Human vs. Artificial Intelligence: Analyzing CPT Coding Accuracy in Vascular Surgeries

Brandon Madris MD, Keivan Ranjbar MD, Kumudini Myla, Andre Critsinelis MD, Stephanie D. Talutis MD, MPH, Shivani Kumar MD, Payam Salehi MD, PhD
Division of Vascular Surgery, Cardiovascular Center, Tufts Medical Center, Boston, MA.

Background

- Accurate Current Procedural Terminology (CPT) coding is critical for financial stability, legal compliance, and patient care in the U.S. healthcare system. Errors in CPT coding can lead to claim denials, revenue loss, audits, and administrative inefficiencies. Misclassification can also impact patient care by disrupting treatment records and insurance coverage.
- Manual CPT coding remains error-prone, with misclassification rates ranging from 20% to 50% [4], increasing the burden on human coders. Artificial intelligence (AI), particularly large language models (LLMs) like ChatGPT-4 and Perplexity Pro, has emerged as a potential tool for improving coding accuracy and efficiency. However, in these broad use LLMS, challenges remain including over-coding, under-coding and misclassification.
- **Objective:** This study evaluates the performance of ChatGPT-4 and Perplexity Pro in matching CPT codes from vascular surgery cases at Tufts Medical Center, comparing their accuracy to the finance department's CPT coding, which serves as the reference standard for this project.

Methods

Study Design & Data

- Compared **ChatGPT-4** and **Perplexity AI** for CPT coding accuracy.
- Analyzed **120 vascular surgery cases** from April 2024.
- Finance department CPT codes used as the reference standard.

Evaluation Criteria

- AI-generated codes classified as **Exact Match, Partial Match, or No Match**.
- **Two input formats:** Full operative notes & brief summaries.

Statistical Analysis

- Used **SPSS v29** for **Crosstabs & Cohen's Kappa** to assess AI agreement.
- Calculated **match rates, partial matches, and non-matches** for each AI system.

Ethical Compliance

- Used **de-identified** patient data; adhered to privacy standards.

Results

- **Accuracy Variation:**
- ChatGPT over-reported CPT codes.
- Perplexity AI under-reported in full-note cases but over-reported in brief notes.
- **Full Operative Notes Performance:**
- ***CPT-Level Accuracy:***
  - ChatGPT: 45.2%
  - Perplexity AI: 43.7%
- ***Case-Level Accuracy:***
  - ChatGPT: 29.2% exact match, 49.2% partial match, 21.6% no match.
  - Perplexity AI: 39.2% exact match, 33.3% partial match, 27.5% no match.
- **Brief Summaries Performance:**
- ***CPT-Level Accuracy:***
  - ChatGPT: 64.75% (43% increase).
  - Perplexity AI: 60.2% (38% increase).
- ***Case-Level Accuracy:***
  - ChatGPT: 51.6% exact match (77% increase).
  - Perplexity AI: 37.5% exact match (4% decrease).
- **Agreement Analysis (Cohen's Kappa):**
- ***CPT-Level Agreement:***
  - ChatGPT (brief note) vs. Finance: κ = 0.45 (highest agreement).
  - Perplexity AI (brief note) vs. Finance: κ = 0.39.
- ***Case-Level Agreement:***
  - ChatGPT (brief note) vs. Finance: κ = 0.341.
  - Perplexity AI (brief note) vs. Finance: κ = 0.273.

•Key Findings:

- Brief summaries improved accuracy and agreement for both AI models.
- ChatGPT had the highest agreement with the finance department's codes.
- Overall agreement remained fair to moderate, emphasizing the need for human oversight in AI-assisted CPT coding.

