

# Can Natural Language Processing and Artificial Intelligence Automate The Generation of Billing Codes From Operative Note Dictations?

Global Spine Journal  
 2023, Vol. 13(7) 1946–1955  
 © The Author(s) 2022  
 Article reuse guidelines:  
[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)  
 DOI: 10.1177/21925682211062831  
[journals.sagepub.com/home/gsj](https://journals.sagepub.com/home/gsj)



Jun S. Kim, MD<sup>1</sup> , Andrew Vivas, MD<sup>2</sup>, Varun Arvind, BS<sup>1</sup>, Joseph Lombardi, MD<sup>3</sup>, Jay Reidler, MD<sup>4</sup>, Scott L Zuckerman, MD<sup>3</sup>, Nathan J. Lee, MD<sup>3</sup> , Meghana Vulapalli, BS<sup>5</sup> , Eric A Geng, BS<sup>1</sup> , Brian H. Cho, BS<sup>1</sup>, Kazuaki Morizane, MD, PhD<sup>6</sup>, Samuel K. Cho, MD<sup>1</sup> , Ronald A. Lehman, MD<sup>3</sup>, Lawrence G. Lenke, MD<sup>3</sup>, and Kiehyun Daniel Riew, MD<sup>5</sup>

## Abstract

**Study Design:** Retrospective Cohort Study.

**Objectives:** Using natural language processing (NLP) in combination with machine learning on standard operative notes may allow for efficient billing, maximization of collections, and minimization of coder error. This study was conducted as a pilot study to determine if a machine learning algorithm can accurately identify billing Current Procedural Terminology (CPT) codes on patient operative notes.

**Methods:** This was a retrospective analysis of operative notes from patients who underwent elective spine surgery by a single senior surgeon from 9/2015 to 1/2020. Algorithm performance was measured by performing receiver operating characteristic (ROC) analysis, calculating the area under the ROC curve (AUC) and the area under the precision-recall curve (AUPRC). A deep learning NLP algorithm and a Random Forest algorithm were both trained and tested on operative notes to predict CPT codes. CPT codes generated by the billing department were compared to those generated by our model.

**Results:** The random forest machine learning model had an AUC of .94 and an AUPRC of .85. The deep learning model had a final AUC of .72 and an AUPRC of .44. The random forest model had a weighted average, class-by-class accuracy of 87%. The LSTM deep learning model had a weighted average, class-by-class accuracy Of 59%.

**Conclusions:** Combining natural language processing with machine learning is a valid approach for automatic generation of CPT billing codes. The random forest machine learning model outperformed the LSTM deep learning model in this case. These models can be used by orthopedic or neurosurgery departments to allow for efficient billing.

## Keywords

machine learning, natural language processing, long short term, memory, random forest model learning, billing, surgery, current procedural terminology codes, coding

<sup>1</sup>Department of Orthopedics, Icahn School of Medicine at Mount Sinai, New York, NY, USA

<sup>2</sup>Department of Neurological Surgery, UCLA Medical Center, Los Angeles, CA, USA

<sup>3</sup>Department of Orthopedics, Columbia University Irving Medical Center- Och SpineHospital, New York, NY, USA

<sup>4</sup>Department of Orthopedics, University of Pennsylvania, Philadelphia, PA, USA

<sup>5</sup>Department of Neurological Surgery, Weill Cornell Medical Center- Och Spine Hospital, New York, NY, USA

<sup>6</sup>Department of Orthopedics, Hayashi Hospital, Echizen, Japan

## Corresponding Author:

Jun S. Kim, MD, Department of Orthopedics, Icahn School of Medicine at Mount Sinai 425 West 59th Street, 5th Floor, New York, NY 10019, USA.

Email: [Jun.Kim@mountsinai.org](mailto: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>).

## Introduction

Natural language processing is a subfield of machine learning useful for processing free text. Because 80% of the electronic medical record is recorded in unstructured text, natural language processing is a valuable resource for synthesizing this data.<sup>1</sup> Applications include radiology note generation and data extraction from imaging reports.<sup>2,3</sup> Machine learning models have also been used to predict Current Procedural Terminology (CPT) codes in anesthesiology and surgical pathology.<sup>4,5</sup> Accurate generation of these codes has become critical, as the last 2 decades have seen a shift to using ICD-10 and CPT codes for characterizing admissions, outcomes, and justifying reimbursement.

To our knowledge, no prior studies have examined the use of machine learning algorithms for generating automated CPT codes in spine surgery. In this study, we sought to validate which natural language processing-based machine learning model is better able to predict CPT codes in spine surgery from operative dictations. This study also serves as a pilot to determine, with a larger sample size, if machine learning algorithms achieve human accuracy in predicting CPT codes within spinal surgery.

## Methods

### Study Design and Setting

This study is a retrospective cohort study for the development and validation of a singular machine learning model to predict CPT codes from operative dictations. All ethical regulations and concerns for patients' privacy were followed during this study. The Columbia University Irving Medical Center Institutional Review Board approved the present study and granted waiver for consent of patient data.

### Participants and Data Sources

This was a retrospective analysis of operative notes from a large, single-center academic institution's database of cases from a single senior surgeon. Inclusion criteria included consecutive patients who underwent elective anterior cervical spine surgery by a senior surgeon from 9/2015 to 1/2020. Analysis was limited to CPT codes that appeared more than 50 times to limit the impact of uncommon codes in this relatively smaller dataset.

Gold standard labels (i.e., CPT codes) for each case were obtained via the billing department under the supervision of a spine focused, senior billing administrator. Operative notes were entered into the dataset once all identifying patient information was removed from the original operative note and after gold standard labels were obtained. The de-identification process did not affect the natural text since no identifying information is used within the main dictation of the operative note.

