Surgical classification using natural language processing of informed consent forms in spine surgery

Michael D. Shost, BS,1,2 Seth M. Meade, BSE,1–3 Michael P. Steinmetz, MD,2,3 Thomas E. Mroz, MD,2,3 and Ghaith Habboub, MD2,3

1Case Western Reserve University, School of Medicine; 2Center for Spine Health, Neurologic Institute, Cleveland Clinic Foundation; and 3Cleveland Clinic Lerner College of Medicine, Cleveland, Ohio

OBJECTIVE In clinical spine surgery research, manually reviewing surgical forms to categorize patients by their surgical characteristics is a crucial yet time-consuming task. Natural language processing (NLP) is a machine learning tool used to adaptively parse and categorize important features from text. These systems function by training on a large, labeled data set in which feature importance is learned prior to encountering a previously unseen data set. The authors aimed to design an NLP classifier for surgical information that can review consent forms and automatically classify patients by the surgical procedure performed.

METHODS Thirteen thousand two hundred sixty-eight patients who underwent 15,227 surgeries from January 1, 2012, to December 31, 2022, at a single institution were initially considered for inclusion. From these surgeries, 12,239 consent forms were classified based on the Current Procedural Terminology (CPT) code, categorizing them into 7 of the most frequently performed spine surgeries at this institution. This labeled data set was split 80%/20% into train and test subsets, respectively. The NLP classifier was then trained and the results demonstrated its performance on the test data set using CPT codes to determine accuracy.

RESULTS This NLP surgical classifier had an overall weighted accuracy rate of 91% for sorting consents into correct surgical categories. Anterior cervical discectomy and fusion had the highest positive predictive value (PPV; 96.8%), whereas lumbar microdiscectomy had the lowest PPV in the testing data (85.0%). Sensitivity was highest for lumbar laminectomy and fusion (96.7%) and lowest for the least common operation, cervical posterior foraminotomy (58.3%). Negative predictive value and specificity were > 95% for all surgical categories.

CONCLUSIONS Utilizing NLP for text classification drastically improves the efficiency of classifying surgical procedures for research purposes. The ability to quickly classify surgical data can be significantly beneficial to institutions without a large database or substantial data review capabilities, as well as for trainees to track surgical experience, or practicing surgeons to evaluate and analyze their surgical volume. Additionally, the capability to quickly and accurately recognize the type of surgery will facilitate the extraction of new insights from the correlations between surgical interventions and patient outcomes. As the database of surgical information grows from this institution and others in spine surgery, the accuracy, usability, and applications of this model will continue to increase.

https://thejns.org/doi/abs/10.3171/2023.3.FOCUS2371

KEYWORDS machine learning; natural language processing; curriculum development; classification

NATURAL language processing (NLP) is a powerful machine learning tool that allows automated parsing and categorization of text with numerous potential applications. NLP classifier systems are trained on an initial text data set containing words or phrases with corresponding category labels, and the trained model is then validated on novel text in a testing data set in which the labeled category is not known to the model. During the training period, the model is shown the text and the correct category and determines the relative importance of each feature in the target text for each category. When presented with unseen text data, the model conducts a comparison of feature importance for each potential class and then categorizes the input text after considering the sum of all positive and negative predictors.

Multiple studies have demonstrated the utility of NLP

ABBREVIATIONS AUC = area under the curve; CPT = Current Procedural Terminology; EMR = electronic medical record; NLP = natural language processing; PPV = positive predictive value; ROC = receiver operating characteristic.


INCLUDE WHEN CITING DOI: 10.3171/2023.3.FOCUS2371.
in medicine.\textsuperscript{1–3} One study applied NLP to medication reviews by searching for positive or negative sentiments in patient testimonies and subsequently created a rating system based on patient experiences.\textsuperscript{4} Another study used an NLP system to evaluate the free-form text from surgical trainee evaluation forms and categorize them into levels of competency.\textsuperscript{5} A study from Google attempted to predict in-hospital mortality and readmission from patient records.\textsuperscript{6} Furthermore, prior studies in spine surgery have leveraged this technology for automated detection of complications (such as durotomy and venous thromboembolism) and preoperative prediction of intraoperative vascular injury during anterior spine procedures, among other applications.\textsuperscript{7–9} Given the significant flexibility of NLP systems, the vastness of healthcare data, and the ability of trained NLP systems to quickly parse text data, there are many opportunities for further implementation in healthcare.