Table 1. Comparison of AI models vs. finance department for CPT code accuracy

Match Type		CPT codes Matched (out of 261)	Cases Matched (out of 120)		
			Exact Match	Partial Match	No Match
Full operative notes	ChatGPT and Finance Department	118 (45.2%)	35 (29.2%)	59 (49.2%)	26 (21.6%)
	Perplexity AI and Finance Department	114 (43.7%)	47 (39.2%)	40 (33.3%)	33 (27.5%)
	Both AI Models and Finance Department	88 (33.7%)	29 (24.2%)	33 (27.5%)	20 (16.7%)
Brief operative notes	ChatGPT and Finance Department	169 (64.75%)	62 (51.6%)	41 (34.2%)	17 (14.2%)
	Perplexity AI and Finance Department	157 (60.2%)	45 (37.5%)	58 (48.3%)	17 (14.2%)
	Both AI Models and Finance Department	127 (48.65%)	41 (34.2%)	34 (28.3%)	6 (5%)

Conclusions

- This study demonstrates the potential and limitations of AI-assisted CPT coding in vascular surgery. Both ChatGPT-4 and Perplexity Pro performed better with structured, concise documentation, with ChatGPT showing greater accuracy. However, issues like overreporting, underreporting, and misclassification continue to highlight the need for human oversight. Given that 80% of medical bills contain errors, costing $210 billion annually[10], and hospitals spend $19.7 billion contesting 15% of denied claims[11], AI has the potential to enhance billing accuracy and reduce administrative burdens. Additionally, another survey showed nearly 45% of insured adults receive unexpected bills, 17% experience coverage denials, and 47% report worsening health due to billing issues[12]. Standardizing documentation could improve AI performance in billing efficiency, reduce financial losses, and minimize legal risks, but further research is needed to optimize its implementation and reliability in medical coding.

References

1. Association AM. CPT® Overview and Code Approval Process. American Medical Association. 2024.

2. BellMedEx. Pitfalls of Inaccurate Coding and Billing in Healthcare. BellMedEx. 2023.

3. Invensis. Top 7 Consequences of Incorrect Medical Coding and Billing. Invensis Blog. 2022.

4. Venkatesh KP, Raza MM, Kvedar JC. Automating the overburdened clinical coding system: challenges and next steps. NPJ Digit Med. 2023;6(1):16.

5. Kim JS, Vivas A, Arvind V, Lombardi J, Reidler J, Zuckerman SL, et al. Can Natural Language Processing and Artificial Intelligence Automate The Generation of Billing Codes From Operative Note Dictations? Global Spine J. 2023;13(7):1946-55.

6. Ayub S, Scali ST, Richter J, Huber TS, Beck AW, Fatima J, et al. Financial implications of coding inaccuracies in patients undergoing elective endovascular abdominal aortic aneurysm repair. J Vasc Surg. 2019;69(1):210-8.

7. Takefuji Y. Generative AI for analysis and identification of Medicare improper payments by provider type and HCPC code. Explor Res Clin Soc Pharm. 2023;12:100387.

8. Uppalapati VK, Nag DS. A Comparative Analysis of AI Models in Complex Medical Decision-Making Scenarios: Evaluating ChatGPT, Claude AI, Bard, and Perplexity. Cureus. 2024 Jan 18;16(1):e52485. doi: 10.7759/cureus.52485. PMID: 38371109; PMCID: PMC10874112.

9. Zhu C, Attaluri PK, Wirth PJ, Shaffrey EC, Friedrich JB, Rao VK. Current Applications of Artificial Intelligence in Billing Practices and Clinical Plastic Surgery. Plast Reconstr Surg Glob Open. 2024 Jul 1;12(7):e5939. doi: 10.1097/GOX.0000000000005939. PMID: 38957712; PMCID: PMC11216662.

10. Healthline. 80 percent of hospital bills have errors—Are you being overcharged? [Internet]. 2024. Available from: https://www.healthline.com/health-news/80-percent-hospital-bills-have-errors-are-you-being-overcharged

11. Becker's Hospital Review. Claims denials are costing hospitals nearly $20B per year. [Internet]. 2023. Available from: https://www.beckershospitalreview.com/finance/claims-denials-are-costing-hospitals-nearly-20b-per-year.html

12. Grossi G. Survey Exposes Pervasive Billing Errors, Aggressive Tactics in US Health Insurance. [Internet]. 2024. Available from: https://www.ajmc.com/view/survey-exposes-pervasive-billing-errors-aggressive-tactics-in-us-health-insurance