### Analysis Platform

All analyses were performed on secure computer clusters. Code was written in Python3 using numpy, pandas, and scikit-learn libraries. Code can be found at GITHUB REPO.

### Supervised Learning of Current Procedural Terminology Code Prediction

Natural language processing is a machine learning technique that allows an algorithm to learn from words. We trained 2 algorithms on operative notes to determine which approach was better. The first algorithm was a more modern, deep neural network approach called a long short term memory network. Long short term memory networks take in data that comes in a series such as words or data that is linearly oriented in time. They have "memory" in that they use data from the past to make predictions about the future. For instance, words that appear earlier in the sentence or paragraph help make predictions about words that come later. In another example, these networks have been applied to stock prices where historical stock prices help predict future stock prices. We compared this more modern approach to another machine learning algorithm called a random forest model that is essentially a classification technique that uses layered decision trees. The idea behind a tree is to search for a feature-value pair (i.e., word-CPT code pair) within the data and split it in such a way that will generate the "best" 2 child subsets. By the end of training, the algorithm has learned how to map a set of features (words) to targets (CPT codes). A real-life analogy to this would be if a young surgeon were trying to decide the best approach to cervical myelopathy patient. The young surgeon may ask their mentor or colleague for advice. Their mentor may ask them details about the patient's history, exam, medical conditions, and imaging. Based on the answers and "rules" to guide his decision, the mentor gives the young surgeon some advice. The young surgeon may repeat this process with a few other senior surgeons and decide on the surgery that was most commonly recommended.

To train each algorithm, each operative note was accompanied by a confirmed list of CPT codes. Both algorithms were tasked with "reading" each operative note and then learning the CPT codes associated with that note. We trained both algorithms over multiple iterations to reinforce the association between an operative note and a collection of CPT codes. Once we believed that training was complete, we tested both models or algorithms on a collection of notes that each had never "read" before. This allowed us to evaluate the performance of each model and assess how generalizable it is.

To test our model, we randomly allocated operative notes into a training and validation cohort (70%) and the remainder into a testing cohort (30%). The same feature data processing protocol was applied to both cohorts. Following optimization of the model on the training/validation cohort, the model was evaluated in a blinded fashion on the testing cohort.

### Statistical Methods

Algorithm performance was measured by performing receiver operating characteristic analysis and calculating the area under

**Note:** dict daniel k riew md admit date 08202018 discharge date procedure date 08202018 surgeon daniel k riew md preoperative diagnosis cervical spondylosis with myelopathy c4 through 7 postoperative diagnosis cervical spondylosis with myelopathy c4 through 7 procedure 1 anterior cervical corpectomy of c5 removing 75% of the vertebral body on the right side 2 anterior cervical disectomy fusion at c6-7 3 anterior cervical plating c4 through 7 with a biomet maxan green plate 4 anterior cervical fusion c4-5 5 anterior cervical fusion additional level c5-6 6 structural fresh frozen iliac crest allografting 7 local autologous bone grafting 8 demineralized bone matrix allograft plus extra-extra small bmp grafting 9 microdissection for corpectomy assistant joseph osorio md anesthesia general endotracheal blood loss 75 ml complication none indication for operation patient is from florida was referred to us by dr nate lowen as she had multilevel cervical spondylosis with myelopathy she had undergone full conservative treatment and having failed this elected to proceed with surgery risks and benefits explained to the patient she accepted and wished to proceed description of procedure after successful induction of general anesthesia and intubation the patient was prepped and draped in the usual manner a standard smith-robinson approach was utilized to expose the anterior cervical spine thru a left-sided approach we elevated the longus colli localized the correct level and placed in retractors after exposure we started out with a c5 corpectomy and we decompressed 75% of the vertebral body leaving a thin wedge of bone posteriorly on the right side but otherwise thoroughly decompressing the 4-5 and 5-6 levels and then we reconstructed with a struct fresh frozen iliac crest allograft to bridge from c4 down to c6 on the left side there was a little bit of room left over so we used a unicortical small wedge graft to put into the uncinate at both levels and then turned our attention to the c6-7 level where we did a thorough disectomy and put in 1 tricortical and 1 bicortical bone graft side by side along with local autograft demineralized bone matrix we used an extra- extra small bmp that was divided about 50% into the c6-7 level the other 50% at c4-5 and c5-6 that we put into small cavities that we cut into the iliac crest allograft we then took an anterior cervical plate biomet maxan green plate put 2 screws into c4 into c5 in the remaining 25% of the bone on the left side and 2 screws into c6 2 screws into c7 obtained an x-ray which confirmed that everything was in good position we achieved hemostasis put bone wax in front of the plate tissue to seal everything in vancomycin powder hemostatic agents and depo-medrol 40 mg then we put in the drain and closed the wound in layers i used the microscope for better visualization of the spinal cord and nerve roots during the entire procedure and the corpectomy was performed using microdissection techniques and principles she tolerated the operation well both dr osorio and i were present for the entire operation attestation i was present for the all key parts of the procedure i understand that section 1842b7d of the social security act generally prohibits medicare physician fee schedule payment for the services of assistants at surgery in teaching hospitals when qualified residents are available to furnish such services i certify that the services for which payment is claimed were medically necessary and that no qualified resident was available to perform the services i further understand that these services are subject to post-payment review by the medicare carrier dd 08202018 dt 08202018 mdq232723 232723750820 electronic signatures riew k daniel k md signed on 21-aug-2018 14 24 authored transcriptionist it entered on 20-aug-2018 00 00 entered last updated 21-aug-2018 14 24 by riew k daniel k md

Predicted  
[ '20931', '20936', '22585', '20930', '63081', '22856', '22858', '22846', '22551' ]  
True  
[ '20931', '20936', '22585', '20930', '63081', '22554', '22846', '22551' ]

**Figure 1.** Attention Map generated by the long short term memory model. Words highlighted in red signify words that the algorithm deemed important for generating a CPT code prediction. Darker red suggests that more weight was given to those specific words. Predicted CPT codes and the actual billed CPT are shown at the bottom.

