Artificially Intelligent Billing in Spine Surgery: An Analysis of a Large Language Model

Bashar Zaidat, BS[1] [iD], Yash S. Lahoti, MSE[1], Alexander Yu, BS[1] [iD], Kareem S. Mohamed, BS[1], Samuel K. Cho, MD[1], and Jun S. Kim, MD[1]

Abstract

Study Design: Retrospective cohort study.

Objectives: This study assessed the effectiveness of a popular large language model, ChatGPT-4, in predicting Current Procedural Terminology (CPT) codes from surgical operative notes. By employing a combination of prompt engineering, natural language processing (NLP), and machine learning techniques on standard operative notes, the study sought to enhance billing efficiency, optimize revenue collection, and reduce coding errors.

Methods: The model was given 3 different types of prompts for 50 surgical operative notes from 2 spine surgeons. The first trial was simply asking the model to generate CPT codes for a given OP note. The second trial included 3 OP notes and associated CPT codes to, and the third trial included a list of every possible CPT code in the dataset to prime the model. CPT codes generated by the model were compared to those generated by the billing department. Model evaluation was performed in the form of calculating the area under the ROC (AUROC), and area under precision-recall curves (AUPRC).

Results: The trial that involved priming ChatGPT with a list of every possible CPT code performed the best, with an AUROC of .87 and an AUPRC of .67, and an AUROC of .81 and AUPRC of .76 when examining only the most common CPT codes.

Conclusions: ChatGPT-4 can aid in automating CPT billing from orthopedic surgery operative notes, driving down healthcare expenditures and enhancing billing code precision as the model evolves and fine-tuning becomes available.

Keywords

spine, current procedural terminology, natural language processing, large language model, chatGPT, prompt engineering

Introduction

The process of assigning Current Procedural Terminology (CPT) billing codes from operative notes has long been acknowledged as a resource-intensive task and susceptible to human error.[1] The development of the CPT system can be traced back to the American Medical Association's effort to establish a universal framework for encoding medical procedures. These codes, each comprising 6 digits, are instrumental in conveying in-depth information regarding performed surgical procedures. Challenges may surface in scenarios where a singular procedure aligns with multiple CPT codes. Accurate coding is essential for proper billing and reimbursement within the United States healthcare system, but the manual generation of CPT codes is often time-consuming and can cause unnecessary administrative expenses. Studies have revealed that operational costs constitute roughly 15% to 25% of the complete national healthcare spending, amounting to an approximate annual total of $600 billion to $1 trillion in 2019.[2] Billing and coding expenditures emerge as prominent

[1] Department of Orthopaedic Surgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Corresponding Author:
Jun S. Kim, MD, Icahn School of Medicine at Mount Sinai, Mount Sinai West, 425 West 59th Street, New York, NY 10019, USA.
Email: jun.kim@mountsinai.org

factors driving this economic load.[3] Thus, there is an escalating demand for innovative solutions aimed at streamlining this process while increasing precision.

Orthopedic surgery involves a wide range of procedures, each with its unique set of CPT codes, making coding a complex task for human coders. Furthermore, spine surgery billing can be particularly intricate due to multi-level procedures and the vast number of bone grafts, intervertebral devices, plates, and other implants. AI-powered natural language processing (NLP) machine learning algorithms like ChatGPT-4 have garnered attention for their ability to comprehend free text. Additionally, ChatGPT-4 has been proposed as a tool to supplement clinical decision-making and educate patients about common pathologies, treatment options, and surgical outcomes in the realm of orthopedics. Since 80% of the electronic medical record is recorded in unstructured text,[4] ChatGPT-4, the latest and largest iteration of ChatGPT trained on billions of text sources, opens up promising possibilities for automating coding tasks in the healthcare domain. By harnessing the language generation capabilities of ChatGPT-4, it is possible to extract vital information from operative notes and automatically generate accurate CPT codes. This automation has the potential to significantly alleviate the workload of healthcare professionals, ultimately optimizing the efficiency and precision of billing in the United States healthcare system.