In settings such as clinical research or surgical department analysis, increasing the efficiency of surgical data review would provide numerous benefits. A common workflow for spine surgery outcomes research may be to identify a patient population that underwent a given surgery or surgeries of interest and compare how patient or surgical factors impacted their outcome. Determining which patients should be included in the study, however, often involves manually parsing through consent forms or operative reports for pertinent details. For large retrospective studies, this may require multiple data reviewers and days of manual data review. Leveraging the power of a trained NLP model could significantly increase the efficiency of this process with minimal risk of misclassification. The time from research idea formulation to data collection would be reduced, and the barrier to conduct retrospective studies would also be reduced as a team of data reviewers would not be necessary. Furthermore, as we consider the future of predictive modeling, the ability to collect a large repository of categorized patient information, surgical information, and surgical outcomes will enable valuable insights into the factors that drive postoperative outcomes. Ultimately, this could benefit both surgeons and patients and improve preoperative surgical planning. In this study, we present our focused effort to categorize surgical cases from consent forms using an NLP classifier.

**Methods**

**Study Design**

We conducted a retrospective study of 13,268 patients at least 18 years of age who underwent 15,227 spine surgeries within the Cleveland Clinic Foundation system between January 1, 2012, and December 31, 2022. Patients were excluded if there was no text available in the consent form or no associated Current Procedural Terminology (CPT) codes for NLP dictionary creation. This study was approved as exempt human subject research and was determined to be of minimal risk by the Cleveland Clinic IRB. The IRB approved an exemption of informed consent because this study is of minimal risk and consists only of previously collected clinical data.

**Consent Forms**

Consent form text was utilized as the data source for our NLP model to interpret and choose a surgical classification. The consent forms were presented to the model in both the training and testing data sets without any preprocessing, meaning there was a wide range of variations in consent form text for each surgery. For example, some consent forms used acronyms for procedures (TLIF), whereas others wrote the full surgery title (transforaminal lumbar interbody fusion). The spinal region, likewise, was abbreviated in some consent forms and explicitly stated in others. Additionally, some consents included blood transfusions or nonspine procedures, such as upper endoscopy, to be performed during the same operative hospital encounter. This information was not removed from the consents, leading to a more robust NLP capable of distinguishing important predictors of spine surgery categorization. The rationale for this study design was twofold: it allows for more efficient categorization by eliminating the need for preprocessing of consent forms, and it broadens the applicability of the model to many healthcare systems regardless of their consent form’s structure.

**NLP System Design**

The NLP system was created using the TensorFlow open-source package for Python. Seven primary surgical operations were chosen for classification: anterior cervical discectomy and fusion, cervical laminectomy and fusion, cervical laminoplasty, cervical posterior foraminotomy, lumbar laminectomy, lumbar laminectomy and fusion, and lumbar microdiscectomy.

Once the categorizations are decided, an initial surgical classification is required to both train an NLP model and assess the correctness of its categorization. This classification was accomplished by using the most common CPT codes for these procedures at our institution. CPT codes were chosen for this application as they were found to be accurate for these procedures and allow the rapid accumulation of a large training and testing data set without manual review or data preprocessing. Figure 1 demonstrates the conditional logic used to classify surgery type based on CPT codes. The aim of this logic was to reliably reflect their corresponding surgery type. Uncommon, complex, or combination surgeries would not meet any of the criteria outlined in the conditional logic and would instead be assigned to a categorization of “other.”