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.

### Source of Funding

The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper. There was no financial support for this research project.

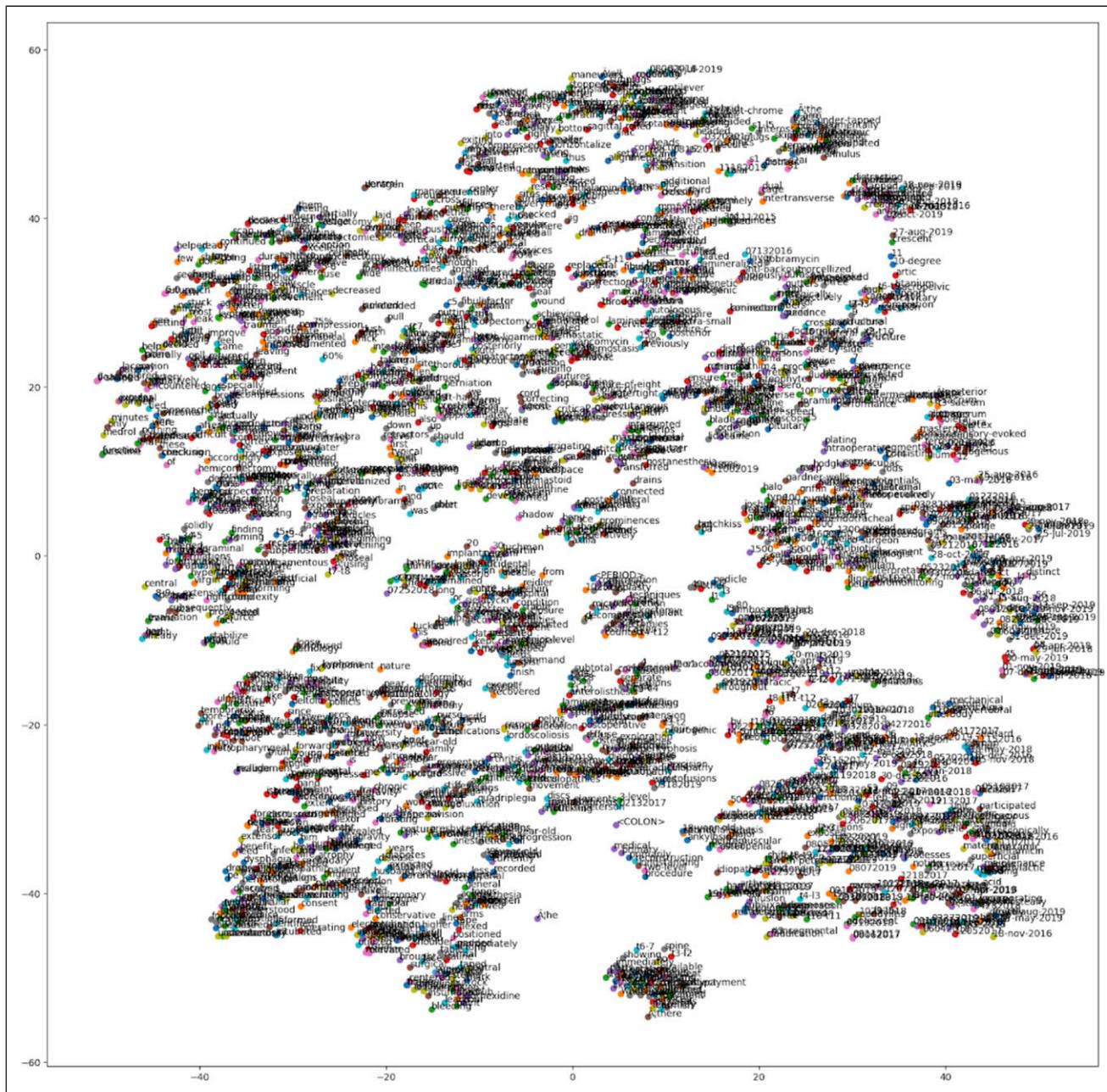
### Results

391 operative dictations fit our inclusion criteria. 15 CPT codes were incorporated into our model as these codes had an appearance in the dataset on more than 50 instances. On average, the models took .1 seconds to process a single operative note and generate a prediction.

To evaluate these 2 machine learning algorithms, we present 3 metrics:

1. Accuracy (0 to 100%)—This percentage represents the CPT codes that were predicted by the algorithm divided by the CPT codes that were verified by a senior coder and subsequently sent out for billing. Accuracy by itself can be a misleading metric to evaluate the performance of an algorithm. For example, we present a hypothetical clinical scenario where a predictive algorithm is trying to distinguish patients with or without cancer. If the real rate of cancer is 1% then a predictive model that predicts “no cancer” will still be correct 99% of the time (accuracy = 99%). In the case where 100 patients were evaluated by this algorithm, then the positive predictive value is 0% and the sensitivity is 0%. Therefore, two other statistical metrics were calculated.
2. Area under the receiver operating curve (0 to 1)—This represents the ability of the model to classify or differentiate between CPT codes. The higher the number, the better the model is at predicting the presence or absence of a CPT code after it has read an operative note.
3. Area under the precision-recall curve (0 to 1)—This number represents the ability of the algorithm to maintain good positive predictive value and good sensitivity.

