Journal of Hand Surgery Global Online 7 (2025) 181–185

Contents lists available at ScienceDirect

Journal of Hand Surgery Global Online

journal homepage: www.JHSGO.org

Original Research

Artificial Intelligence in Surgical Coding: Evaluating Large Language Models for Current Procedural Terminology Accuracy in Hand Surgery

Emily L. Isch, MD, [*] Jamie Lee, BA, [†] D. Mitchell Self, MD, [‡] Abhijeet Sambangi, BS, [§] Theodore E. Habarth-Morales, BS, 1LT, [§] John Vaile, BS, [§] EJ Caterson, MD, PhD [‖]

[*] Department of General Surgery, Thomas Jefferson University, Philadelphia, PA
[†] Drexel University College of Medicine, Philadelphia, PA
[‡] Department of Neurosurgery, Thomas Jefferson University and Jefferson Hospital for Neuroscience, Philadelphia, PA
[§] Sidney Kimmel Medical College at Thomas Jefferson University, Philadelphia, PA
[‖] Department of Surgery, Division of Plastic Surgery, Nemours Children's Hospital Wilmington, DE

ARTICLE INFO

Article history:
Received for publication October 28, 2024
Accepted in revised form November 21, 2024
Available online January 9, 2025

Key words:
AI in surgery
ChatGPT
CPT coding
Current Procedural Terminology
Hand surgery efficiency
Large language models

Purpose: The advent of large language models (LLMs) like ChatGPT has introduced notable advancements in various surgical disciplines. These developments have led to an increased interest in the use of LLMs for Current Procedural Terminology (CPT) coding in surgery. With CPT coding being a complex and time-consuming process, often exacerbated by the scarcity of professional coders, there is a pressing need for innovative solutions to enhance coding efficiency and accuracy.
Methods: This observational study evaluated the effectiveness of five publicly available large language models—Perplexity.AI, Bard, BingAI, ChatGPT 3.5, and ChatGPT 4.0—in accurately identifying CPT codes for hand surgery procedures. A consistent query format was employed to test each model, ensuring the inclusion of detailed procedure components where necessary. The responses were classified as correct, partially correct, or incorrect based on their alignment with established CPT coding for the specified procedures.
Results: In the evaluation of artificial intelligence (AI) model performance on simple procedures, Perplexity.AI achieved the highest number of correct outcomes (15), followed by Bard and Bing AI (14 each). ChatGPT 4 and ChatGPT 3.5 yielded 8 and 7 correct outcomes, respectively. For complex procedures, Perplexity.AI and Bard each had three correct outcomes, whereas ChatGPT models had none. Bing AI had the highest number of partially correct outcomes (5). There were significant associations between AI models and performance outcomes for both simple and complex procedures.
Conclusions: This study highlights the feasibility and potential benefits of integrating LLMs into the CPT coding process for hand surgery. The findings advocate for further refinement and training of AI models to improve their accuracy and practicality, suggesting a future where AI-assisted coding could become a standard component of surgical workflows, aligning with the ongoing digital transformation in health care.
Type of study/level of evidence: Observational, IIIb.

Since their inception, large language models (LLMs) like ChatGPT have emerged as invaluable tools across various surgical disciplines, encompassing hand surgery, otolaryngology, and neurosurgery.[1–6] Existing research has explored LLMs within educational frameworks, focusing on disseminating patient information and augmenting surgical training.[7,8] LLMs hold the potential to fundamentally transform the landscape of medicine, facilitating the generation of predictive text for physician notes, streamlining order completion, and aiding in the accurate selection of International Classification of Disease and Current Procedural Terminology (CPT) codes.[9]

Accurate CPT coding can be arduous and time-intensive and requires specialized training, which has now been further hindered

Corresponding author: Emily L. Isch, MD, 840 Walnut Street, 15th Floor, Philadelphia, PA 19107.
E-mail address: Emily.isch@jefferson.edu (E.L. Isch).

