


Can a Novel Natural Language Processing Model and Artificial Intelligence Automatically Generate Billing Codes From Spine Surgical Operative Notes?

Global Spine Journal
2024, Vol. 14(7) 2022–2030
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/21925682231164935
journals.sagepub.com/home/gsj


Bashar Zaidat, BS¹ , Justin Tang, BS¹ , Varun Arvind, PhD¹, Eric A. Geng, BA¹ , Brian Cho, MD¹, Akiro H. Duey, BS¹ , Calista Dominy, BS¹ , Kiehyun D. Riew, MD² , Samuel K. Cho, MD¹, and Jun S. Kim, MD¹

Abstract

Study Design: Retrospective cohort.

Objective: Billing and coding-related administrative tasks are a major source of healthcare expenditure in the United States. We aim to show that a second-iteration Natural Language Processing (NLP) machine learning algorithm, XLNet, can automate the generation of CPT codes from operative notes in ACDF, PCDF, and CDA procedures.

Methods: We collected 922 operative notes from patients who underwent ACDF, PCDF, or CDA from 2015 to 2020 and included CPT codes generated by the billing code department. We trained XLNet, a generalized autoregressive pretraining method, on this dataset and tested its performance by calculating AUROC and AUPRC.

Results: The performance of the model approached human accuracy. Trial 1 (ACDF) achieved an AUROC of .82 (range: .48-.93), an AUPRC of .81 (range: .45-.97), and class-by-class accuracy of 77% (range: 34%-91%); trial 2 (PCDF) achieved an AUROC of .83 (.44-.94), an AUPRC of .70 (.45-.96), and class-by-class accuracy of 71% (42%-93%); trial 3 (ACDF and CDA) achieved an AUROC of .95 (.68-.99), an AUPRC of .91 (.56-.98), and class-by-class accuracy of 87% (63%-99%); trial 4 (ACDF, PCDF, CDA) achieved an AUROC of .95 (.76-.99), an AUPRC of .84 (.49-.99), and class-by-class accuracy of 88% (70%-99%).

Conclusions: We show that the XLNet model can be successfully applied to orthopedic surgeon's operative notes to generate CPT billing codes. As NLP models as a whole continue to improve, billing can be greatly augmented with artificial intelligence assisted generation of CPT billing codes which will help minimize error and promote standardization in the process.

Keywords

cervical, disc replacement, fusion, ACDF, artificial intelligence, natural language processing, PCDF

Introduction

Administrative tasks are a major source of financial and labor burden across healthcare systems in the United States. These are estimated to account for up to 15%-25% of national healthcare expenditure, amounting to nearly \$1 trillion dollars annually.¹ As healthcare costs continue to rise, it will be essential for healthcare institutions to optimize financial resources and reduce wasteful spending. Among healthcare

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

² Department of Neurological Surgery, Weill Cornell Medical Center- Ochs Spine Hospital, New York, NY, USA

Corresponding Author:

Jun S. Kim, MD, Mount Sinai West, 425 West, 59th St New York, NY 10019, USA.

Email: jun.kim@mountsinai.org



Creative Commons Non Commercial No Derivs CC BY-NC-ND: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License (<https://creativecommons.org/licenses/by-nc-nd/4.0/>) which permits non-commercial use, reproduction and distribution of the work as published without adaptation or alteration, without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

administration, billing and coding related tasks are one of the primary drivers of expenditure. This primarily revolves around the accurate characterization of services, outcomes and reimbursement through Current Procedural Terminology (CPT) codes.

The CPT system was developed by the American Medical Association to provide a standard language and methodology for coding medical procedures to streamline communication across healthcare providers. The six-digit codes are assigned to relay all components of an undertaken surgical intervention; issues can arise when the same procedure has more than one CPT code. For example, a multi-level fusion procedure will have more CPT codes based off of the number of levels fused and what devices or techniques were used.

One technique that may help optimize this process is Natural Language Processing (NLP). NLP is a sub-domain of machine learning focused on the analysis of free-text. The technique has garnered significant interest in clinical medicine owing to the vast quantity of medical documentation, approximately 80% of which is in the form of unstructured text.² We previously developed a robust machine learning model to generate automated CPT codes in spine surgery to near-human accuracy using random forest and deep learning models.³ We now present a second iteration using a newer and more powerful model called XLNet, a bidirectional Long Short-Term Memory (LSTM) model that outperformed the previous best model on almost all of the basic NLP tasks, including sentiment analysis and question answering.⁴ The strength of this model comes from the amount of semantic information it can pull from a piece of text. Our objective is to show that XLNet can successfully automate the generation of CPT codes from operative notes in ACDF, PCDF, and CDA procedures.

Methods

Study Design and Setting

This study is a retrospective cohort study. We used a single natural language processing model (XLNet) to predict CPT codes from billing note data. All ethical regulations and concerns for patients' privacy were followed during this study. The IRB approved the present study and granted waiver for consent of patient data. The protocol number is AAAS8683.