Our study aims to investigate the utility of ChatGPT-4 as a predictive tool for CPT billing codes based on operative notes from orthopedic surgeries. The findings of this study may shed light on ChatGPT-4's effectiveness as an innovative approach to alleviate challenges linked with CPT coding in orthopedic surgery, potentially providing a transformative alternative to the current manual approach.
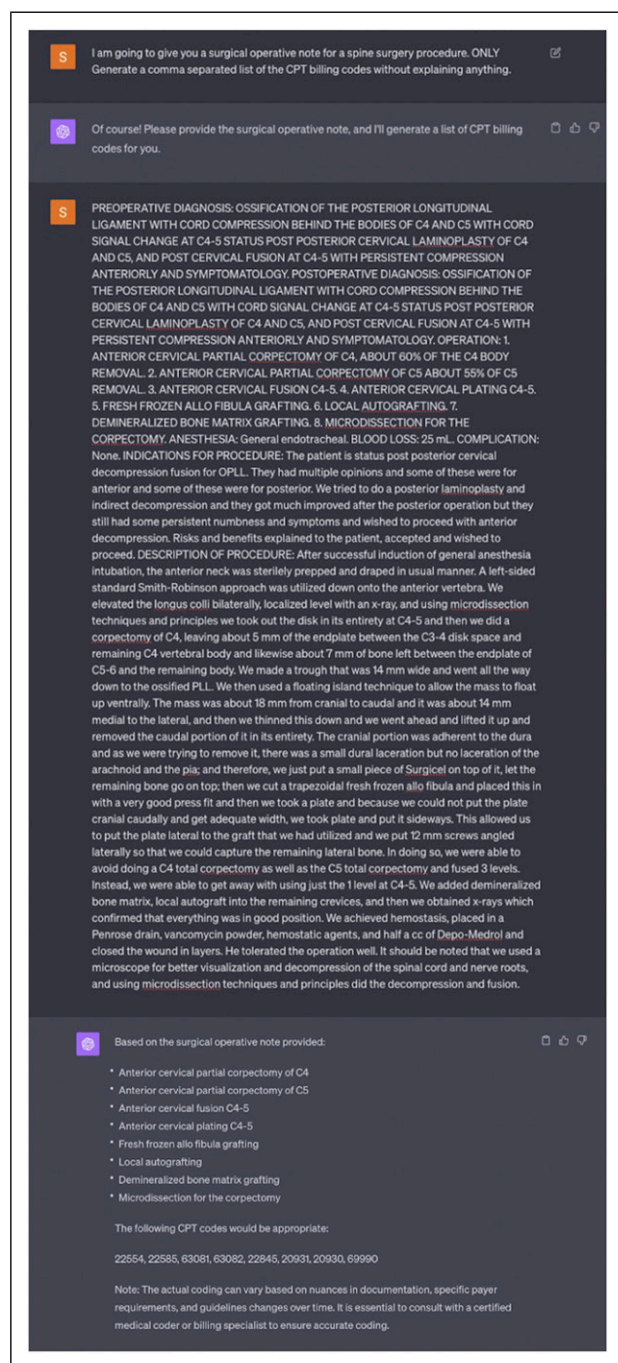
Methods

Study Design

Fifty surgical operative notes were collected from 2 different spine surgeons. All protected patient data was removed from the operative notes. The study uses the large language model, GPT 4.0, to predict CPT Billing codes from surgical operative notes on patients who underwent orthopedic spine surgery. Three distinct trials were executed to assess the efficacy of the language model. For all 3 trials, the same dataset of 50 operative notes was used. A separate ChatGPT session was used for each operative note to minimize potential carry-over effects between operative notes as well as reduce the risk of cumulative knowledge learned from an operative note influencing the CPT billing codes assignments for subsequent notes. Further analysis was performed on a dataset where the CPT codes appearing less than 3 times were removed. The IRB approved the present study and granted waiver for consent of patient data. This study has been IRB-Approved at Columbia University Irving Medical Center under protocol number AAAS8683.
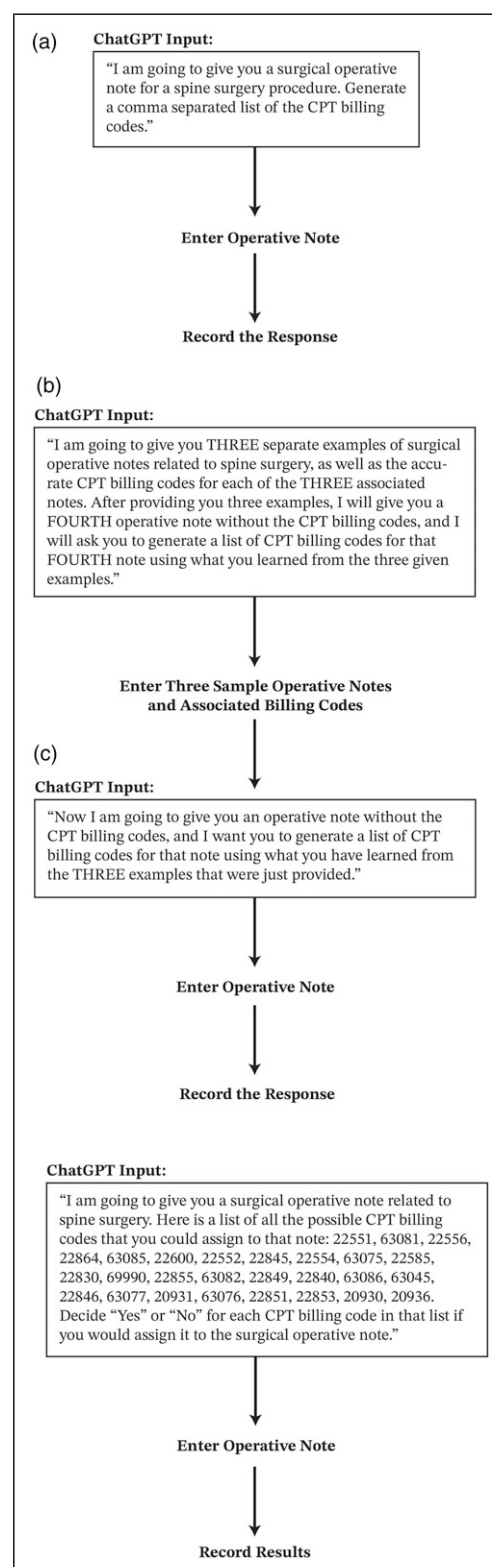
Figure 1. Prompt engineering used in the 3 trails. Figure 1A corresponds to Trial 1 - Open-Ended Prediction. Figure 2A corresponds to Trial 2 - Pre-Training with Examples. Figure 1C corresponds to Trial 3 - Psuedo-Classification Approach.