https://doi.org/10.1016/j.jhsg.2024.11.013

by the widespread shortage of professional coders for surgical specialties.[9] The efficacy of machine learning in predictive coding has been substantiated in non-surgical domains, with machine learning models demonstrating reasonable accuracy in predicting CPT codes from processed pathology reports.[10] In surgical subspecialties such as neurosurgery, the use of LLMs for accurately coding diverse procedures has garnered attention, paralleling the competencies of specialized natural language processing (NLP) models.[11–13]

Hand surgery involves a broad range of procedures performed on patients of all ages and remains the domain of several surgical specialties including hand surgery, trauma surgery, and general surgery. Unsurprisingly, there is substantial variation in CPT coding within hand surgery, and these discrepancies persist even after considering factors such as the experience level of the surgeon, practice environment, and educational background.[14] Leveraging pre-trained and standardized artificial intelligence (AI) models presents a promising solution to mitigating this problem.

The purpose of this study was to explore and quantify the ability of these LLM services to correctly identify CPT codes of operative hand procedures. We hypothesized that there would be variability in accuracy between the different LLM services but that AI would achieve accurate CPT coding for a subset of hand procedures.

Materials and Methods

Study design

This study employed a quantitative approach to evaluate the capacity of publicly accessible LLMs in identifying the appropriate CPT code(s) for a given hand procedure. The research aims to assess the efficacy of these models in automating the process of CPT code assignment.

Ethical considerations

It is important to emphasize that this research does not meet the criteria for human subject research, as it involves the analysis of publicly available textual data without direct involvement or impact on human subjects. Consequently, the need for local institutional review board approval was deemed unnecessary for this study.

Data collection

For this study, five prominent large language model platforms were evaluated: Perplexity.AI, Bard (developed by Google), BingAI (developed by Microsoft), ChatGPT 3.5 (developed by OpenAI), and ChatGPT 4.0 (also developed by OpenAI). All data collection and analysis were conducted from March to June 2024.

Query formulation

A standardized method was employed to assess the ability of these LLMs to determine the appropriate CPT codes for hand procedures. Each language model was presented with a standardized query, which commenced with the question: “What are the appropriate CPT codes for...?” This initial question was followed by the specific name of the procedure under consideration. For the five complex cases in which the procedure involved multiple distinct procedures, they were provided in sequence with the word “AND” between.

Table 1
χ^2 Test Results for Simple Procedures

AI Model	Correct	Incorrect	Total
Perplexity.AI	15	0	15
Bard	14	1	15
Bing AI	14	1	15
ChatGPT 3.5	7	8	15
ChatGPT 4.0	8	7	15
Total	58	17	75
			Test statistic: 21.75
			P value: .000224

Procedure selection

Twenty of the most common hand procedures performed at our tertiary care academic institution were selected, as determined by the senior authors, hand surgery attending surgeons whom have undergone hand surgery training. Fifteen were simple, single-code procedures, and five were complex multicode procedures.

Response recording

The responses generated by each large language model were recorded for subsequent evaluation. To determine the accuracy of the responses, a predefined classification system was applied as follows.

Correct: Responses were designated as “correct” if it provided either all of the accurate CPT codes for procedures with multiple applicable codes or the sole correct CPT code for procedures with a single applicable code. Additionally, if a response presented multiple CPT codes but identified the first code as the most accurate, it was also classified as “correct.” The reference standard for CPT codes was determined by two senior plastic surgeons with over 15 years of experience in hand surgery. Each case was independently reviewed, and discrepancies were resolved with discussions with our billing department.

Partially Correct: Responses were designated as “partially correct” if they provided some, but not all, of the accurate CPT codes for procedures with multiple codes.

Incorrect: Responses were designated as “incorrect” if it failed to include any of the accurate CPT codes for the given procedure.

Results

The data set comprises performance outcomes of AI models on both simple and complex procedures. For simple procedures, AI model was asked to evaluate 15 separate procedures, while for complex procedures, each model was asked to evaluate 5 distinct surgical procedures which each code for two distinct CPT codes. For simple procedures, Perplexity.AI achieved 15 correct outcomes (100%), Bard and BingAI each achieved 14 correct outcomes (93.3%), whereas ChatGPT 4.0 and 3.5 achieved 8 (53.3%) and 7 (46.7%) correct outcomes, respectively. For complex procedures, Perplexity.AI and Bard achieved 3 correct outcomes each (60%), and Bing AI demonstrated five partially correct outcomes (100%). In contrast, both ChatGPT models achieved no correct outcomes (0%).