Participants and Data Sources

922 operative notes were collected from patients who either underwent elective Anterior Cervical Discectomy and Fusion (ACDF) (n = 389), Posterior Cervical Decompression and Fusion (PCDF) (n = 223), combined PCDF and ACDF approach (n = 158), or Cervical Disc Arthroplasty (CDA) (n = 152) from 2015 to 2020 written by three spine surgeons.

CPT codes were generated by the billing code department. This served as our gold standard dataset with which to compare the accuracy of our model. Identifying information was removed and

the operative notes and labels (the associated CPT codes) were entered into our dataset.

Model Design and Data Cleaning

Preprocessing in the form of removing stop words, words that have little semantic value such as "and" and "the," and lemmatization (standardizing variations of the same word by grouping them via replacing plural forms etc.), was performed using the Natural Language Toolkit (NLTK) python package,⁵ and CPT codes with less than 50 operative notes were removed from the labels.

XLNet is a generalized autoregressive pretraining method that has managed to outperform the previous top performing model, BERT, in most of the basic NLP tasks. It functions by first generating a deep representation of the operative note. A bi-directional Long Short-Term Memory (LSTM) layer is then applied to develop a sense of sequential data within the operative note.

XLNet is pre-trained on a plethora of common NLP corpuses such as Wikipedia and BookCorpus. However, these are incredibly generalized and would perform poorly if used on such specific tasks. As a result, we needed to fine-tune our model on our operative note dataset.

Data was randomized with a train-validation set of 70% and a blinded test set of 30%. A grid search was performed to optimize the hyperparameters. We included batch size (of 8, 16, 32, 48, and 64 notes), learning rate (of 1e-2, 1e-3, 2e-3, 3e-3, 1e-4, and 1e-5), and epochs (of 5, 10, and 15). All trials were run with 5-fold cross validation. We found a batch size of 48, learning rate of 1e-5, and epochs of 10 performed the best.

The base sized XLNet model was fine-tuned on this data and used to generate CPT codes which were compared with our department generated codes.

Statistical Analysis

Our data was analyzed using the XLNet library, NLTK library, Numpy, and Pandas.^{4,5} All analyses were performed using Python 3.8.9.

We measured our model's performance using a receiver-operating characteristic analysis and calculating the area under the receiver-operating curve. The area under the receiver-operating curve tells how much a model is capable of distinguishing between classes. Additionally, in the case where you have a situation with a mix of common CPT codes and uncommon CPT codes, predictive algorithms that maintain good positive predictive value without sacrificing sensitivity are challenging to develop. To evaluate this, areas under the precision-recall curves were generated.

Four trials were performed on our dataset: one with ACDF operative notes only (n = 389), another with PCDF notes only (n = 223), one with ACDF and CDA notes (n = 541), and one with all the operative notes (ACDF only, PCDF only, CDA only, and combined PCDF and ACDF) we collected (n = 922).

Our results were measured using three metrics:

1. Accuracy - This was reported as a percentage from 0-100% for each CPT code included in the trial, as well as an average accuracy. This is simply how often the model was able to correctly predict that a specific CPT code should be assigned to an operative note. This cannot be the only metric used to evaluate a machine learning model, however, since it is limited by the nature of the data itself. For example, in a case where 90% of the data is ACDF operative notes and the other 10% are CDA operative notes, the model can achieve 90% accuracy by incorrectly adding ACDF CPT codes to 100% of the notes.
2. Area Under the Receiver-Operating Curve (AUROC) - This was reported as an averaged score from 0 to 1 across all CPT codes, where a higher score represents a better ability to differentiate between the presence or absence of a CPT code in a given operative note (whereas accuracy simply tells us how good the model is at detecting the presence of a CPT code).
3. Area Under the Precision Recall curve - This was also reported as a score from 0 to 1. A high score here indicates

that the model has both good precision and recall. Precision is a representation of quality, measuring how many more relevant predictions were made compared to irrelevant. Recall represents quantity, measuring how many predictions were relevant in total. In this case, a model with high precision will return few CPT codes for a given operative note but most will be correct, whereas a model with high recall will return many CPT codes for a given operative note but many may be incorrect. Therefore, a high scoring model will return many results, and most will be correct.

Heatmap Generation

Data was visualized using an attention heatmap with a darker red color corresponding to greater attention being placed on the word. Attention matrices for specific operative notes were pulled from the model and normalized to generate these figures.