Figure 2. Sample prompt using the methodology of Trial 1.

Table 1. CPT Code Counts.

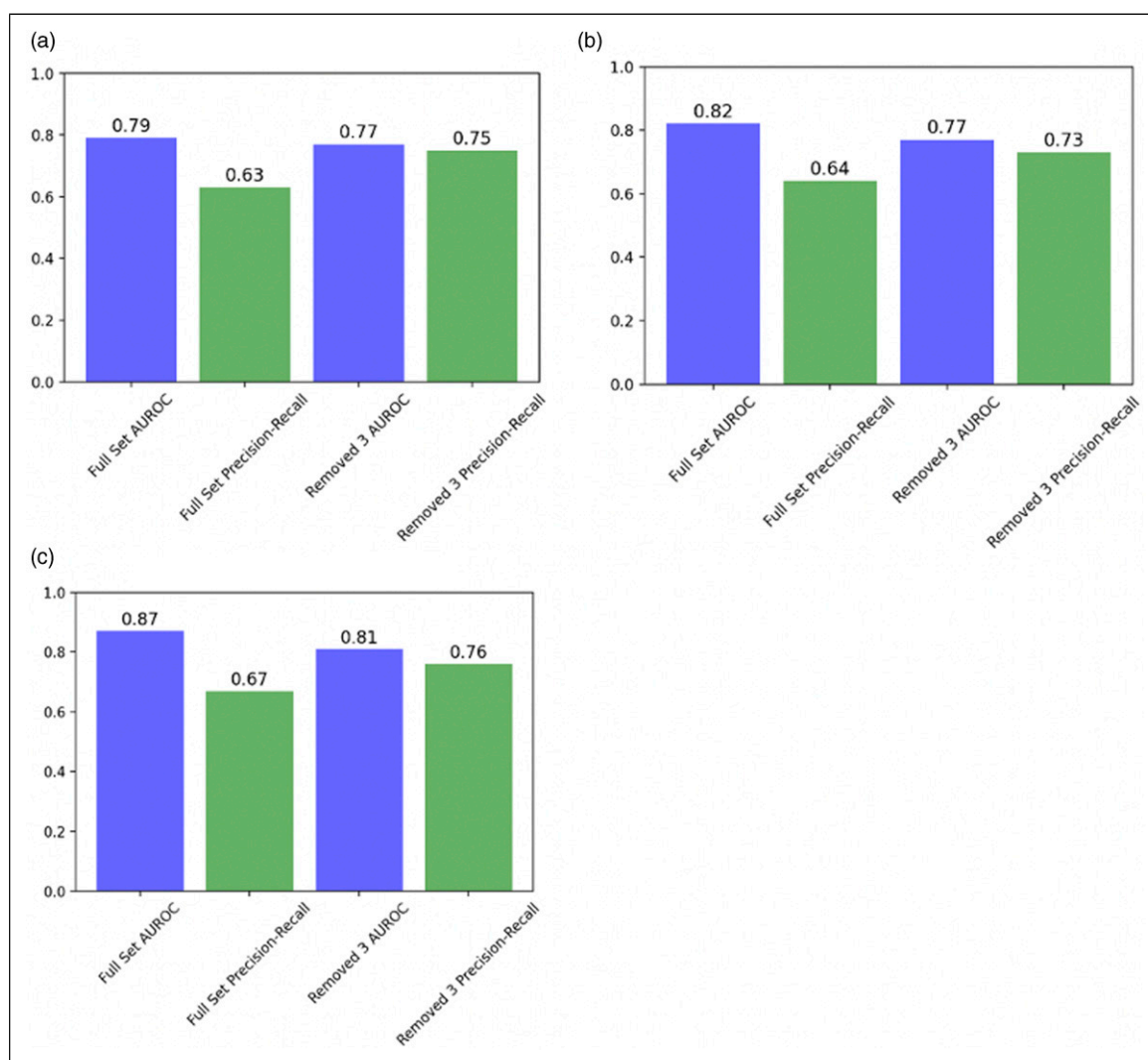| CPT Code	Count
22551	45
63081	6
22556	1
22864	1
22552	29
22845	46
22554	6
63075	3
22585	6
22830	3
69990	6
22855	1
63082	1
22846	3
20931	50
63077	1
63076	1
22851	17
20930	35
20936	48
63078	1
20932	1