To evaluate the association between AI models and their performance outcomes, χ^2 tests were conducted (Tables 1 and 2). For simple procedures, the χ^2 test yielded a statistic of 21.75 with a *P* value of .000224, signifying a statistically significant association between AI models and performance outcomes. A significance threshold of 0.05 was employed. In contrast, the χ^2 test for complex procedures resulted in a statistic of 25.95 with a *P* value of .001071, also indicating a statistically significant association.

Table 2
χ^2 Test Results for Complex Procedures

AI Model	Correct	Partially Correct	Incorrect	Total
Perplexity.AI	3	1	1	5
Bard	3	0	2	5
Bing AI	0	5	0	5
ChatGPT 3.5	0	0	5	5
ChatGPT 4.0	0	1	0	5
Total	6	7	8	25
			Test Statistic: 25.95	
			P value: .001071	

Table 3
Mean Accuracy and Standard Error by AI Model and Complexity

AI Model	Complexity	Mean Accuracy	Standard Error
Perplexity.AI	Simple	1	0.1127
Perplexity.AI	Complex	0.6	0.147
Bard	Simple	0.933	0.1127
Bard	Complex	0.6	0.147
Bing AI	Simple	0.933	0.1127
Bing AI	Complex	0	0.147
ChatGPT 3.5	Simple	0.467	0.1127
ChatGPT 3.5	Complex	0	0.147
ChatGPT 4	Simple	0.533	0.1127
ChatGPT 4	Complex	0	0.147

Table 3 and the Figure illustrate the mean accuracy and standard error of each AI model for simple and complex hand surgery procedures. Perplexity.AI demonstrated the highest mean accuracy for simple procedures at 1.0, with a standard error of 0.1127, indicating perfect accuracy in identifying the correct CPT codes. Bard and Bing AI also performed well in simple procedures, each achieving a mean accuracy of 0.933, with a standard error of 0.1127, showing a minor deviation from Perplexity.AI. For complex procedures, Perplexity.AI and Bard both achieved a mean accuracy of 0.6 with a standard error of 0.147, signifying a drop in performance compared to simpler tasks. Bing AI, despite performing well in simple procedures, recorded a mean accuracy of zero for complex cases, further emphasizing the challenges faced by the models in accurately identifying CPT codes for more intricate surgeries. ChatGPT 3.5 and ChatGPT 4.0 displayed notably lower mean accuracies, particularly for complex procedures, in which both models achieved a mean accuracy of 0, underscoring their limited capability in handling complex surgical coding tasks. Figure 1 visually corroborates these findings, illustrating the clear performance gap between Perplexity.AI and Bard compared to the ChatGPT models, especially as task complexity increases.

Discussion

This study provides a critical evaluation of the efficacy of various LLMs in accurately identifying CPT codes for hand surgery procedures. The results reveal a considerable variation in performance among the models, with Perplexity.AI and Bard demonstrating higher accuracy and reliability compared to others in simple procedures. These findings are indicative of the current capabilities and limitations of AI-assisted coding in surgical disciplines.

Key findings

The study's results underscore the effectiveness of Perplexity.AI and Bard, which outperformed other models in both simple and complex hand surgery procedures. For simple procedures, Perplexity.AI achieved a perfect score, correctly identifying all CPT codes, whereas Bard and Bing AI also performed commendably with only minor inaccuracies. These results are encouraging, as they suggest that certain LLMs have reached a level of maturity that allows them to handle straightforward coding tasks with high precision. The consistency in their performance across simple tasks highlights their potential as reliable tools in clinical settings, where efficiency and accuracy are paramount.