Results

Trial 1 - ACDF Operative Notes Only

389 operative notes were collected of patients who underwent elective ACDF procedures. The overall AUROC

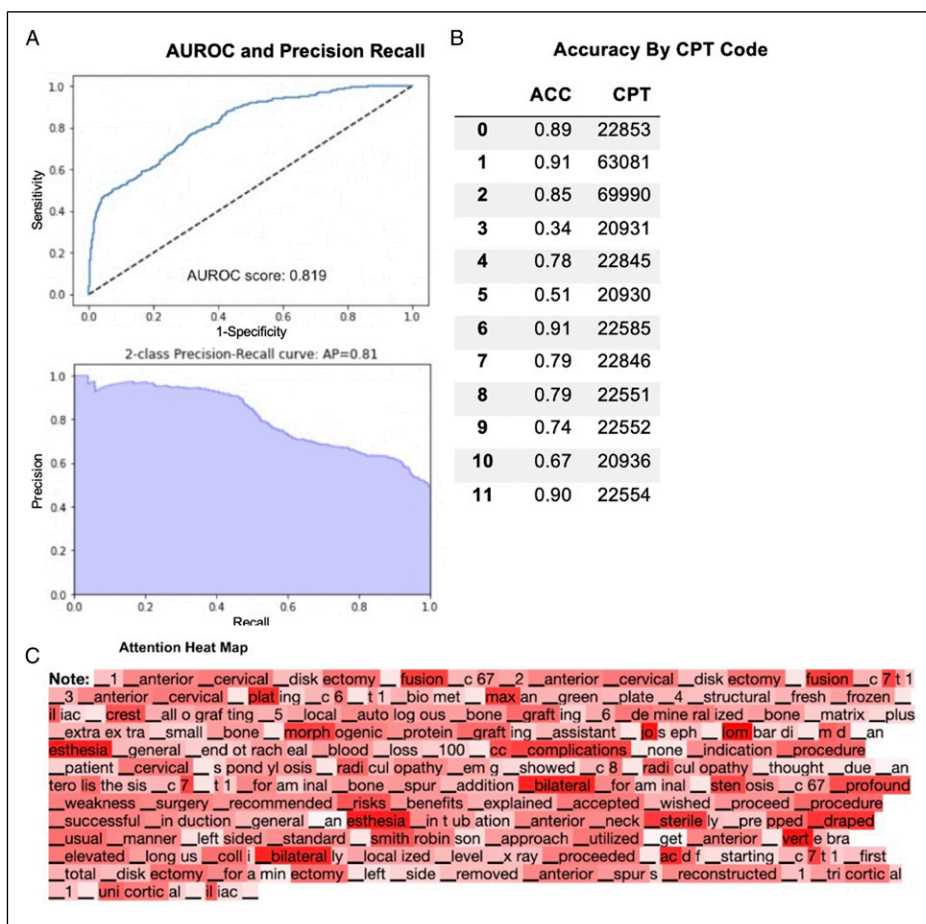


Figure 1. Performance of fine-tuned XLNet model on ACDF clinical operative notes only. (A) AUROC and precision recall. (B) Accuracy by CPT Code. (C) Attention heat map.

score was .82 (range: .48-.93), and the 2-class precision-recall average precision was .81 (range: .45-.97) (Figure 1A). The average class-by-class accuracy was 77% (range: 34%-91%) for the generated CPT codes (Figure 1B). Data was visualized using an attention heatmap with a darker red color corresponding to greater attention being placed on the word (Figure 1C).

Trial 2 - PCDF Operative Notes Only

223 PCDF operative notes were collected and analyzed. The overall AUROC score was .83 (range: .44-.94), and the 2-class precision-recall average precision was .70 (range: .45-.96) (Figure 2A). The average class-by-class accuracy was 71% (range: 42%-93%) for the generated CPT codes (Figure 2B).

Trial 3 - ACDF and CDA Operative Notes

A total of 541 ACDF and CDA Operative Notes were collected for this trial. The overall AUROC score was .95 (range: .68-.99), and the average area under the 2-class precision-recall curve was .91 (range: .56-.98) (Figure 3A). The average class-by-class accuracy was 87% (range: 63%-99%) for the generated CPT codes (Figure 3B).

Trial 4 - All Operative Notes

A total of 922 ACDF, PCDF, and CDA Operative Notes were collected for this trial. The overall AUROC score was .95 (range: .76-.99), and the average area under the 2-class precision-recall curve was .84 (range: .49-.99) (Figure 4A). The average class-by-class accuracy was 88% (range: 70%-99%) for the generated CPT codes (Figures 4B and 5).