When evaluated on the test set, the random forest model with a bag-of-words approach performed with an area under



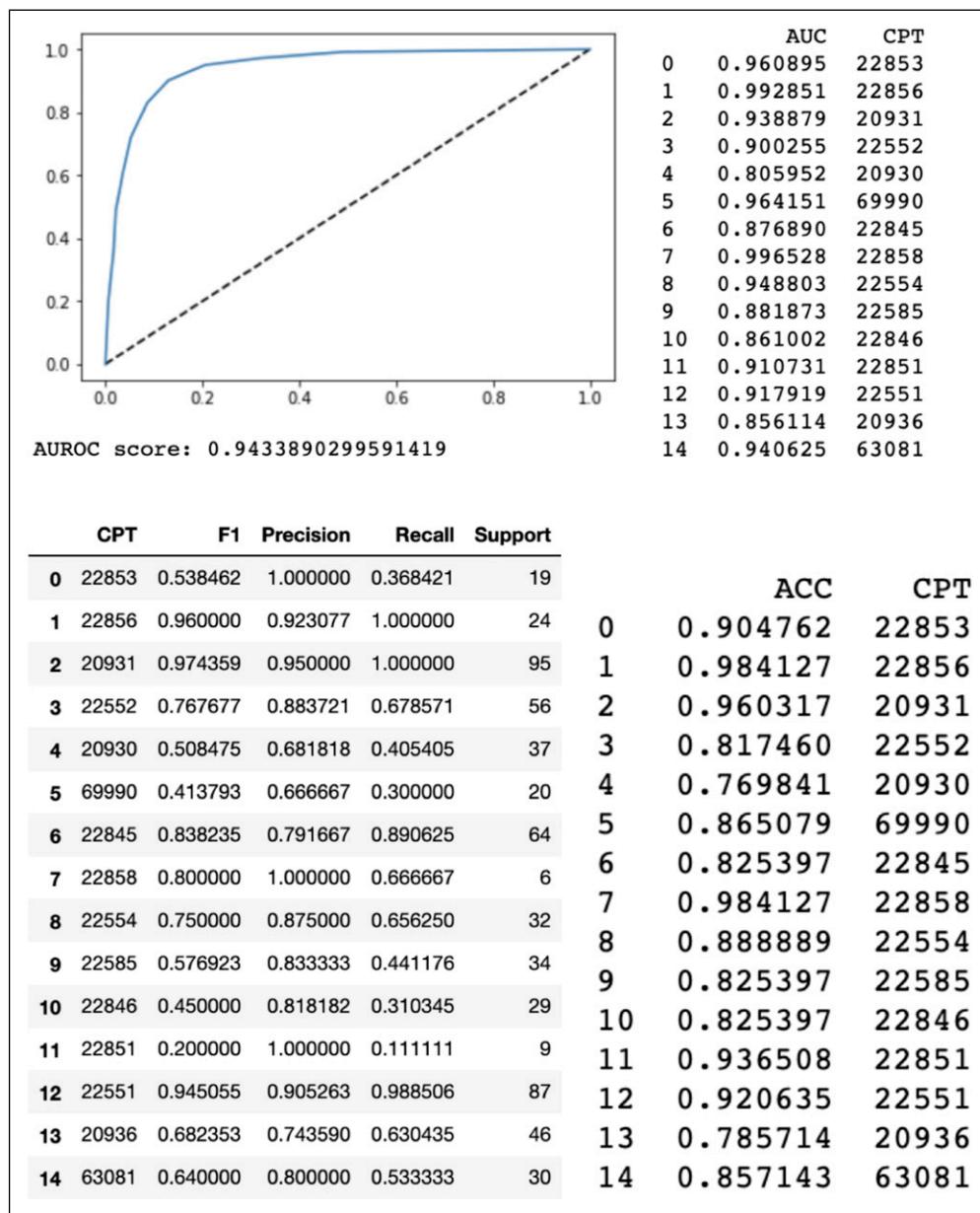
**Figure 2.** Word embedding generated via the word2vec algorithm plotted on a t-sne plot. Words with similar semantic meaning cluster together. This is an essential preprocessing step that turns words into number vectors.

the receiver operating curve of .94 and area under the precision-recall curve of .85. A bag-of-words model is a method of extracting features from text. It converts words into a numerical representation by measuring the occurrence of words within a document. The long short term memory model had an overall area under the receiver operating curve of .72 and area under the precision-recall curve of .44.

When compared to our senior billing coder, the ultimate weighted average of the CPT by CPT code accuracies were 87% and 59% for the random forest machine learning model and

long short-term memory model, respectively. In other words, the random forest machine learning model predicted 87% of the CPT codes that were verified and sent to insurance companies by our senior billing coder. The long short term memory machine learning model predicted 59% of the CPT codes.