Following the initial surgery class labeling by CPT code, the data set of consent form text and surgery type was split 80%/20% into training and testing subsets, respectively. Ten percent of the resulting training data was held for validation. After training and validation, the model was presented with previously unseen consent form text from the testing data set. The CPT code was hidden from the model during testing; after the NLP model’s determination of surgical category, the accuracy of this prediction was determined based on the surgery indicated by the CPT code.

To make a prediction, the model parsed the consent form into subcomponents or features. For each surgical
category, the features were evaluated for their relative importance in predicting that surgery as established from the learning phase. Figure 2 demonstrates a Shapley text plot of this process. An example consent form, “L4/5 transforaminal lumbar interbody fusion with posterolateral fixation and fusion,” was being evaluated against each of the 7 possible surgical categories. Each surgical category weighs feature importance differently according to what was learned from the training data set. This is demonstrated in Fig. 2, in which words in red represent a strong positive predictor for the selected category, blue words represent a strong negative predictor, and white or lighter shades represent weak feature importance that contributes little to the probability of categorization. The probability of correct categorization was compared for each surgery type and the category with the highest overall probability was chosen. In the example in Fig. 2, the model concluded that there was a high probability that this text corresponded to the “Lumbar laminectomy and fusion” group and categorized it accordingly.

Hyperparameter testing was performed to determine the optimal model configuration. The final model was constructed using the following layers: a pretrained embedding layer (nnlm-en-dim50 trained on English Google news) and one hidden layer with dropout layers for regularization.

Data Analysis

The categorization performance for each surgery type was reported with tables and figures to summarize the overall performance and a per-category breakdown. The packages used to evaluate the model included Keras with

FIG. 1. This conditional logic CPT classifier was used to create the initial dictionary of surgical categories. The input is the list of all CPT codes and the output is the surgery type.
Shost et al.

Neurosurg Focus  Volume 54 • June 2023

TensorFlow, scikit-learn, and Shapley packages within Python (Python Software Foundation). Additional statistical analysis was conducted using Microsoft Excel and RStudio.

Results

CPT Classification

After completion of the CPT classification stage, we were able to sort 12,239 surgeries into the 7 most common spine surgeries using the conditional logic outlined in Fig. 1. The results of this classification were as follows: 4271 lumbar laminectomy and fusion (34.9%), 2499 lumbar laminectomy (20.4%), 1941 anterior cervical disectomy and fusion (15.9%), 1643 lumbar microdisectomy (13.4%), 1262 cervical laminectomy and fusion (10.3%), 445 cervical laminoplasty (3.6%), and 178 cervical posterior foraminotomy (1.5%). Notably, 2988 surgeries (20%) were not able to be sorted by the CPT classifier because they had complex procedure code combinations, were less common procedures, or had incomplete billing information. As a result, these consent forms could not be used for training and were not factored into the evaluation of the model’s accuracy as there was no surgical classification available based on CPT code.

NLP Classifier Performance

The overall weighted accuracy of the NLP consent form model was 91%. The surgical category with the highest positive predictive value (PPV) was anterior cervical disectomy and fusion with 97%, whereas lumbar microdisectomy had the lowest PPV at 85%. Sensitivity was highest for cervical laminoplasty (97%) and lumbar laminectomy and fusion (97%). A summary of the model performance for each surgical category is included in Table 1. The confusion matrix (Fig. 3) further illustrates the performance for each surgical category and provides details on false labeling. Two of the most difficult situations for the model to interpret were 1) between lumbar microdisectomy and lumbar laminectomy, and 2) between cervical posterior foraminotomy, cervical laminectomy and fusion, and lumbar laminectomy. The micro-averaged one-versus-rest receiver operating characteristic (ROC) curve shows the performance of each surgical category in the model compared with an untrained, chance-based model (Fig. 4). By this metric, all categories achieved an area under the curve (AUC) of 0.98 or greater. The micro- and macro-averaged ROC curves consider all surgical categories and have AUCs of 1.00 and 0.99, respectively. These values were all greater than the baseline AUC of 0.5 for the chance model.