In contrast, the performance of ChatGPT models (3.5 and 4.0) was notably inferior, especially for complex procedures. These models struggled to accurately identify CPT codes for more intricate surgeries, failing to achieve correct outcomes in any of the complex cases. The high rate of partially correct responses by Bing AI for complex tasks suggests that although some models may grasp the overall context of a procedure, they lack the nuanced understanding necessary for precise coding.

Regardless, this limitation is particularly concerning given the critical role that accurate CPT coding plays in health care. CPT codes are critical in accurate billing, reimbursement, and clinical documentation. Thus, inaccurate CPT coding through LLMs can harm rather than help physicians, as it would require constant monitoring by a physician for potential errors. With the current system, physicians use nearly 12% of their net patient service revenue to cover the costs of excessive administrative complexity.[15] Although LLMs certainly have the potential to drastically get rid of this burden, it is easy to see how much harm they can do if their accuracy is not ensured.

The χ^2 analysis further substantiated the observed discrepancies, with statistically significant associations between the AI models and their performance outcomes for both simple ($P = .000224$) and complex procedures ($P = .001071$). This statistical significance not only validates the findings but also underscores the importance of model selection in clinical practice. The variability in performance across different LLMs suggests that health care providers should exercise caution when integrating these tools into their workflows, particularly for complex cases that demand a higher level of coding accuracy.

In this study, no a priori sample size estimation was performed given its exploratory nature. A post hoc power analysis was conducted to evaluate the adequacy of the sample size for detecting differences among AI models. The analysis revealed a statistical power of approximately 29.1% for detecting the observed effect size (0.533) between the best-performing model (Perplexity.AI) and the lowest-performing model (ChatGPT 3.5). This falls below the conventional 80% threshold for sufficient power, indicating that the study may be underpowered to confidently detect significant differences. As such, *P* values are presented with caution, and findings are interpreted as exploratory. This limitation underscores the need for future studies with larger sample sizes to validate and extend these findings.

Implications for Clinical Practice

The findings of this study have important implications for the future integration of AI in surgical coding. The high accuracy of models like Perplexity.AI and Bard in simple procedures suggests that these tools could be effectively employed to automate routine coding tasks, thereby reducing the burden on human coders and minimizing the risk of errors. This could lead to more efficient processing of surgical cases, faster billing cycles, and potentially improved financial outcomes for health care institutions. This may imply saving up to several billion-dollars each year—the current cost spent in the identification of proper CPT billing, decreasing excessive costs for both physicians and patients.[15]

However, the challenges faced by all models, particularly in complex procedures, highlight the current limitations of AI in fully replacing human coders. The inability of the ChatGPT models to accurately code complex surgeries indicates that although LLMs can