We also analyzed accuracies on a separate CPT by CPT code basis compared to our senior billing coder. The random forest machine learning algorithm had the highest performance as graded by accuracy on codes 22 856 (total disc arthroplasty first interspace) (98.4%), 20 931 (application of structural bone



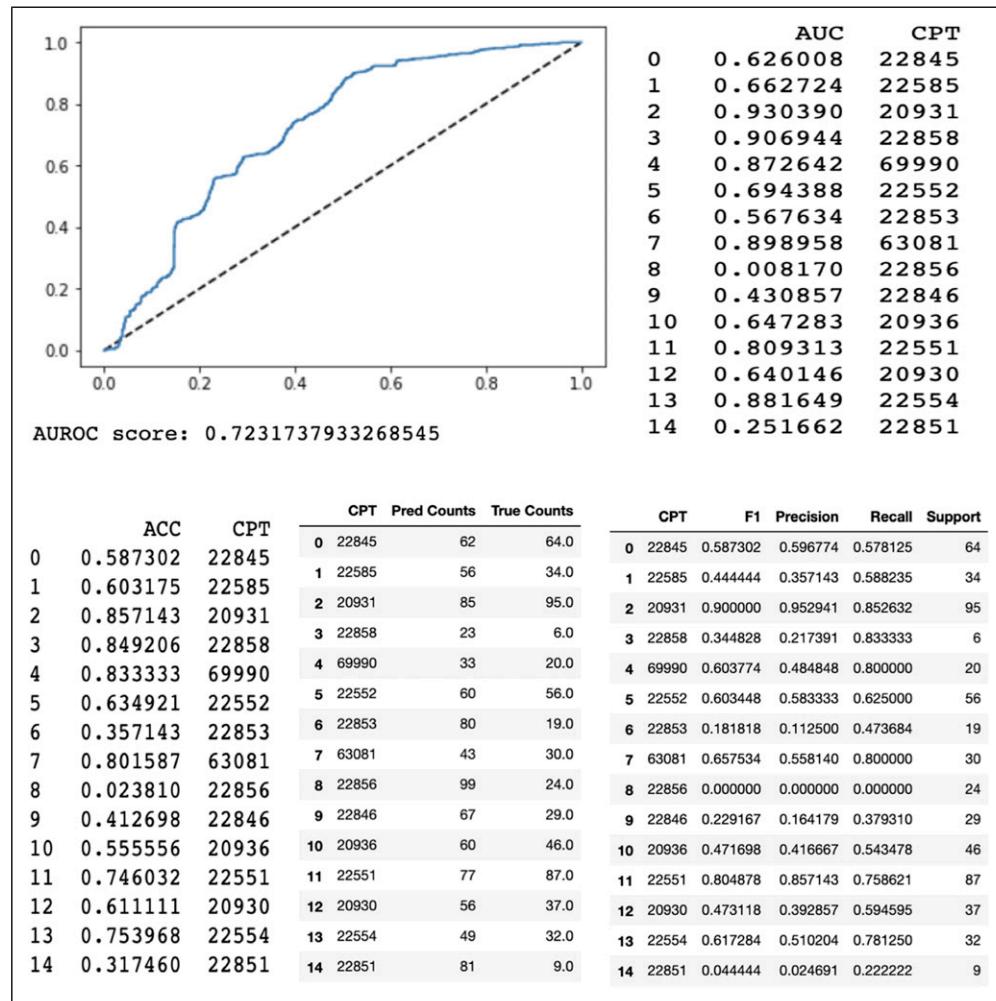
**Figure 3.** Random Forest Model a classification technique that uses layered decision trees.

graft) (96.0%), and 22 858 (total disc arthroplasty second interspace) (98.4%). The long short term memory had the highest accuracies on codes 20 931 (application of structural bone graft) (89%), 69 990 (use of operating microscope) (84%), and 22 551 (anterior interbody fusion and decompression) (75%).

Again, we analyzed the area under the receiver operating curve on a CPT by CPT code basis. The random forest model performed with the highest area under the receiver operating curve for codes 22 856 (total disc arthroplasty first interspace) (.99), 22 858 (total disc arthroplasty second interspace) (.99), and 69 990 (use of operating microscope) (.96). The long short term memory had the highest area under the receiver operating curve for codes 20 931 (application of structural bone graft)

(.92), 22 554 (arthrodesis, anterior interbody technique without decompression) (.89), 69 990 (use of operating microscope) (.88).

Finally, we analyzed the area under the precision-recall curve on a CPT by CPT code basis. The random forest had the highest area under the precision-recall curve for codes 20 931 (application of structural bone graft) (.96), 22 858 (total disc arthroplasty second interspace) (.93), and 22 856 (total disc arthroplasty first interspace) (.94). The long short-term memory had the highest AURPC for codes 22 551 (anterior interbody fusion and decompression) (.89), 22 845 (anterior instrumentation 2–3 segments) (.79), and 20 931 (application of structural bone graft) (.95). All res



**Figure 4.** Long Short-Term Memory networks take in data that comes in a series such as words or data that is linearly oriented in time. They have “memory” in that they use data from the past to make predictions about the future.