Hyperparameter optimization was achieved with hand

---

**TABLE 1. Overall NLP model performance by surgical category**

<table>
<thead>
<tr>
<th>Surgical Category</th>
<th>Total (n)</th>
<th>PPV</th>
<th>NPV</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anterior cervical disectomy and fusion</td>
<td>388</td>
<td>96.8%</td>
<td>99.0%</td>
<td>94.8%</td>
<td>99.4%</td>
</tr>
<tr>
<td>Cervical laminectomy and fusion</td>
<td>252</td>
<td>86.7%</td>
<td>98.9%</td>
<td>90.5%</td>
<td>98.4%</td>
</tr>
<tr>
<td>Cervical laminoplasty</td>
<td>89</td>
<td>93.5%</td>
<td>99.9%</td>
<td>96.6%</td>
<td>99.7%</td>
</tr>
<tr>
<td>Cervical posterior foraminotomy</td>
<td>36</td>
<td>91.3%</td>
<td>99.4%</td>
<td>58.3%</td>
<td>99.9%</td>
</tr>
<tr>
<td>Lumbar laminectomy</td>
<td>500</td>
<td>83.9%</td>
<td>97.0%</td>
<td>88.6%</td>
<td>95.6%</td>
</tr>
<tr>
<td>Lumbar laminectomy and fusion</td>
<td>854</td>
<td>96.5%</td>
<td>98.2%</td>
<td>96.7%</td>
<td>98.1%</td>
</tr>
<tr>
<td>Lumbar microdisectomy</td>
<td>329</td>
<td>85.0%</td>
<td>96.8%</td>
<td>79.0%</td>
<td>97.8%</td>
</tr>
</tbody>
</table>

NPV = negative predictive value.

The average weighted accuracy of the NLP model was 91%.
tuning of the model. A summary of the parameter testing results is included in Table 2. The model’s accuracy and loss were plotted for both training and validation data in Fig. 5. The optimal number of epochs (an epoch is when the data set is passed through the given algorithm one time) and training length were determined using the “early stopping” callback to monitor validation loss. After analyzing the convergence of validation and training loss and balancing the minimization of loss and maximization of accuracy, model 3 was selected as the final model for this project.

Discussion

NLP has increased in popularity in recent years as its power and flexibility continue to be demonstrated in cases of varied use.3,15,16 Manual chart review is an especially time-intensive task in the research process for most clinical studies, and having multiple data reviewers with varying levels of training often leads to heterogeneous data collection. A standardized machine learning approach would streamline this process and assist institutions without large databases or significant teams of data reviewers to quickly produce reliable research information from clinical data. This study describes a strategy to simplify data acquisition through NLP modeling. By training an NLP system with the data available in the electronic medical record (EMR), we were able to demonstrate 91% overall accuracy in identifying an operation without any user input or specially created dictionary.

The CPT classifier demonstrated in Fig. 1 was used to build the surgical dictionary that trained our NLP model and demonstrates an easy and approachable way to quickly categorize data into common surgical categories. This approach relies on CPT codes to make surgical assignments and is a great tool for surgical trainees and institutions to track surgical volume of most common spine surgeries. CPT codes are sufficiently accurate for common operations that do not include modifiers or multiple surgical techniques. For this reason, it is a trustworthy resource for training an NLP model on the language for each surgical category. However, approximately 20% of the surgeries at our institution have complex or incomplete CPT coding that was not able to be interpreted by a CPT classifier alone. Our NLP model is one possible solution for this problem because it has already been trained on categorized data and can be deployed to interpret this “other” category without any additional input.

