Artificially Intelligent Billing Performance in Hip and Pelvis Operative Procedures: A Comparative Analysis of Compact Large Language Models

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

INTRODUCTION:
This study evaluated the performance of compact large language models (LLMs) in automating the assignment of Current Procedural Terminology (CPT) codes from hip and pelvis surgical operative notes. Compact LLMs differ from conventional large models in that their smaller parameter size enables them to operate securely on local hardware, protecting patient data while maintaining strong language understanding capabilities.
METHODS: Three compact LLMs, DeepSeek-R1, Mistral-Nemo, and LLaMA 3.3, were used to evaluate 925 operative notes labeled with CPT codes from 139 orthopaedic providers performing hip and pelvis procedures. The dataset included 116 unique CPT codes. For each operative note, two types of prompts were generated: one prompt presented the model with the note and the correct CPT code, and a separate prompt presented the same note with a randomly selected incorrect CPT code. Each CPT code prompt also included the corresponding American Medical Association (AMA) CPT code description. The models were tasked with providing a confidence score from 0 to 100, where 0 indicated no confidence and 100 indicated complete confidence that the CPT code correctly matched the described procedure. Model performance was evaluated across three independent trials using area under the receiver operating characteristic curve (AUROC) and Brier score.
RESULTS:
Mistral-Nemo demonstrated the highest performance out of all models, achieving mean AUROC of 0.97–0.98 and Brier scores of 0.040–0.043 across trials. LLaMA 3.3 also performed strongly, with AUROC of 0.933–0.934 and Brier scores of 0.075–0.076. In contrast, DeepSeek-R1 demonstrated near-random discrimination (AUROC 0.48–0.52) and poor calibration (Brier scores 0.28–0.29). In terms of processing speed, models evaluated operative notes at an average rate of 54 cases per minute, highlighting its potential advantage for real-time integration into clinical workflows.
DISCUSSION AND CONCLUSION: Compact LLMs such as Mistral-Nemo and LLaMA 3.3 offer substantial potential for automating CPT coding in orthopedic care. Their ability to operate locally while maintaining high accuracy and calibration could reduce administrative burden, improve efficiency, and support more cost-effective healthcare delivery.