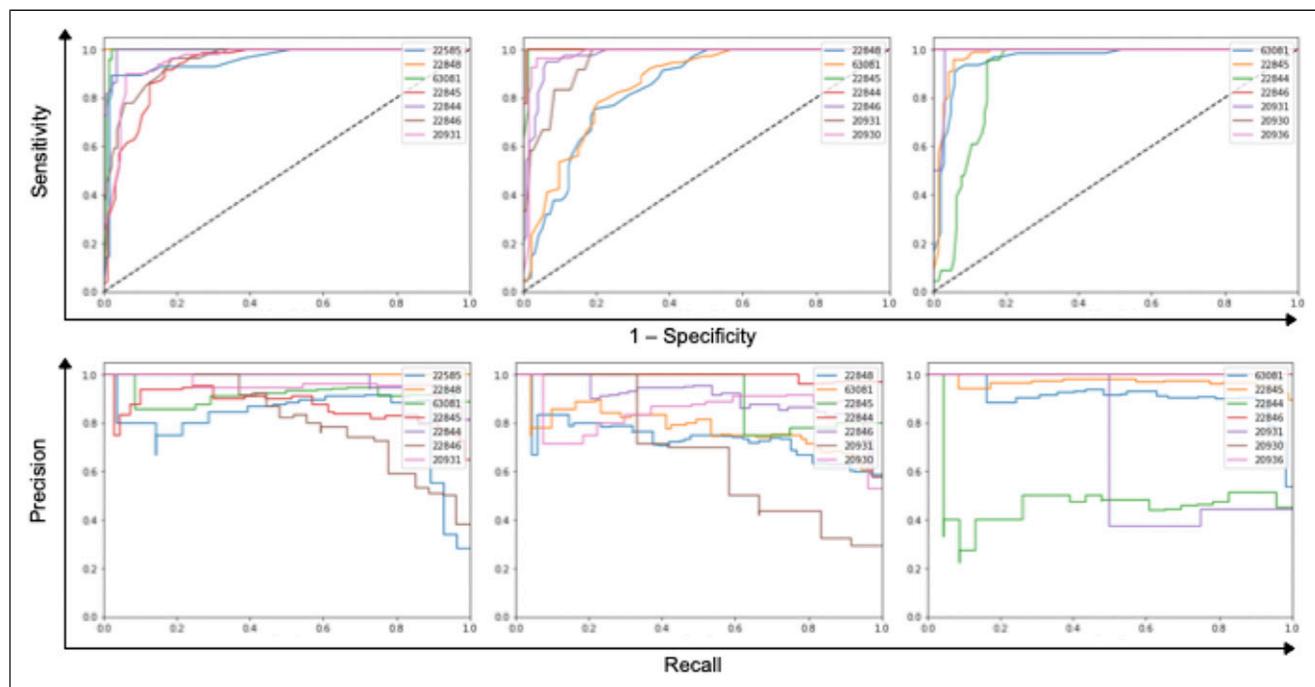
Figure 1 showed that the long short term memory model looked closely at certain words in the operative dictation to generate a prediction of cpt codes. Example key phrases included “anterior cervical,” “corpectomy,” “plating,” and “c5-6 structural...grafting.” We also noticed that for operative dictations that incorporated anterior cervical corpectomy, the algorithm looked at the percentage of total bone that was removed. This helps further validate our model as it provides transparency for the model’s analytical process. (Figures 2, 3, 4 and 5)

## Discussion

As the world population grows and ages, the demand for medical services will continue to increase; yet, declining reimbursement rates have narrowed the gap between overhead costs and fiscal gains for physician groups. A diminishing profit margin has forced providers in both academic and private sectors to re-examine all components of the medical revenue cycle.<sup>6</sup> Billing and insurance

related tasks alone may consume 14% of physician group revenue, and the financial impact of billing and insurance related tasks was on the order of 25 billion dollars in the US market.<sup>7-10</sup> One way to decrease overhead expenditures and “cost to collect” is to automate tasks historically performed by ancillary staff.

In 1995, Larkey and Croft first developed automated systems to assign coding on the basis of dictated inpatient discharge summaries using a probabilistic information retrieval system based on an inference net model.<sup>11</sup> Since then, a variety of machine learning models for billing and coding have been developed.<sup>12-19</sup> In this study, combining natural language processing with machine learning models resulted in accurate, timely generation of CPT codes based on machine interpretation of spine surgery operative notes. We compared two different machine learning approaches: A deep learning, long short-term memory model and a Random Forest model. We were able to achieve near-human accuracy, precision, and recall with a Random Forest model to correctly generate CPT codes from operative dictations.



**Figure 5.** (Top) Receiver operating characteristic curves for the random forest stratified by CPT code predicted on the test set. (Bottom) Precision-recall curves for the random forest stratified by CPT code predicted on the test set.

**Table I.** CPT Codes Used to Determine Machine Learning Algorithm Efficacy.

CPT	Description
22845	Anterior instrumentation (2–3 vertebral segments)
22846	Anterior instrumentation (4–7 vertebral segments)
20930	Application of bone graft (morcclized) or placement of osteopromotive material
20931	Application of bone graft (structural)
20936	Application of bone graft (local)
22551	Anterior interbody fusion with discectomy and decompression below C2
22552	Anterior interbody fusion with discectomy and decompression below C2, each additional interspace
22856	Total disc arthroplasty(1st interspace)
22858	Total disc arthroplasty (2nd interspace)
22853	Insertion of interbody cage (1st interspace)
69990	Use of operating microscope
22851	Application of intervertebral biomechanical device to vertebral defect or interspace
22554	Arthrodesis, anterior interbody technique, including minimal discectomy without decompression
22585	Arthrodesis, anterior interbody technique, including minimal discectomy without decompression (each additional interspace)
63081	Vertebral corpectomy (partial or complete), anterior approach with decompression of spinal cord and or nerve roots

**Table 2.** Metrics Used to Evaluate Machine Learning Algorithms Ability CPT Codes Overall.

	Long Short Term Memory Network	Random Forest Machine Learning
Accuracy to senior billing coder	59.0%	87.0%
Area under the receiver operating curve	.72	.94
Area under the precision-recall curve	.44	.85

**Table 3.** Metrics Used to Evaluate Machine Learning Algorithms Ability by CPT Codes.

	Long Short Term Memory Network	Random Forest Machine Learning
Accuracy to senior billing coder	59.0%	87.0%
Accuracy by CPT code		
22 856	—	98.4%
20 931	—	96.0%
22 858	—	98.4%
20 931	89.0%	—
69 990	84.0%	—
22 551	75.0%	—
Area under the receiver operating curve by CPT code		
22 856	—	.99
22 858	—	.99
69 990	.88	.96
20 931	.92	—
22 554	.89	—
Area under the precision-recall curve by CPT code		
20 931	—	.96
22 858	—	.93
22 856	—	.94
22 551	.89	—
22 845	.79	—
20 931	.95	—