Trial 2 - Pre-Training With Examples

Trial 2 consisted of pre-training ChatGPT using 3 separate examples. The model was first prompted to train from 3 distinct operative notes and their respective, accurate CPT billing codes. ChatGPT was then prompted to generate CPT billing codes for each independent operative note using the knowledge acquired from the pre-training examples (Figure 1B). This trial served to assess ChatGPT's predictive performance on CPT billing code assignment when pre-trained.

Trial 3 - Pseudo-Classification Approach

A pseudo-classifier approach was employed in the third trial. ChatGPT was provided a comprehensive list of all possible CPT billing codes related to orthopedic spine surgery and was prompted to assign the appropriate CPT billing code for each operative note using a "yes" or "no" response for each provided CPT billing code (Figure 1C).

AUROC and AUPRC Measurement

The area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) were the metrics used to assess the performance of ChatGPT. Using both AUROC and AUPRC is important for thorough model assessment. AUROC gauges class distinction across various thresholds, while Precision-Recall helps gauge

Trial 1 - Open-Ended Prediction

In Trial 1, ChatGPT was prompted to generate a list of CPT billing codes when provided an operative note (Figure 1A). This trial serves to evaluate ChatGPT's baseline performance solely based on each independent operative note. A sample prompt can be seen in Figure 2.

Figure 3. AUROC/AUPRC comparison for various datasets across trails. (A) Trail 1 - AUROC/AUPRC Comparison. (B) Trail 2 - AUROC/ AUPRC Comparison. (C) Trail 3 - AUROC/AUPRC Comparison.

positive class identification, especially in imbalanced data. This dual approach offers a holistic view of model robustness and informs threshold choices based on specific application requirements. Higher values of AUROC and AUPRC indicate improved performance of the machine learning model's ability to predict accurate CPT billing codes for the operative notes (Table 1).

Results

Trial 1

The full dataset analysis in Trial 1 yielded an AUROC score of .79 and an AUPRC score of .63 (Figure 3A). When considering a subset of CPT billing codes that appeared at least 3

times, the AUROC decreased to .77, with a corresponding increased AUPRC of .75.

Trial 2

In Trial 2, the full dataset analysis yielded from pre-training ChatGPT resulted in an AUROC score of .82 and an AUPRC score of .64 (Figure 3B). Within the subset of frequently occurring CPT codes, the AUROC score was .77, with an AUPRC of .73.