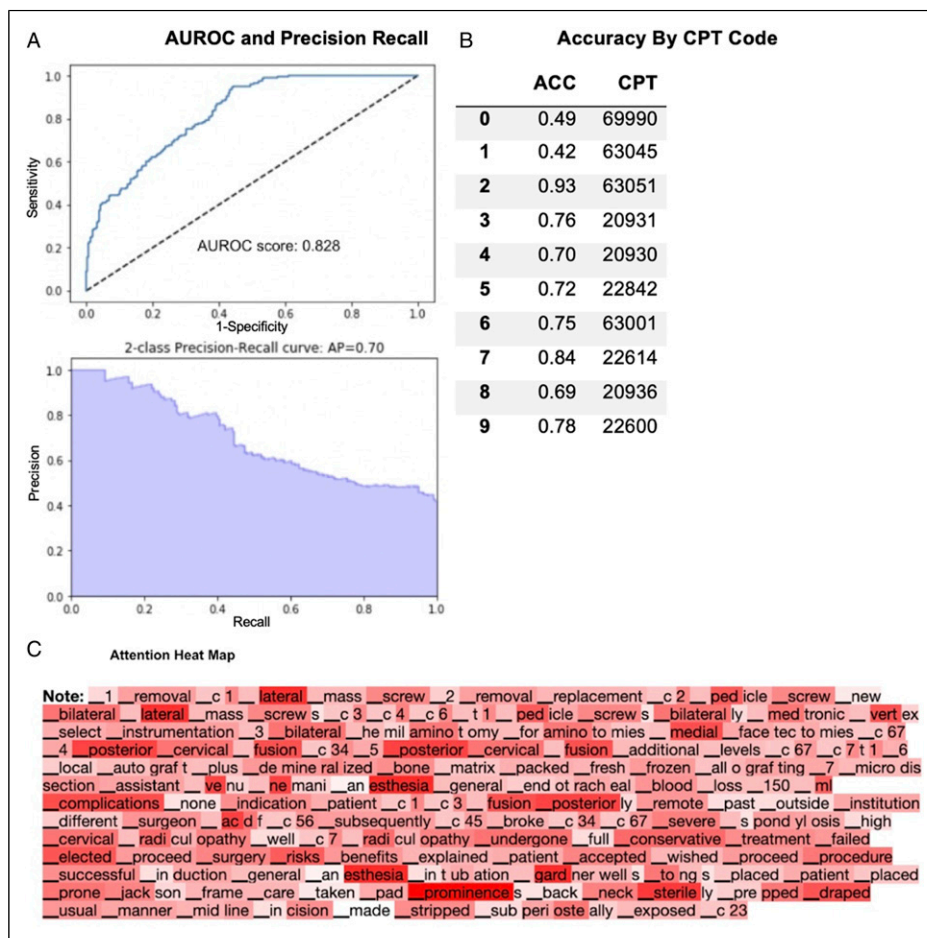


Figure 2. Performance of fine-tuned XLNet model on PCDF clinical operative notes only. (A) AUROC and precision recall. (B) Accuracy by CPT Code. (C) Attention heat map.

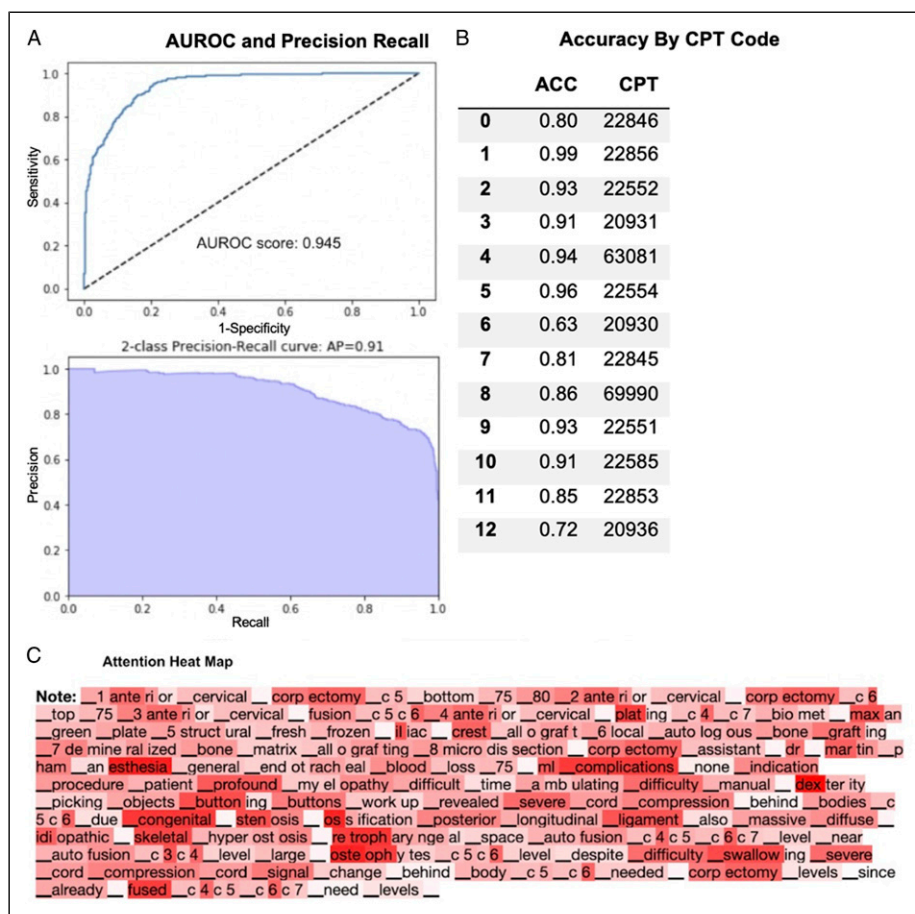


Figure 3. Performance of fine-tuned XLNet model on ACDF and CDA clinical operative notes only. (A) AUROC and precision recall. (B) Accuracy by CPT Code. (C) Attention heat map.