The ability of the NLP model to correctly interpret and categorize surgical information will continue to increase as more data and categorizations are added to the training dictionary. For example, while all lumbar fusion surgeries are currently categorized together under lumbar laminectomy and fusion, additional data would allow us to create separate categorizations for operations such as anterior lumbar interbody fusions, lateral lumbar interbody fusions, and transforaminal lumbar interbody fusions. With this level of detail, surgeons would be able to stratify their previous patients by surgical approach to compare how similar patients recovered after surgery, and this informa-
tion may inform how they counsel a patient on their likely postoperative outcome or surgical options during a preoperative clinic visit. As additional data sources such as operative reports are added, it will be possible to expand beyond surgical classifications to surgical information such as the number of operative levels and what operation was performed at each level in complex multilevel multitechnique procedures. This level of detail will never be possible with the CPT system as presently constructed and demonstrates an area of growth and opportunity for NLP and machine learning technology.

**Strengths of the Study**

This work represents a dramatic improvement in efficiency for a routine and simple task performed in retrospective spine surgical research, the assessment of fellows’ caseloads, and other important applications. The performance of the most common surgical procedure in

![ROC Curve](image)

**FIG. 4.** The one-versus-rest ROC curve demonstrates the performance of our NLP model for each surgery type compared with a model based on chance. Micro- and macro-averaged ROC curves for all surgical categories are included. Performance is estimated from the AUC. Ant. = anterior; Cerv. = cervical; Disc. = discectomy; Lami., Lamin. = laminectomy; Lum. = lumbar; Post. = posterior.

<table>
<thead>
<tr>
<th>Model No.</th>
<th>Fully Connected Layer (Neurons)</th>
<th>Best Epoch</th>
<th>Training Loss</th>
<th>Training Accuracy</th>
<th>Validation Loss</th>
<th>Validation Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64</td>
<td>1033</td>
<td>0.206</td>
<td>0.934</td>
<td>0.265</td>
<td>0.923</td>
</tr>
<tr>
<td>2</td>
<td>128</td>
<td>784</td>
<td>0.206</td>
<td>0.934</td>
<td>0.225</td>
<td>0.920</td>
</tr>
<tr>
<td>3</td>
<td>256</td>
<td>800</td>
<td>0.192</td>
<td>0.938</td>
<td>0.223</td>
<td>0.923</td>
</tr>
<tr>
<td>4</td>
<td>256 128</td>
<td>649</td>
<td>0.241</td>
<td>0.923</td>
<td>0.231</td>
<td>0.919</td>
</tr>
<tr>
<td>5</td>
<td>512 512</td>
<td>575</td>
<td>0.214</td>
<td>0.930</td>
<td>0.224</td>
<td>0.921</td>
</tr>
<tr>
<td>6</td>
<td>512 512 256</td>
<td>578</td>
<td>0.243</td>
<td>0.921</td>
<td>0.222</td>
<td>0.922</td>
</tr>
<tr>
<td>7</td>
<td>512 512 512</td>
<td>576</td>
<td>0.230</td>
<td>0.924</td>
<td>0.225</td>
<td>0.917</td>
</tr>
<tr>
<td>8</td>
<td>256 256 256 256 713</td>
<td>684</td>
<td>0.241</td>
<td>0.920</td>
<td>0.230</td>
<td>0.915</td>
</tr>
<tr>
<td>9</td>
<td>512 512 256 256 512 512 601</td>
<td></td>
<td>0.234</td>
<td>0.921</td>
<td>0.221</td>
<td>0.918</td>
</tr>
</tbody>
</table>

**TABLE 2. Hyperparameter tuning results for the NLP model**
our database, lumbar laminectomy and fusion, was the most accurate result, which suggests that the accuracy of an NLP model will continue to improve as the dictionary of surgical data continues to expand. Additionally, much of the error in this model came from slight differences in wording, such as identifying a laminectomy and fusion, not from choosing the wrong spinal level or classifying a microdiscectomy as a laminectomy. These errors illustrate clear areas to target for further refinement that will result in even greater accuracy rates. Notably, NLP models are robust for different text formats, meaning that this model would behave similarly for consent forms across institutions. Any text that can be extracted from an EMR and presented to the NLP algorithm could be interpreted as is or used to continue to train and enhance the accuracy of the model.