Trial 3

The full dataset analysis using the pseudo-classification approach in Trial 3 yielded an AUROC score of .87 and an AUPRC score of .67. Within the subset of frequently
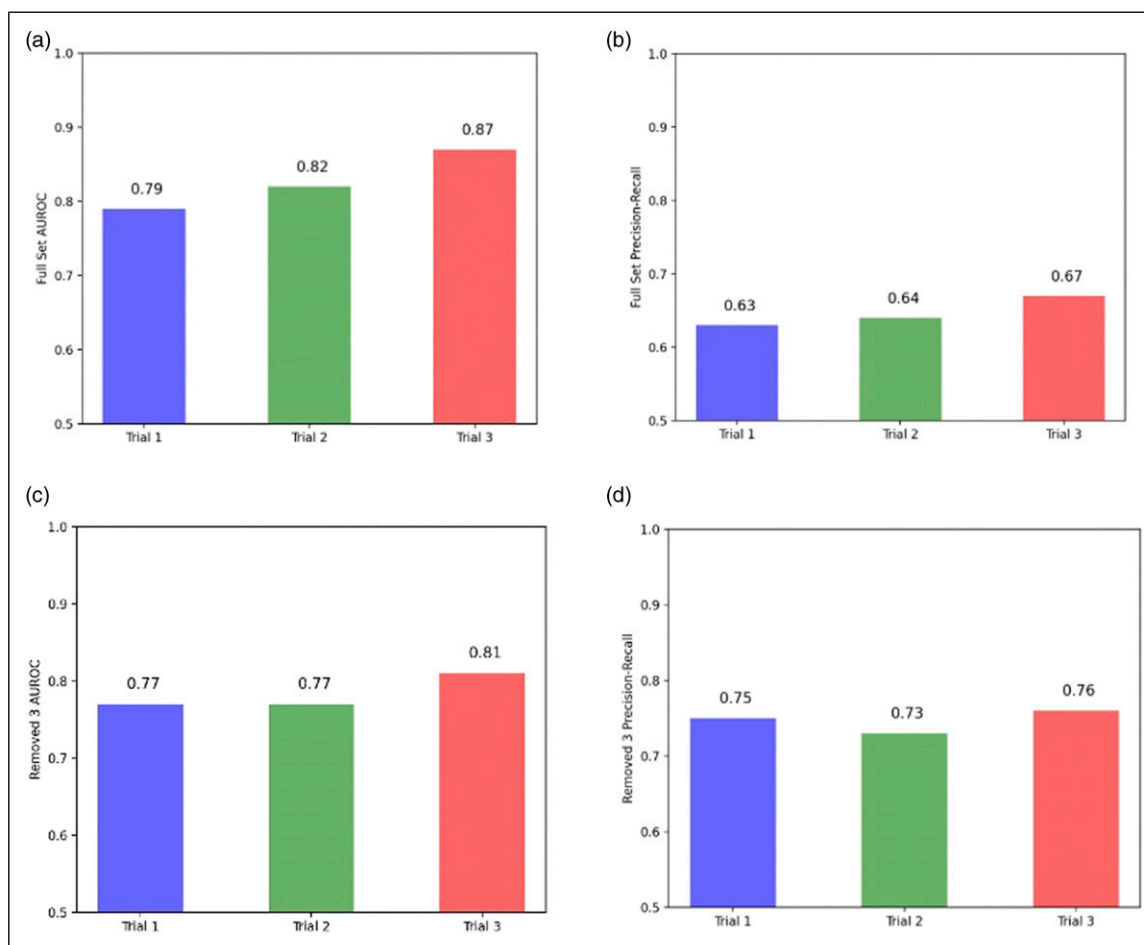
Figure 4. Comparison of AUROC and AUPRC between trails. (A) Full Set AUROC Comparison Across Trails. (B) Full Set Precision-Recall Comparison Across Trails. (C) Removed 3 AUROC Comparison Across Trails. (D) Removed 3 Precision-Recall Comparison Across Trails.

occurring CPT codes of 3 or more, the AUROC score reached .81, with an AUPRC of .76 (Figure 3C).

Comparison of Trials

The results demonstrate notable differences in ChatGPT's performance to accurately assign CPT billing codes across the 3 trials. The pseudo-classification approach in Trial 3 demonstrated the highest performance with an AUROC score of .87. This was followed by the pre-training approach of Trial 2 with an AUROC score of .82, leaving Trial 1 with the lowest AUROC score of .79 (Figure 4A). A similar trend was observed in the full set precision-recall analysis, with Trial 3 exhibiting the highest AUPRC score of .67. This was followed by an AUPRC score of .64 and .63 from Trials 2 and 3, respectively (Figure 4B). When looking at frequently occurring CPT billing codes that appeared at least 3 times in the dataset, Trial 3 maintained the highest AUROC score of .81, surpassing the AUROC score of .77 shared by Trials 2 and 3 (Figure 4C). Trial 3 again demonstrated the highest performance of .76 when looking at precision-recall for

frequently occurring CPT billing codes. This was followed by Trial 1's AUPRC score of .75 and Trial 2's score of .73 (Figure 4D).

Discussion

Background

This study is the third in a series of NLP experiments investigating the reliability and effectiveness of AI in automated CPT code generation. Our research previously demonstrated that deep learning models, such as pre-trained XLNet, can be fine-tuned with clinical notes to achieve human-level performance of CPT billing code generation from operating notes.[5,6] In this study, we demonstrate that without model fine-tuning and using a limited training dataset, the ChatGPT-4 general model can achieve comparable performance to previous deep learning performance specifically trained for this task. The creation of large language models has altered the way in which clinicians can interact with the vast quantity of data generated in medical practice. ChatGPT-4 has entered the

mainstream this past year as 1 of the most widely adopted AI technologies and will have profound impacts on clinical practice. To the authors' knowledge, this is the first study evaluating ChatGPT-4 performance for the task of CPT code generation.