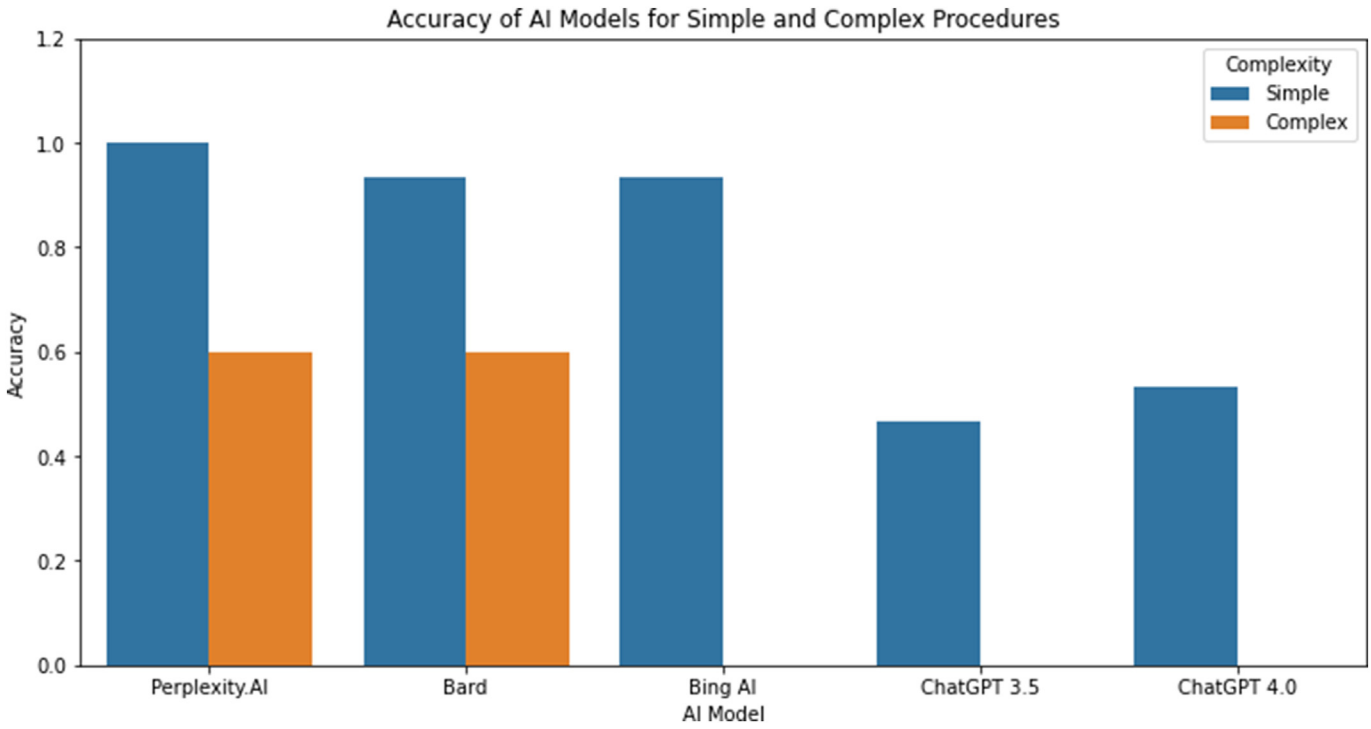

Figure 1. Accuracy of AI models for simple and complex procedures

assist in the coding process, they are not yet capable of independently managing the intricacies of surgical coding. These models should be viewed as supplementary tools rather than replacements for human expertise, particularly in cases where coding accuracy is critical.

Future directions

The study also points to several avenues for future research and development. First, further refinement of LLMs, particularly models like ChatGPT 3.5 and 4.0, is necessary to enhance their performance in complex coding tasks. This could involve additional training with more specialized medical data sets, as well as the development of algorithms that can better interpret the multifaceted nature of surgical procedures. The improvement of performance with pre-training has been well-established; previous study by Zaidat et al[16] found an improvement in ChatGPT 4.0's accuracy from 63 percent to 73 percent when it was primed with three operative notes and associated spinal CPT codes, and to 76% when it was exposed to a list of every possible CPT code. XLNet, a NLP algorithm, increased in accuracy from 71% to 88% when a "generalized autoregressive pre-training method" was given.[12] Although it may require an extra step, an improvement in a LLM's accuracy is promising as well given that they get preemptive training by a physician to introduce sample CPT codes.

Moreover, expanding the data set to include a broader range of surgical specialties and procedures would help validate the generalizability of the findings and ensure that the models can perform reliably across different contexts. The applicability of LLMs has already been studied in various surgical fields including hand surgery, otolaryngology, craniofacial surgery, and neurosurgery; it is only a matter of time LLMs primed with various surgical data emerge. Future studies should thus explore the integration of AI-assisted coding with existing clinical documentation systems to streamline workflows and enhance interoperability.

Finally, there is a need to investigate the underlying factors contributing to the performance discrepancies among different LLMs. Understanding the specific strengths and weaknesses of each model will allow for more targeted improvements and ensure that AI tools are optimized for the unique demands of surgical coding.

The findings highlight the variability in AI model performance based on task complexity. Perplexity.AI and Bard consistently performed better, suggesting their robustness and reliability for both simple and complex procedures. Conversely, the performance of ChatGPT models indicates the need for further refinement and training, especially for complex tasks. These insights are crucial for health care providers and decision-makers when selecting AI tools for clinical applications. The study underscores the importance of evaluating AI models not just broadly, but in the context of specific task complexities to ensure optimal outcomes.

