ANESTHESIOLOGY

Classification of Current Procedural Terminology Codes from Electronic Health Record Data Using Machine Learning

Michael L. Burns, M.D., Ph.D., Michael R. Mathis, M.D., John Vandervest, M.S., Xinyu Tan, Ph.D., Bo Lu, B.S., Douglas A. Colquhoun, M.B. Ch.B., M.Sc., M.P.H., Nirav Shah, M.D., Sachin Kheterpal, M.D., M.B.A., Leif Saager, Dr. med., M.M.M.

ANESTHESIOLOGY 2020; 132:738–49

EDITOR'S PERSPECTIVE

What We Already Know about This Topic

- The ability to process anesthesiology procedure code data in an accurate manner is important for clinical and research considerations. Advanced data science techniques present opportunities to improve coding and develop classification tools.

What This Article Tells Us That Is New

- The application of machine learning and natural language processing techniques facilitate a more rapid creation of accurate real-time models for Current Procedural Terminology code classification. The potential benefits of this approach include performance optimization and cost reduction for quality improvement, research, and reimbursement tasks that rely on anesthesiology procedure codes.

ABSTRACT

Background: Accurate anesthesiology procedure code data are essential to quality improvement, research, and reimbursement tasks within anesthesiology practices. Advanced data science techniques, including machine learning and natural language processing, offer opportunities to develop classification tools for Current Procedural Terminology codes across anesthesia procedures.

Methods: Models were created using a Train/Test dataset including 1,164,343 procedures from 16 academic and private hospitals. Five supervised machine learning models were created to classify anesthesiology Current Procedural Terminology codes, with accuracy defined as first choice classification matching the institutional-assigned code existing in the perioperative database. The two best performing models were further refined and tested on a Holdout dataset from a single institution distinct from Train/Test. A tunable confidence parameter was created to identify cases for which models were highly accurate, with the goal of at least 95% accuracy, above the reported 2018 Centers for Medicare and Medicaid Services (Baltimore, Maryland) fee-for-service accuracy. Actual submitted claim data from billing specialists were used as a reference standard.

Results: Support vector machine and neural network label-embedding attentive models were the best performing models, respectively, demonstrating overall accuracies of 87.9% and 84.2% (single best code), and 96.8% and 94.0% (within top three). Classification accuracy was 96.4% in 47.0% of cases using support vector machine and 94.4% in 62.2% of cases using label-embedding attentive model within the Train/Test dataset. In the Holdout dataset, respective classification accuracies were 93.1% in 58.0% of cases and 95.0% among 62.0%. The most important feature in model training was procedure text.

Conclusions: Through application of machine learning and natural language processing techniques, highly accurate real-time models were created for anesthesiology Current Procedural Terminology code classification. The increased processing speed and *a priori* targeted accuracy of this classification approach may provide performance optimization and cost reduction for quality improvement, research, and reimbursement tasks reliant on anesthesiology procedure codes.

(ANESTHESIOLOGY 2020; 132:738–49)

Anesthesiology professional fee billing is a complex process requiring accurate documentation by clinical providers and timely coordination among administrative personnel. Billing staff are responsible for selecting Current Procedural Terminology (CPT) codes to describe anesthesia care provided during each procedure and enable reimbursement using relative base unit values.[1–3] Anesthesiology base CPT codes are determined by surgical procedures performed. The process of assigning CPT codes is complex and labor-intensive, requiring various resources including specialized coding personnel for health record data extraction, transcription, translation, assignment, validation, and auditing.[4] The process can be costly: Professional billing costs are estimated to represent 13.4% of professional revenue for ambulatory surgical procedures and 3.1% for inpatient surgical procedures, equating to an estimated $170 to $215 per case for billing and insurance-related activities.[5] Error rates in medical coding can be high: Even with specialized

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are available in both the HTML and PDF versions of this article. Links to the digital files are provided in the HTML text of this article on the Journal's Web site (www.anesthesiology.org). This study has been presented in part at the 2017 Annual Meeting of the American Society of Anesthesiologists, October 21 to 25, 2017, in Boston, Massachusetts.