Discussion

Summary, Background, Significance

In this study, we developed and validated a machine learning model for the automated generation of CPT codes for not only ACDF, but also PCDF and CDA procedures. This study represents the second iteration of our previous pilot study, which only looked at ACDF CPT codes and used a random forest with a bag of words approach and LSTM model. Key points of improvement include the use of a more recent and advanced algorithm and the inclusion of PCDF and CDA. Overall, the XLNet model demonstrated higher performance and was broadened to include posterior fusion and cervical disc replacement.

Healthcare costs continue to rise annually, placing significant financial strain upon hospitals and provider groups across the United States. Efforts to control costs have seen the rise of value-based reimbursement models, which incentivize efficient resource utilization over volume. It will be imperative for medical institutions to optimize all parts of the healthcare delivery process. Approximately 62% of administrative-related costs are attributable to billing and coding.^{6,7} Some estimates suggest that it may

consume up to 14% of revenue for a physician group and thus represents a significant portion of overhead costs.⁸ The cost of these tasks is particularly high in the United States when comparing globally. One study by Morra et al showed that US physicians spent nearly 4 times more in billing and administration related tasks than corresponding Canadian physicians.⁹ With comparable administrative costs to Canada, the authors estimated an annual savings of \$27.6 billion in the United States.⁹ Automation can lessen the required number of support staff, leading to reduced labor costs. Moreover, automated generation of CPT codes can also help with improving coding accuracy and reduce error rates. Inaccuracies in billing can lead to lost revenue (ie: downcoding) or potentially fraudulent claims (ie: upcoding). According to the Centers of Medicaid and Medicare services (CMS), \$95 billion issued payments were related to fraudulent billing in 2016.^{10,11} An automated algorithm can serve as a second layer of reinforcement against miscoding.

Research into the application of machine learning for automated code generation is a recent undertaking in the medical field with few studies conducted on the topic. To the authors' knowledge, ours is the second study demonstrating the use of machine learning for orthopedic surgery CPT code

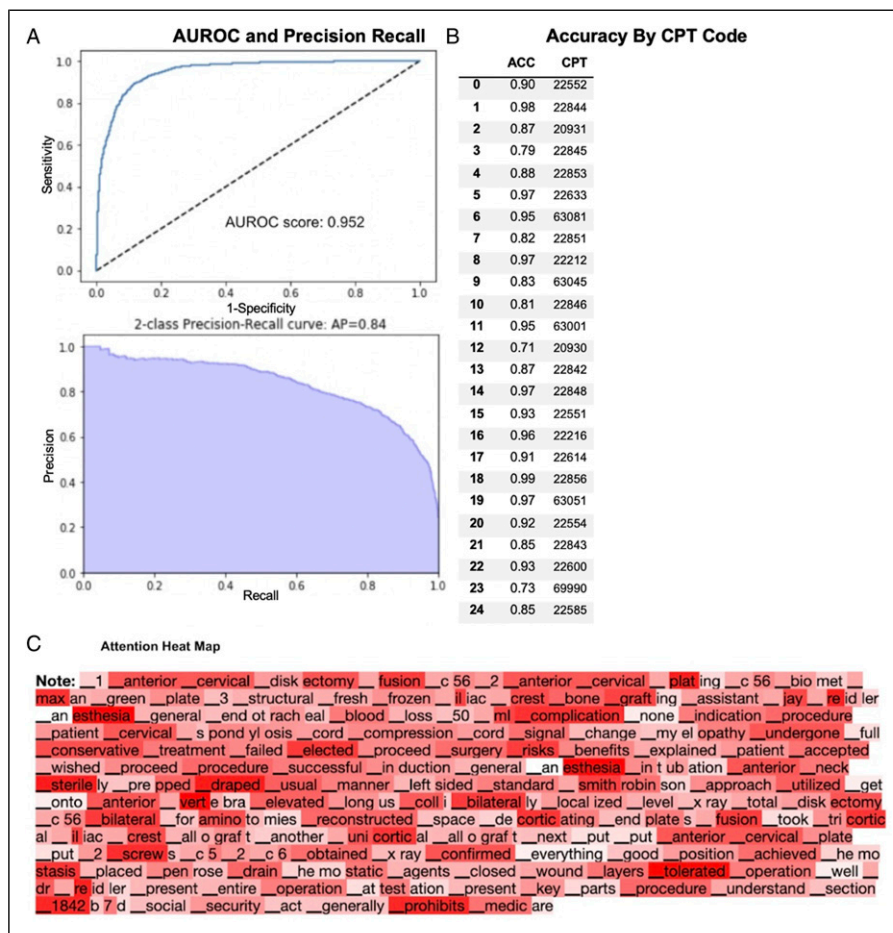


Figure 4. Performance of fine-tuned XLNet model on ACDF, PCDF, and CDA clinical operative notes only. (A) AUROC and precision recall. (B) Accuracy by CPT Code. (C) Attention heat map.

generation through operative notes, and the first to use an advanced NLP model such as XLNet. Two previous studies have applied natural language processing machine learning algorithms to CPT code generation via pathology reports¹² and electronic health data.¹³ Levy et al reported high accuracy of CPT generation from pathology reports using the XGBoost and BERT models and 5 primary CPT codes¹² while Burns et al reported good accuracy of CPT generation from electronic health records using support vector machines and neural network label-embedding attentive models and 285 primary CPT codes.¹³ We add to this small but growing body of literature by analyzing an alternative source of CPT code generation – surgeon operative notes – and using a novel NLP model – XLNet – with comparable accuracies to the previous studies.