### Evaluating Various Models for Current Procedural Terminology Code Generation

The random forest model outperformed the long short term memory model in all metrics (accuracy, area under the receiver operating curve, area under the precision-recall curve). The random forest model was less prone than the long short term memory model to guess the wrong CPT code and also less prone to guess too many or too few CPT codes per note. We used a relatively small dataset of 391 operative notes from a single institution to train these 2 models. Not surprisingly, during hyperparameter and model tuning, we noticed that the deep learning model (long short term memory) was prone to overfitting, which is a modeling error that occurs when an algorithm is fit too closely to a limited set of data points. This leads to poor generalizability of the algorithm. These findings are a direct result of the small size of this dataset. The algorithm complexity was also curtailed by the few number of unique words in this compilation of text. For these reasons, we found the “simpler” random forest model with a bag-of-words approach outperformed the long short term memory model. As we increase the number of surgeon operative notes and expand the number of CPT codes, we predict that the deep learning approach (i.e., long short-term memory) will outperform the random forest model, similar to prior studies.<sup>14</sup>

Figure 1 provided some interesting insight into the words that the long short term memory model found important to generate its prediction. The long short term memory took into account word modifiers. For instance, we see that the model

considers the diagnosis in its prediction of CPT codes. The attention decoder took into account that the patient had “cervical spondylosis with myelopathy.” The decoder also read “anterior cervical plating c4...7...with a ... plate.” A lot of the focus of the long short term memory model was in the beginning portion of the note where surgeons typically dictate their procedure summary. These findings further validate our long short term memory model as it provides transparency for the model’s analytical process. In a different way, the bag-of-words method is a simple word counter that the random forest model used to generate its predictions of CPT code. It considers only that the word appears in the body of the operative note and none of the context of the word or its meaning. This “buzz word” approach that the random forest model used was quite effective in generating accurate predictions.

### Benefits of ML Models for Automated Coding

While few studies address the utility of artificial intelligence in improving cost efficacy in medicine, various machine learning models can be used to optimize almost all steps within the revenue cycle. Combining artificial intelligence with electronic medical records can ultimately yield increases in billing for equivalent care provided achieved with less support staff.<sup>20</sup> Charge lag time, percentage of clean claims submitted (or percentage of denials), the period of time charges remain in accounts receivable, and collection ratio may likewise be improved with AI based algorithms that improve the accuracy of documentation, coding, and billing.<sup>6</sup> Most major practices

use a number of FTE's or third party billers to ensure appropriate billing occurs in a timely fashion. These are justifiable expenses, as inaccurate coding commonly results in a loss of revenue. False positives (upcoding, inappropriate or fraudulent coding) may be just as damaging as false negatives (downcoding, or failure to capture a code). In 2019, the Centers for Medicare and Medicaid Services reported that improper payment amounted to \$ 28.91 billion.<sup>21</sup>

Worst case scenarios may result in a loss of payer contracting, fraud investigations, or sanctions. Given that current fraud protection experts are increasingly using statistical based methods to screen for fraud, it is reasonable for surgeons to internally audit and do the same. Natural language processing may serve as a valuable check to ensure that coders are not upcoding or unbundling charges, constituting inadvertently fraudulent practices.<sup>22</sup> Likewise, precise diagnostic codes and electronic medical record analysis may be useful in overturning denials from payers, again maximizing the revenue stream from services provided.

The use of machine learning algorithms to generate accurate diagnostic and billing codes is not only germane to profit margins. Accurate billing and coding are critical to outcome and socioeconomic studies. Systematic reviews show considerable improvement in the quality of clinical care when using electronic medical record based predictive analytics.<sup>23</sup> Automated identification of cases based on structured data often lacks sensitivity, due to variations in coding. Natural language processing may also be a valuable resource in instances where there is an absence of CPT or ICD codes corresponding to a particular diagnosis or procedure. The use of natural language processing to characterize electronic records such as the operative note may be useful to ensure appropriate billing, but also for reporting surgical outcomes and in automating cohort creation.<sup>24</sup>

Prior studies have shown a dramatic improvement in characterization of patient diagnoses from clinical notes absent specific billing or diagnosis codes, with over 95% specificity and positive predictive values.<sup>25</sup>

### **Limitations**

This study is limited by the relatively small compilation of words and small dataset. We also limited the CPT codes in an effort to avoid the stark class imbalance issue that was present in this dataset. In future studies, we plan to increase the number of operative notes and the variability in CPT codes that are included. Finally, the use of supervised learning models may be subject to regional, institutional, or practice specific bias, which limit portability of supervised algorithms to other health care systems. All of these limitations point out that this is a preliminary result that requires much further work before it can be practically used. However, our findings also suggest that with larger numbers and further refinement, machine learning has a great potential to save time, reduce

expenses and achieve a level of accuracy that equals or surpasses humans. (Tables 1, 2 and 3).

### **Conclusions**

This preliminary, pilot study demonstrates the use of a natural language processing-based machine learning algorithm for automated CPT coding within spine surgery. The Random Forest machine learning model had a CPT by CPT accuracy of 87%. The area under the receiver operating curve and area under the precision-recall curve were also .94 and .85, respectively. Despite the limitations of our dataset, the Random Forest model achieved near-human ability to correctly generate CPT codes from operative dictations.

### **Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### **Funding**

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

### **ORCID iDs**

Jun S. Kim  <https://orcid.org/0000-0002-6114-2673>  
 Nathan J. Lee  <https://orcid.org/0000-0001-9572-5968>  
 Meghana Vulapalli  <https://orcid.org/0000-0003-1197-0400>  
 Eric Geng  <https://orcid.org/0000-0003-0736-3245>  
 Samuel K. Cho  <https://orcid.org/0000-0001-7511-2486>

### **Ethical Approval**

This study has been IRB-Approved at Columbia University Irving Medical Center under protocol number AAAS8683.