Submitted for publication June 8, 2019. Accepted for publication December 17, 2019. Published online first on February 4, 2020. From the Department of Anesthesiology, University of Michigan Medical School, Ann Arbor, Michigan (M.L.B., M.R.M., J.V., X.T., B.L., D.A.C., N.S., S.K., L.S.); Department of Anaesthesiology, University Medical Center Goettingen, Goettingen, Germany (L.S.).

teams, error rates as high as 38% for standard CPT coding in anesthesia have been described,[6] well above the 2018 overall fee-for-service error rate reported from Comprehensive Error Rate Testing by the Centers for Medicare and Medicaid Services (8.1%).[7] Modest gains in process efficiency can have large effects on revenue: A decrease by 10 days in accounts receivable resulted in a 3.0% revenue gain for a single academic anesthesiology practice.[8] Although crosswalk from surgical to anesthesia CPTs exists, surgical CPT data are frequently unavailable real-time as a result of business, political, or technical obstacles; when available, surgical CPT data have similar lag times to anesthesia CPT generation. Efficient billing processes are key to maintaining financial viability within departments. Additionally, billing data are vitally important in quality improvement and research projects to allow reproducible case inclusion, exclusion, and risk adjustment.[9]

As electronic health record adoption has increased, healthcare data have become more available. Data science techniques have also advanced, including methods for creating classification models using machine learning, and processing and analyzing human language using natural language processing. Machine learning and natural language processing have been applied to a variety of clinical applications including disease prediction,[10] gene expression profiling,[11] and medical imaging.[12] Such techniques are beginning to be applied within clinical anesthesiology[13] and intensive care,[14] including applications predicting bispectral index,[15] hypotension,[16] and postoperative mortality.[17] Although applications exist in medical coding, including assignment of International Classification of Diseases diagnostic codes,[18,19] there remains a paucity of work to apply these techniques to anesthesia billing. Anesthesia billing is a classification problem in which text and other variables are translated into a single numerical code from a limited set of choices. Natural language processing and machine learning tools excel at these tasks.

Using data science techniques applied to perioperative electronic health record data across multiple centers, anesthesia CPT code classification models were developed *via* multiple machine learning methods and evaluated. We hypothesized that machine learning and natural language processing could be used to develop an automated system capable of classifying anesthesia CPT codes with accuracy exceeding current benchmarks. This classification modeling could prove beneficial in efforts to optimize performance and reduce costs for research, quality improvement, and reimbursement tasks reliant on such codes.

Materials and Methods

Study Design

Institutional Review Board approval for this multicenter study was obtained for this retrospective observational study (HUM00152875, Ann Arbor, Michigan) and followed multidisciplinary guidelines for reporting machine learning–based classification models in biomedical research.[20] The study design was presented, approved, and registered within the multicenter research committee on August 14, 2017, before accessing the data.[21] This design included study outcomes, data collection, and statistical analyses.

Case Selection

This study included all patients, adults and pediatrics, undergoing elective or emergent procedures with an institution-assigned valid anesthesia CPT code and an operative date between January 1, 2014 and December 31, 2016 from 16 contributing centers in the Multicenter Perioperative Outcomes Group database. This data set includes both academic hospitals and community-based practices across the United States. Methods for data collection, validation, and multicenter integration within the Multicenter Perioperative Outcomes Group are previously described,[22,23] and data from this group have been used in multiple published studies.[24–26] All sites submitting valid data were eligible for inclusion; cases with missing procedure text were excluded. No additional exclusion criteria were applied. This data set is called Train/Test.

A second and distinct data set was created using cases on patients undergoing elective or urgent procedures with a valid institution-assigned CPT code between October 1, 2015 and November 1, 2016 from a single Multicenter Perioperative Outcomes Group institution not included in the Train/Test data set. This Holdout data set was used for external validation of the models created in this study. Figure 1 shows a flow diagram of the data sets used and the experimental design of this study.