Model Performance

The model performed well in all the trials, but especially in the combined ACDF and CDA and all Operative Notes trials, which had larger datasets and more varied CPT codes. The

lower AUROC and area under the precision recall curves seen in Trials 1 and 2 is likely due to the model overfitting the data. Because the data was homogenous, with only one procedure type being presented, it aligned itself too closely to those points. This resulted in a failure to generalize properly. However, in trials 3 and 4, this problem seemed to diminish as we added in not only more data, but also less homogenous data. The greater contextual understanding afforded by models such as XLNet allows it to flourish with tasks such as this.

Words such as “posterior” and “anterior” are given a lot of attention, which makes sense given how important they would be when distinguishing between CPT codes. The cervical vertebra numbers were also given attention as well as the ‘t’ in ‘t1’, likely because the model would use that to decide when to add the CPT codes for each additional level after the first level.

Comparison to Previous Model

Compared to our pilot study, we saw important improvements. The previous long short-term memory (LSTM) model had an AUROC score of .72 and an area under the precision curve of .44.

CPT Codes	Counts
22552	253
22844	83
20931	539
22845	316
22853	172
22633	61
63081	146
22851	75
22212	68
63045	80
22846	140
63001	102
20930	355
22551	425
22216	93
22614	193
22856	113
63051	89
22554	163
22843	67
22600	138
69990	188
22585	162
20936	400

Figure 5. CPT counts across all operative notes.

Compared to the performance of the LSTM model for only ACDF operative notes, the XLNet model performed much better across all metrics for all trials. However, when compared to the random forest with bag-of-words approach, the random forest performed better by about .1 with an AUROC score of .94 and area under the precision-recall curve of .85. This is likely due to the problem discussed previously with overfitting, since we achieve better or similar scores when there is more data that is more varied. Moreover, a model such as XLNet utilizes bi-directional LSTM layers, which take into account the context of words around the word of interest, and store it in a sort of memory, which is advantageous when data is heterogeneous.

Our results are consistent with or outperform other deep learning models in code generation from clinical notes.¹⁴ The higher AUROC and area under the precision-recall curve demonstrate that the LSTM model performs better when given a larger

dataset. Our use of unstructured operative notes shows that these models can understand and even perform similarly to humans when generating billing codes.

Benefits

XLNet has been lauded as the latest and greatest model for understanding language as it outperforms the previous state-of-the-art model, BERT.⁴ Already, XLNet has been shown to outperform previous models in predicting prolonged mechanical ventilation.¹⁵ Unstructured clinical data has great potential compared to current methods of structuring clinical data, since it allows for a more holistic overview without loss of data in the process, such as temporal information. However, the extraction of this unstructured data is costly, whether in the time it takes a physician or billing team to sort through it, or the computation cost involved in training a machine learning model from scratch.

Each physician has their own style of writing notes, so it makes sense that structured methods of understanding operative notes would be a poor approach. This is especially true for billing. Automated generation of billing codes may allow for greater accuracy and faster billing. NLP models, in particular, are helpful in discovering fraud, both in prescription and in detecting protected health information.^{16,17}

Despite our limited sample size, we still managed to find examples of when the model identified notes that were missing from the ground truth. For example, in the operative note in [Figure 6](#), the model correctly identified the CPT code 20936, which codes for an autograft, that was missed by the billing department.

Limitations

A major limitation of this study was the size of the dataset. When compared to corpora (large text bodies used to train the XLNet Base Model) such as the Wikipedia corpus, which has millions of entries, our sample size of about 900 operative notes is very limited. This discrepancy is especially seen in the attention heatmaps, as certain phrases or words like “sterile,” that are seemingly irrelevant, are given attention that is likely a vestige of the previous training, since we are only fine-tuning over a few iterations. These words are also more common in terms of everyday usage, so the pretrained model is more familiar with them. Moreover, a more varied dataset gives the model an opportunity to learn and understand the context of words it may have never encountered in the previous corpora. We would likely see even better performance, as more data becomes available. While the source dataset was already small, we additionally limited the CPT codes to account for class imbalance in the dataset, such as having many more common codes, like 22551 for ACDF procedures, compared to those without or with rarer CPT codes, and we excluded codes with less than 50 notes. This may have restricted the model’s ability to differentiate some of the minutiae which can be seen in some of the very low accuracies and AUROC scores for CPT codes that had fewer notes. As such, the model does not