Future research should explore the underlying reasons for performance discrepancies among AI models. It should investigate the potential for model improvement, particularly for those like ChatGPT 3.5 and ChatGPT 4.0, in handling complex tasks. Additionally, expanding the data set and including more AI models would help validate and extend the findings. By continuing to refine and evaluate AI models, the health care industry can leverage these tools more effectively, ultimately improving patient outcomes and operational efficiency.

Conflicts of Interest

No benefits in any form have been received or will be received related directly to this article.

References

1. Gupta R, Park JB, Bisht C, et al. Expanding cosmetic plastic surgery research with ChatGPT. *Aesthet Surg J*. 2023;43(8):930–937.

2. Sharma SC, Ramchandani JP, Thakker A, Lahiri A. ChatGPT in plastic and reconstructive surgery. *Indian J Plast Surg*. 2023;56(4):320–325.
3. Zalzal HG, Cheng J, Shah RK. Evaluating the current ability of ChatGPT to assist in professional otolaryngology education. *OTO Open*. 2023;7(4):e94.
4. Revercomb L, Patel AM, Choudhry HS, Filimonov A. Performance of ChatGPT in otolaryngology knowledge assessment. *Am J Otolaryngol*. 2024;45(1):104082.
5. Liu J, Zheng J, Cai X, Wu D, Yin C. A descriptive study based on the comparison of ChatGPT and evidence-based neurosurgeons. *iScience*. 2023;26(9):107590.
6. Mannam SS, Subtirelu R, Chauhan D, et al. Large language model-based neurosurgical evaluation matrix: a novel scoring criteria to assess the efficacy of ChatGPT as an educational tool for neurosurgery board preparation. *World Neurosurg*. 2023;180:e765–e773.
7. Sevgi UT, Erol G, Doğruel Y, Sönmez OF, Tubbs RS, Güngor A. The role of an open artificial intelligence platform in modern neurosurgical education: a preliminary study. *Neurosurg Rev*. 2023;46(1):86.
8. Mishra A, Begley SL, Chen A, et al. Exploring the intersection of artificial intelligence and neurosurgery: let us be cautious with ChatGPT. *Neurosurgery*. 2023;93(6):1366–1373.
9. DiGiorgio AM, Ehrenfeld JM. Artificial intelligence in medicine & ChatGPT: de-tether the physician. *J Med Syst*. 2023;47(1):32.
10. Levy J, Vattikonda N, Haudenschild C, Christensen B, Vaickus L. Comparison of machine-learning algorithms for the prediction of Current Procedural Terminology (CPT) codes from pathology reports. *J Pathol Inform*. 2022;13:3.
11. O'Malley GR, Jr., Sarwar SA, Cassimatis ND, et al. Can publicly available artificial intelligence successfully identify Current Procedural Terminology Codes for common procedures in neurosurgery? *World Neurosurg*. 2024;183:e860–e870.
12. Zaidat B, Tang J, Arvind V, et al. Can a novel natural language processing model and artificial intelligence automatically generate billing codes from spine surgical operative notes? *Global Spine J*. 2024;14(7):2022–2030.
13. Ewings EL, Konofaos P, Wallace RD. Variations in current procedural terminology coding for craniofacial surgery: A need for review and change. *J Craniofac Surg*. 2017;28(5):1224–1228.
14. Jazayeri HE, Khavanin N, Yu JW, et al. Variability in Current Procedural Terminology codes for craniomaxillofacial trauma reconstruction: a national survey. *J Craniofac Surg*. 2020;31(4):996–999.
15. Blanchfield BB, Heffernan JL, Osgood B, Sheehan RR, Meyer GS. Saving billions of dollars–and physicians' time–by streamlining billing practices. *Health Aff (Millwood)*. 2010;29(6):1248–1254.
16. Zaidat B, Lahoti YS, Yu A, Mohamed KS, Cho SK, Kim JS. Artificially Intelligent billing in spine surgery: an analysis of a large language model. *Global Spine J*. 2023:21925682231224753.