Limitations of the Study

Misclassification affected approximately 9% of the data. Part of this misclassification was due to the occasional inaccuracy of CPT codes. While this project endeavored to only train on the most clearly defined CPT codes, there were still instances of inappropriate or inaccurate coding that introduced error to the model. For applications in which higher accuracy is needed, misclassification could be improved by manually reviewing all text and assigning surgical categories in the training and testing data set. However, for most research applications, the accuracy achieved by this model will result in the overwhelming majority of surgeries being appropriately categorized, with manual review reserved for the small group of misclassified or more complex cases to ensure high accuracy.

While the model demonstrated high accuracy overall, there were patterns in the categorizations that may be opportunities for future improvement. Specifically, the model demonstrated difficulty in discerning between lumbar microdiscectomies and lumbar laminectomies, as well as between cervical posterior foraminotomy and cervical laminectomy. The consent forms for these groups were found to contain many overlapping terms related to the decompression and offer a clear example in which additional information on approach and the extent of an operation from the operative report could reduce misclassification.

Another inherent limitation in this study design is that consent forms may not fully correspond with the surgery performed in the event that the surgical plan changes intraoperatively. Incorporation of more definitive data, such as the operative report, would reduce this source of potential misclassification and allow for the extraction of additional surgical information, but this project demonstrates that consent form data alone are adequate for accurate and efficient surgery identification. Not all systems may initially benefit from our method, such as those with written consent forms or without EMRs. However, with the continued expansion of EMRs and software tools such as optical character recognition that process handwritten documents into digital formats, we believe that the utility of machine learning and NLP models will be available to many hospital systems.

Future Directions

The advancement of NLP systems relies on increasing the amount of surgical information per patient as well as increasing the amount of patient data in the training data set. This can be accomplished by contributing additional data sources, most notably the operative report, as well as collaborating with other health systems to build a large, diverse repository of deidentified patient data. With the near-universal use of electronic health records, healthcare systems across the country and the world possess vast amounts of patient data. While this amount of data quickly becomes overwhelming for any human reviewer to examine and parse through, machine learning offers the opportunity to extract novel insights that can improve surgical care and advance the scientific process.

The application of this system in research would allow investigators to rapidly proceed from idea formulation to data collection on patients of interest. In the clinical setting, this model would allow surgeons to quickly overview their operative practice and review patient outcomes by surgical type. Training with additional data sources, such as operative reports, would also allow for the collection of complication data, operative approach, and surgical level,
to name a few possibilities. In the era of probabilistic models, this amount of granular data is valuable for the creation of predictive systems that can analyze prior surgeries along with preoperative factors and postoperative results to produce insights that can benefit surgical planning and patient expectations. By considering all of a given patient’s preoperative factors and the outcomes of similar patients who underwent different procedures, an advanced predictive model would be able to provide recommendations based on the data gathered from an NLP model.

Conclusions
Processing surgical information from healthcare data sources is an ideal task for a machine learning application, and NLP will dramatically increase the efficiency of research and surgical data review. This project demonstrates the rapid ability of an NLP algorithm to parse, learn, and output surgical information from consent forms. Continued implementation of this work will enable faster and more accurate evaluation of surgical spine procedures.

References

Disclosures
Dr. Steinmetz reported receiving royalties from Zimmer/Biomet, Elsevier, and Globus, and an honorarium from Cerapedics outside the submitted work. Dr. Mroz reported receiving royalties from Stryker outside the submitted work.

Author Contributions
Conception and design: Habboub, Shost. Acquisition of data: Habboub, Shost. Analysis and interpretation of data: all authors. Drafting the article: Habboub, Meade, Mroz. Critically revising the article: Habboub, Meade, Mroz. Reviewed submitted version of manuscript: all authors. Statistical analysis: Habboub, Shost. Administrative/technical/material support: Meade.

Correspondence
Ghaith Habboub: Center for Spine Health, Cleveland Clinic, Cleveland, OH. hhabboug@ccf.org.