PREOPERATIVE DIAGNOSIS: QUESTIONABLE C4 COMPRESSION FRACTURE IN THE REMOTE PAST WITH NO HISTORY OF SUCH FRACTURE, BUT WITH SEVERE KYPHOSIS AT C3-4, CONGENITAL STENOSIS WITH SEVERE CORD COMPRESSION C4-6 DUE TO HERNIATED DISKS AND OSTEOPHYTES AND CORD COMPRESSION ALSO AT C3-4 AND C6-7 WITH MYELOPATHY. POSTOPERATIVE DIAGNOSIS: QUESTIONABLE C4 COMPRESSION FRACTURE IN THE REMOTE PAST WITH NO HISTORY OF SUCH FRACTURE, BUT WITH SEVERE KYPHOSIS AT C3-4, CONGENITAL STENOSIS WITH SEVERE CORD COMPRESSION C4-6 DUE TO HERNIATED DISKS AND OSTEOPHYTES AND CORD COMPRESSION ALSO AT C3-4 AND C6-7 WITH MYELOPATHY.

OPERATION: 1. ANTERIOR CERVICAL CORPECTOMY OF C5. 2. ANTERIOR CERVICAL FUSION, C4-5. 3. ANTERIOR CERVICAL FUSION ADDITIONAL LEVEL, C5. 4. ANTERIOR CERVICAL DISKETOMY AND FUSION AT C3-4. 5. ANTERIOR CERVICAL DISKETOMY AND FUSION, ADDITIONAL LEVEL, C6-7. 6. ANTERIOR CERVICAL PLATING C4 THROUGH 7 WITH A BIOMET MAXAN GREEN PLATE. 7. STAND-ALONE BIOMET SOLITAIRE CAGE AT C3-4, 18 X 15 X 8 MM TALL. 8. STRUCTURAL FRESH FROZEN ILIAC CREST ALLOGRAFTING. 9. LOCAL AUTOLOGOUS BONE PLUS DEMINERALIZED BONE MATRIX PLUS A SMALL BMP GRAFTING. 10. MICRODISSECTION FOR CORPECTOMY. ANESTHESIA: General endotracheal. BLOOD LOSS: 100 mL.

COMPLICATIONS: None. INDICATION PROCEDURE: After successful induction of general anesthesia intubation, the anterior neck was sterilely prepped and draped in the usual manner. A left-sided standard Smith-Robinson approach was utilized to get down to the anterior vertebra. We elevated the longus colli bilaterally, localized the level with an x-ray, and started out at the C3-4 level. We took out the disk in its entirety and shaped the C4 vertebra as it had likely a compression fracture in the remote past. We thoroughly decompressed the central canal, went through the posterior longitudinal ligament to visualize the dura, took down the posterior osteophytes, corrected the kyphosis, and then we reconstructed with a Biomet Solitaire C 8 mm tall, 18 x 15 mm, and this was packed with part of a small BMP along with demineralized bone matrix, and cancellous allograft, local autograft. We then turned our attention to doing the C5 corpectomy and we made a width that was about 18 mm wide and left a thin wedge of bone posteriorly, but otherwise thoroughly decompressed both the C4-5 and the C5-6 areas and we left just a small amount of uncinat and lateral wall on the right side, but the left side was taken all the way out to just the medial portion of the foramen transverse area. At the conclusion of the corpectomy we saw nice dural pulsations. We then reconstructed with a tricortical iliac crest allograft as a strut graft to bridge from C4 down to C6. We packed this with part of a small BMP, tamped it in place, put the corpectomy bone next to it and then we put in some uncortical allograft fresh frozen into the remaining uncinat on the right side and then turned our attention to C6-7 where we did a total diskectomy, went through the posterior longitudinal ligament, and we made sure that everything was well decompressed, and then we reconstructed with 2 tricortical iliac crest allografts packed with remaining BMP. We added demineralized bone matrix and packed the Schmorf's node bony defect with this demineralized bone matrix and corpectomy bone and then we took the fresh frozen iliac crest allograft, put that in, and we took an anterior cervical plate Biomet MaxAn Green plate, put 2 screws into C4, 2 into C6, 2 into C7, put 1 screw on the lateral wall of C5, obtained a fluoroscopic image that showed that everything was in good position. We achieved hemostasis, placed in hemostatic agents, Depo-Medrol, antibiotic powder, and a drain, and closed the wound in layers.

True CPT Codes: 63081, 69990, 22551, 22552, 22554, 22585, 22846, 22853, 20931

Predicted CPT Codes: 63081, 69990, 22551, 22552, 22554, 22585, 22846, 22853, 20931, 20936

Figure 6. Sample operative note with predicted output.

incorporate all possible procedure types through the included CPT codes. Moreover, the model only has knowledge of specific pre-selected procedures, and the performance observed here may not hold for other CPT codes not tested.

Another limitation of our dataset is the lack of available surgeons' coding. While we have demonstrated the ability of XLNet to achieve similar accuracy to human coders with small error margins, it is likely that significant errors exist among the human coders themselves. A true measure of the accuracy of XLNet would compare the generated codes to those of the surgeons and determine how this error rate compares to that of human coders and surgeon codes. While such data was not available to us, our present study still contributes important results for the future of machine generated coding.