The current approach to CPT billing has generated a tremendous amount of excess healthcare costs. Billing and coding costs consume 62% of the administrative-related hospital budget and up to 14% of revenue for physician groups.[7,8] When looking across the US, the financial burden of insurance and billing tasks amounted to over $25 billion dollars[9]. Despite the heavy investment, inaccuracies in coding, such as downcoding, can decrease revenue for physicians or lead to fraudulent billing. Within government provided insurance, the Centers of Medicaid and Medicare service have reported $95 billion in billing payments were fraudulent.[10,11] Given the complexity and error-prone process, Morra et al finds that physicians in the US spend 4× longer in billing and documentation management related tasks than physicians in Canada.[12] These automated tools can greatly reduce labor costs spent on manually transcribing billing codes and assist with the execution of administrative tasks. With an increasing demand for medical service, a system must be set in place that can standardize the billing code generation process and include checks to detect and correct coding mistakes. As hospitals generate more clinical data and access to computational resources expands, we enter an era where automated AI tools can improve throughput of patient volume, minimize clerical errors, and alleviate the growing physician workload.

Model Performance

Overall the GPT model performed within expectations, with the "pseudo-classifier" trial showing promising results. Similarly, the second trial with the attempt to manually fine-tune the model performed better than the first iteration. However, there was a large gap in both AUROC and AUPRC scores between different CPT codes, with some having excellent results in the .9 range and others in the .2 range. To combat the class imbalance faced in the dataset, overall AUROC and AUPRC were calculated after removing the CPT codes that appeared in less than 3 operative notes. For these more common operative notes there was a noticeable difference in AUPRC scores, and only a marginal decrease in the AUROC scores, suggesting that the GPT model is more robust when generating the more common CPT codes. Moreover, because GPT is trained on a general dataset, it is likely biased towards the more common codes and lacks some of the nuance required for generating more specific codes.

The trials involving prompt engineering performed better than the simple prompt. For "Trial 2" this meant feeding 3 operative notes into the model and giving the ground truth CPT codes. GPT is a transformer based model and thus is able to take context into account. Moreover, it has been demonstrated that more data enables better performance of transformer models.[13] Therefore, this trial served to function as a sort of method of "pre-training" the model as the API was not in public access for GPT-4 at the time of this manuscript. There was an overall improvement in both the AUROC and AUPRC when compared to the simple prompt used in the first trial.

The third trial performed even better. For this trial, GPT was given a list of all the possible CPT codes in the dataset, and instructed to determine "Yes" or "No" for each of those CPT codes for a given note. This served as a sort of classification task, defined as a model deciding how probable it is that a given item belongs to a specific class, in this case how likely it is that a specific operative note should have a CPT code assigned to it. For this prompt the AUROC and AUPRC were higher than the "pre-trained" version as well as the simple prompt. This is likely because GPT was presented with an easier task, instead of having to parse through the note and generate its own list of CPT codes from the hundreds of possible codes in its dataset it only had to decide from a much smaller subset.

This study is unique in that the GPT model has likely never seen surgical operative notes with their associated CPT billing codes, which presents unique challenges. Thus, it functions almost as a "zero-shot" model for this specific task, meaning that it has little context in its training data related to generating CPT codes from operative notes. Although, this is not truly "zero-shot" learning, since there is likely some data in the training dataset related to CPT billing, as GPT was trained using the CommonCrawl dataset, which includes data from a large number of web pages, some of which probably mention surgical billing, but notably does not include PubMed.[14] This may explain why it performed better in the second and third trial as it was given more information helping to combat the large imbalance in its training set.

This represents the third iteration of a series of models that have been evaluated for automatically generating billing codes from surgical operative notes.[5,6] This study is unique in that the GPT model uses a unidirectional transformer rather than the bi-directional transformer structure seen in the previous best-performing model, XLNet. The GPT model is also a much larger and more powerful model than XLNet, and has had a rapid rise in popular use gaining over 100 million users in its first few weeks. Kim et al. compared a Random Forest classifier to an LSTM model and found that the Random Forest model was able to achieve an AUROC of .94, but describe poor performance of the LSTM model which was likely the result of the small homogenous dataset used.[5] Zaidat et al. followed up on this study by including a larger dataset with a larger variety of CPT codes, and found that the XLNet model was able to outperform the random forest with greater robustness demonstrated by an AUROC of .95 and AUPRC of .91.[6]

The results of this study are promising as the best iteration of the GPT model was able to achieve a relatively high AUROC score of .87, with robustness for the most common

CPT codes represented by an AUPRC of .76. The discrepancy between the results of this study and previous models is likely because models trained on medical data have been shown to perform better at medical tasks.[15] We were not able to fully fine-tune the GPT model which would likely have shown even stronger results.