### **References**

1. Verspoor K, Martin-Sanchez F. Big data in medicine is driving big changes. *Yearb Med Inform.* 2014;15;23(1):14-20. doi:[10.15265/IY-2014-0020](https://doi.org/10.15265/IY-2014-0020).
2. Han Z, Wei B, Leung S, Chung J, Li S. Towards automatic report generation in spine radiology using weakly supervised framework. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018 185–193; September 16-20, 2018; Granada, Spain: Springer International Publishing; 2018.
3. Tan WK, Hassanpour S, Heagerty PJ, et al. Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain. *Acad Radiol.* 2018;25(11):1422-1432. doi:[10.1016/j.acra.2018.03.008](https://doi.org/10.1016/j.acra.2018.03.008).
4. 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](https://doi.org/10.1097/ALN.0000000000003150).

5. Ye J. Construction and utilization of a neural network model to predict current procedural terminology codes from pathology report texts. *J Pathol Inform.* 2019;10:13. doi:[10.4103/jpi.jpi\\_3\\_19](https://doi.org/10.4103/jpi.jpi_3_19)
6. Manley R, Satiani B. Revenue cycle management. *J Vasc Surg Cases.* 2009;50(5):1232-1238. doi:[10.1016/j.jvs.2009.07.065](https://doi.org/10.1016/j.jvs.2009.07.065).
7. Lang D. *Consultant report-natural language processing in the health care industry*, Vol. 6. Cincinnati Children's Hospital Medical Center; 2007.
8. 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-w554. doi:[10.1377/hlthaff.28.4.w544](https://doi.org/10.1377/hlthaff.28.4.w544).
9. Kahn JG, Kronick R, Kreger M, Gans DN. The cost of health insurance administration in CA: estimates for insurers, physicians, and hospitals. *Health Aff.* 2005;24(6):1629-1639. doi:[10.1377/hlthaff.24.6.1629](https://doi.org/10.1377/hlthaff.24.6.1629).
10. Casalino LP, Nicholson S, Gans DN, et al. What does it cost physician practices to interact with health insurance plans? *Health Affairs.* 2009;28(4):w533-w543. doi:[10.1377/hlthaff.28.4.w533](https://doi.org/10.1377/hlthaff.28.4.w533).
11. Larkey LS, Croft WB. *Automatic Assignment of Icd9 Codes to Discharge Summaries*; 1995. <https://www.academia.edu/download/30740467/10.1.1.49.816.pdf>.
12. PestianBrew JPC, Matykiewicz P, Hovermale DJ, Johnson N, Cohen KB, Duch W. A shared task involving multi-label classification of clinical free text. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing 97–104; June 29, 2007; Prague Czech Republic: Association for Computational Linguistics; 2007.
13. Behta M, Friedman G, Manber M, Jordan D. Evaluation of an automated inferencing engine generating ICD-9CM codes for physician and hospital billing. *AMIA Annu Symp Proc.* 2008; 873.
14. 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. Epub 2019 May 25. doi:[10.1016/jcmpb.2019.05.024](https://doi.org/10.1016/jcmpb.2019.05.024).
15. Farkas R, Szarvas G. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinf.* 2008 Apr 11; 9(suppl 3): S10. doi:[10.1186/1471-2105-9-S3-S10](https://doi.org/10.1186/1471-2105-9-S3-S10)
16. Fette G, Krug M, Kaspar M, et al. Estimating a bias in ICD encodings for billing purposes. *Stud Health Technol Inf.* 2018; 247:141-145.
17. Atutxa A, de Ilarrazza AD, Gojenola K, Oronoz M, Perez-de-Viñaspre O. Interpretable deep learning to map diagnostic texts to ICD-10 codes. *Int J Med Inf.* 2019;129:49-59. doi:[10.1016/j.ijmedinf.2019.05.015](https://doi.org/10.1016/j.ijmedinf.2019.05.015).
18. Medori J, Fairon C. Machine learning and features selection for semi-automatic ICD-9-CM encoding. Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents 84–89; June 5, 2010; Los Angeles, CA: Association for Computational Linguistics; 2010.
19. Pakhomov SVS, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *J Am Med Inf Assoc.* 2006;13(5):516-525. doi:[10.1197/jamia.M2077](https://doi.org/10.1197/jamia.M2077).
20. 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](https://doi.org/10.1016/j.jamcollsurg.2007.02.074).
21. Hammon M, Hammon WE. Benefiting from the government CERT audits. *J Oklahoma State Med Assoc.* 2005;98(8):401-402.
22. Haddad Soleimani 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](https://doi.org/10.1007/s40199-018-0227-z).
23. Black AD, Car J, Pagliari C, et al. The impact of eHealth on the quality and safety of health care: a systematic overview. *PLoS Med.* 2011;8(1):e1000387. doi:[10.1371/journal.pmed.1000387](https://doi.org/10.1371/journal.pmed.1000387).
24. Kimia AA, Savova G, Landschaft A, Harper MB. An introduction to natural language processing. *Pediatr Emerg Care.* 2015;31(7):536-541. doi:[10.1097/PEC.0000000000000484](https://doi.org/10.1097/PEC.0000000000000484).
25. Afzal N, Mallipeddi VP, Sohn S, et al. Natural language processing of clinical notes for identification of critical limb ischemia. *Int J Med Inf.* 2018;111:83-89. doi:[10.1016/j.ijmedinf.2017.12.024](https://doi.org/10.1016/j.ijmedinf.2017.12.024).