Future studies could benefit from increasing the dataset size by increasing the number of operative notes or the variability of included

CPT codes and obtaining surgeon codes for accuracy analysis. Further limitations arise from the structure of the XLNet model, which is designed to efficiently ingest unstructured data, but we did not include any structured data. An improved model architecture that could ingest both forms of data would greatly increase the size of the dataset and could improve the performance of the model. Moreover, our dataset may be biased, as it comes from a large academic hospital system and may incur some of the biases involved in such institutions.

Conclusion

Our study of the utility of natural language-processing based machine-learning techniques for automated CPT coding reveal promising results, especially given the limitations in dataset size. The combined ACDF + CDA trial and all operative notes combined trials achieved results nearing that of human accuracy

with AUROC and area under the precision-recall curve values of .95 and .91 (trial 3) and .95 and .84 (trial 4), respectively. The success of our early efforts in NLP-ML assisted CPT coding strongly suggests that future work on these methods may lead to further enhancements in the accuracy of the model. These methods have the potential to greatly reduce the human time costs of manual CPT coding.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: JK, Stryker: Paid consultant, SK, FAAOS; AAOS: Board or committee member; American Orthopaedic Association: Board or committee member; AOSpine North America: Board or committee member; Cervical Spine Research Society: Board or committee member; Globus Medical: IP royalties; North American Spine Society: Board or committee member; Scoliosis Research Society: Board or committee member; Stryker: Paid consultant; KR, FAAOS; America: Stock or stock Options; AOSpine: Board or committee member; AxioMed: Stock or stock Options; Benvenue: Stock or stock Options; Biomet: IP royalties; Paid presenter or speaker; Clinics in orthopedics: Editorial or governing board; European Spine Journal: Editorial or governing board; Expanding Orthopedics, PSD: Stock or stock options; Global Spine Journal: Editorial or governing board; HAPPE Spine: Unpaid consultant; Medtronic: Paid presenter or speaker; Neurosurgery: Editorial or governing board; North American Spine Society: Board or committee member; Nuvasive: Paid consultant; Paid presenter or speaker; Paradigm Spine: Stock or stock Options; Spinal Kinetics: Stock or stock Options; Spine: Editorial or governing board; Spine Surgery Today: Editorial or governing board; Spineology: Stock or stock Options; Vertiflex: Stock or stock Options. The following individuals have no conflicts of interest or sources of support that require acknowledgement: BZ, JT, VA, BC, AD, EG, CD.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Bashar Zaidat  <https://orcid.org/0000-0002-8823-720X>
 Justin Tang  <https://orcid.org/0000-0003-4544-3262>
 Eric A. Geng  <https://orcid.org/0000-0003-0736-3245>
 Akiro H. Duey  <https://orcid.org/0000-0003-0308-0512>
 Calista Dominy  <https://orcid.org/0000-0002-6572-585X>
 Kiehyun D. Riew  <https://orcid.org/0000-0002-0718-3423>

References

- Chernew M, Mintz H. Administrative expenses in the US health care system: Why so high? *JAMA*. 2021;326(17):1679-1680. doi:10.1001/jama.2021.17318.
- Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. *Yearb Med Inform*. 2014;9(1):14-20. doi:10.15265/IY-2014-0020.
- 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? *Global Spine J*. 2022;21925682211062831. doi:10.1177/21925682211062831.
- Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. Advances in Neural Information Processing Systems. 2019. doi:10.48550/ARXIV.1906.08237.
- Bird S, Edward L, Ewan K. *Natural Language Processing With Python*. 2009, O'Reilly Media, Inc.
- 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):691-697. doi:10.1001/jama.2017.19148.
- 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.
- Sakowski JA, Kahn JG, Kronick RG, Newman JM, Luft HS. Peering into the black box: Billing and insurance activities in a medical group. *Health Aff*. 2009;28(4):w544. doi:10.1377/hlthaff.28.4.w544.
- 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.
- CMS Needs to Fully Align its Antifraud Efforts With the Fraud Risk Framework*. U.S. Government Accountability Office. 2017.
- Drabiak K, Wolfson J. What should health care organizations do to reduce billing fraud and abuse? *AMA J Ethics*. 2020;22(3):E221-E231. doi:10.1001/amajethics.2020.221.
- 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. doi:10.4103/jpi.jpi_52_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;132:738-749. doi:10.1097/ALN.0000000000003150.
- Huang J, Osorio C, Sy LW. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Comput Methods Programs Biomed*. 2019;177:141-153. doi:10.1016/j.cmpb.2019.05.024.
- Huang K, Singh A, Chen S, et al. *Clinical XLNet: Modeling Sequential Clinical Notes And Predicting Prolonged Mechanical Ventilation*. arXiv. 2019. arXiv preprint arXiv:1912.11975.
- Oh SH, Kang M, Lee Y. Protected health information recognition by fine-tuning a pre-training transformer model. *Health Inform Res*. 2022;28(1):16-24.
- 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.