Benefits of ML Assisted Billing

The key benefit to the application of NLP to CPT code generation lies in improvements in coding accuracy and efficiency. As NLP increases the accuracy of code generation in a shorter amount of time, it can decrease the amount of charge lag time[16] and decrease the amount of needed support staff[17] without compromising the quality of care. Inappropriate or fraudulent coding, whether upcoding or downcoding, is a persistent issue in the medical field that totaled $28.91 billion in 2019 in the United States.[18] NLP models have previously shown aptitude at detecting fraud in prescriptions[19] and protected health information,[20] and as these models continue to improve, it is likely that they will confer similar benefits to code generation.

Several previous studies have already demonstrated promising results of using NLP models to automatically generate CPT codes from medical information.[5,6,11,21] Traditional methods of looking up information often involve large search engines; however, the development of large language models such as ChatGPT represent the latest evolution in the field of information retrieval. These models are also unique in that they are able to learn from previous responses and feedback as well as provide natural sounding responses to more complex queries. Even still, there is a paucity in the literature related to how the latest iterations of these models perform on specialized medical tasks such as billing.

While fine-tuned Machine Learning (ML) models are currently out-performing ChatGPT, ChatGPT has the advantage of a much larger training dataset size, so there is a higher potential ceiling for ChatGPT compared to these fine-tuned models.[22] This is because large training datasets confer greater potential to ChatGPT to understand and synthesize a larger variety of clinical and non-clinical contexts. Whereas a fine-tuned model may currently achieve greater accuracy in a specific medical setting, ChatGPT may be able to be more broadly applied across medical contexts without the need to use as much data when fine-tuning a new model for each specific context. Thus, GPT may be less costly data-wise and computationally in the future while achieving better results. The results of this study confirm that minimal additional fine tuning of prompts yields consistently better results from ChatGPT's responses, even approaching the results seen in the fine-tuned ML models, suggesting that better yields may be achieved with less effort from ChatGPT compared to the greater amount of effort required to train a fine-tuned model on a new large dataset.

Limitations and Next Steps

One limitation of this study is the small sample size. Future studies would benefit from a greater number of operative notes with more varied CPT codes. Because this study was based on only 3 surgeons at a single institution and only included spine surgeries, these results are not necessarily generalizable across institutions, and may not apply to all types of surgery. Our results suggest that any successful application of GPT to CPT code generation would require proper fine-tuning of the model to achieve the most accurate results.

Any study or application of ChatGPT must acknowledge the time discrepancy in its training data. At the time of this study, ChatGPT had only been trained on data up to 2021, so it is possible that recent changes or updates in billing have occurred or will occur in the future that would not be reflected in ChatGPT's responses. Additionally, as it is trained on a large, non-specific corpus, any billing code data the model previously saw was likely biased towards more common codes. The fact that ChatGPT is trained on non-medical text may also be another reason for limited AUROC and AUPRC scores. Providers should be cautioned that the accuracy is not high enough yet to merit its application without human supervision, until it becomes possible to properly fine-tune the model with a data-set validated by professional billers.

Large language models represent a natural progression in the field of information retrieval which has previously been dominated by search engines. To further explore the potential of ChatGPT-4 and advance its use in healthcare, several avenues of future research and development can be considered. These include efforts to fine-tune the model on healthcare-specific data, integrate real-time clinical data, and develop interfaces that allow ChatGPT-4 to seamlessly integrate with existing healthcare systems.

Conclusion

Overall the results of this study demonstrate that GPT has great potential for assisting with automating CPT billing from surgical operative notes. As the model develops and the ability to fine-tune it is made available, it will be able to help lessen the overhead and labor costs of the billing process, while introducing greater standardization and fraud prevention.

