meniscectomy (29880), and ACL reconstruction (29888). For each operative note, two separate prompts were generated with one pairing the note with its correct CPT code, and the other pairing it with a randomly selected incorrect code. For each CPT prompt, the corresponding AMA CPT code description was included in some trials and omitted in others to evaluate the impact of code descriptions on model performance. Trials were conducted in two formats: (1) binary response trials prompting the model for a “Yes” or “No” answer, and (2) confidence score trials prompting the model to provide a score from 0 to 100 indicating its confidence that the CPT code matched the described procedure.

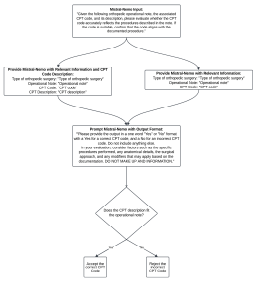
RESULTS:

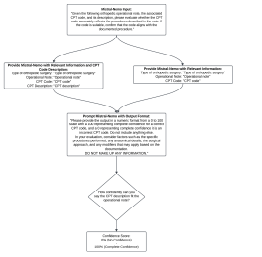
In binary response trials, Mistral-Nemo correctly identified 90% of correct CPT codes (n=1,000) and rejected 99.8% of incorrect codes (n=1,000, $p<0.001$), achieving a precision of 99.8% and a recall of 90%. Performance was further analyzed for the three most common CPT codes, which accounted for 747 of 1,000 operative notes: 29881 (95.6% accuracy, n=430), 29880 (95.5% accuracy, n=160), and 29888 (82.5% accuracy, n=157, all $p<0.001$). In confidence score trials, Mistral-Nemo achieved an AUROC of 0.96–0.97 with descriptions included, and an AUROC of 0.51 without descriptions, indicating strong dependence on structured code definitions. The model processed 111 operative notes per minute, demonstrating high throughput and efficiency.

DISCUSSION AND CONCLUSION:

Mistral-Nemo model’s speed and accuracy in correctly identifying CPT codes in our study offers substantial potential in automating the coding process. It can be further developed to help enhance the efficiency and accuracy of procedural coding in orthopedic surgery.







Trial	With CPT Description		Without CPT Description		Chi-Square Significance Test
	Correct Predictions%	Correct Rejection%	Correct Predictions%	Correct Rejection%	
1	90.30%	99.80%	47.00%	56.80%	$p < .001$
2	90.30%	99.80%	47.00%	56.80%	$p < .001$
3	90.30%	99.70%	47.00%	57.40%	$p < .001$

Table 1: Comparison of Prediction and Rejection Rates With and Without CPT Description