Declaration of conflicting interests

of interest or sources of support that require acknowledgement: *Bashar Zaidat, Yash S. Lahoti, Alexander Yu, Kareem S. Mohamed.*

ORCID iDs

Bashar Zaidat ⓘ https://orcid.org/0000-0002-8823-720X
Alexander Yu ⓘ https://orcid.org/0000-0002-7246-2269

References

1. Duszak R, Blackham WC, Kusiak GM, Majchrzak J. CPT coding by interventional radiologists: a multi-institutional evaluation of accuracy and its economic implications. *J Am Coll Radiol*. 2004; 1(10):734-740. doi:10.1016/j.jacr.2004.05.003

2. Tollen L, Keating E, Weil A. *How Administrative Spending Contributes to Excess US Health Spending*. Health Affairs; 2020. Published online February 20, 2020. doi:10.1377/forefront.20200218.375060. https://www.healthaffairs.org/content/forefront/administrative-spending-contributes-excess-us-health-spending

3. Shrank WH, Rogstad TL, Parekh N. Waste in the US health care system. *JAMA*. 2019;322(15):13978. doi:10.1001/jama.2019.13978

4. Verspoor K, Martin-Sanchez F. Big data in medicine is driving big changes. *Yearbook of Medical Informatics*. 2014;23(1): 14-20. doi:10.15265/iy-2014-0020

5. Kim JS, Vivas A, Arvind V, et al. Can natural language processing and artificial intelligence automate the generation of billing codes from operative note dictations? *J Global Spine*. Published online February. 2022;28:219256822211062831. doi:10.1177/21925682211062831

6. 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? *J Global Spine*. Published online March. 2023;18:219256822311649. doi:10.1177/21925682231164935

7. Tseng P, Kaplan RS, Richman BD, Shah MA, Schulman KA. Administrative costs associated with physician billing and insurance-related activities at an academic health care system. *JAMA*. 2018;319(7):691. doi:10.1001/jama.2017.19148

8. Kahn JG, Kronick R, Kreger M, Gans DN. The cost of health insurance administration in California: estimates for insurers, physicians, and hospitals. *Health Aff*. 2005;24(6):1629-1639. doi:10.1377/hlthaff.24.6.1629

9. Casalino LP, Nicholson S, Gans DN, et al. What does it cost physician practices to interact with health insurance plans? *Health Aff*. 2009;28(Supplement 1):w533-w543. doi:10.1377/hlthaff.28.4.w533

10. Drabiak K, Wolfson J. What should health care organizations do to reduce billing fraud and abuse? *AMA Journal of Ethics*. 2020; 22(3):E221-231. doi:10.1001/amajethics.2020.221

11. 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 Inf*. 2022;13:100165. doi:10.4103/jpi.jpi_52_21

12. Morra D, Nicholson S, Levinson W, Gans DN, Hammons T, Casalino LP. US physician practices versus Canadians: spending nearly four times as much money interacting with payers. *Health Aff*. 2011;30(8):1443-1450. doi:10.1377/hlthaff.2010.0893

13. Popel M, Bojar O. Training tips for the transformer model. *Prague Bull Math Linguist*. 2018;110(1):43-70. doi:10.2478/pralin-2018-0002

14. Buck C, Heafield K, Van Ooyen B. N-gram counts and language models from the common crawl. *Proc. Int. Conf. Lang. Resources Eval*. 2014;4:3579-3584.

15. Huang J, Osorio C, Sy LW. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Comput Methods Progr Biomed*. 2019;177:141-153. doi:10.1016/j.cmpb.2019.05.024

16. Manley R, Satiani B. Revenue cycle management. *J Vasc Surg*. 2009;50(5):1232-1238. doi:10.1016/j.jvs.2009.07.065

17. Grieger DL, Cohen SH, Krusch DA. A pilot study to document the return on investment for implementing an ambulatory electronic health record at an academic medical center. *J Am Coll Surg*. 2007;205(1):89-96. doi:10.1016/j.jamcollsurg.2007.02.074

18. Hammon M, Hammon WE. Benefiting from the government CERT audits. *J Oklahoma State Med Assoc*. 2005;98(8): 401-402. PMID: 16206870.

19. Haddad Soleymani M, Yaseri M, Farzadfar F, Mohammadpour A, Sharifi F, Kabir MJ. Detecting medical prescriptions suspected of fraud using an unsupervised data mining algorithm. *Daru*. 2018;26(2):209-214. doi:10.1007/s40199-018-0227-z

20. Oh SH, Kang M, Lee Y. Protected health information recognition by fine-tuning a pre-training transformer model. *Healthc Inform Res*. 2022;28(1):16-24. doi:10.4258/hir.2022.28.1.16

21. Burns ML, Mathis MR, Vandervest J, et al. Classification of current procedural Terminology codes from electronic health record data using machine learning. *Anesthesiology*. 2020; 132(4):738-749. doi:10.1097/aln.0000000000003150

22. Caruccio L, Cirillo S, Polese G, Solimando G, Sundaramurthy S, Tortora G. Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. *Expert Syst Appl*. 2023; 235:121186. doi:10.1016/j.eswa.2023.121186