Model Features

Features are model inputs, whereas labels are outputs. To maximize the number of cases included in the study and

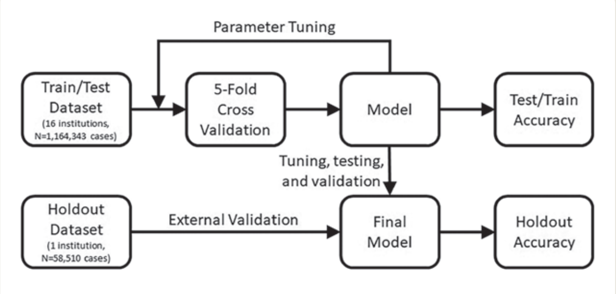

Fig. 1. Machine learning study design flowchart. Shown is a flow diagram of the experimental design of this study. The Train/Test data set is used to create each model, whereas the Holdout data set is used as an external validation. Each model is trained using fivefold cross validation. Parameter tuning occurred with each of the 20 iterations of model training. The single institution from the Holdout dataset was not included in the 16 institutions included in the Train/Test dataset.

allow for broad and easy application of the models, the features used in each model were limited to perioperative electronic health record data commonly found in anesthesia records: age, sex, American Society of Anesthesiologists (ASA; Schaumburg, Illinois) Physical Status, emergent status, procedure text, procedure duration, and the derived procedure text length (number of words in procedure text). Institution-assigned anesthesia CPT codes were used as labels for each case, and each case represents an instance for machine learning modeling. Continuous features underwent scaling through normalization to achieve properties of a standard normal distribution with a mean of zero and a SD of one.

Primary Outcome

Submitted claim data from billing specialists was used as a reference standard to train and test models. The primary outcome of this study is classification accuracy of institution-assigned anesthesia CPT code. Accuracy is defined as (Number of Correct Anesthesia CPT Classifications) / (Total Number of Anesthesia CPT Classifications). To measure the quality of the reference standard, 500 random cases were randomly selected from the Train/Test data set and adjudicated by manual review of operative notes and anesthesia records, performed by an anesthesiologist domain expert (M.L.B.), and from which a sample of 50 cases were reviewed by the University of Michigan departmental billing manager.

Data Preparation and Natural Language Processing

Procedural text is the short text assigned to each case, describing the procedure(s) carried out. Natural language processing techniques were used to process text data into forms usable for machine learning models. As procedure text is typically hand entered, it is subject to misspellings and frequently contains medical abbreviations and acronyms. Top misspelled words by frequency were physician hand audited for validity and placed into a dictionary which was used for text processing. To aid in processing and decrease vocabulary size, procedure text was standardized through removal of numbers, punctuations, and common English stop words (*e.g.*, "a," "an," "the," *etc.*). Common medical abbreviations and acronyms were expanded using domain knowledge from an anesthesiologist (M.L.B.), and a unique spelling correction library was created using approximate string distance and co-occurrence algorithms. The spelling correction library was then manually adjudicated by an anesthesiologist (M.L.B.). After text processing, term matrices were created with single and multi-word phrases using n-grams.[27,28] Steps to transform text into numerical values used in machine learning models included term frequency-inverse document frequency and word2vec.[29–31] Details of natural language processing and text transformation can be found in Supplemental Digital Content 1 (http://links.lww.com/ALN/C202).

Supervised Machine Learning Methods

In supervised machine learning methods all data used in training have labels, meaning that each case used in training has inputs and outputs. Five unique supervised machine learning classification models were compared: random forest,[32] long short-term memory,[33] extreme gradient boosting,[34] support vector machine,[35] and label-embedding attentive model.[18] Each model was chosen for potential advantageous properties, including ease of implementation/interpretation (random forest and support vector machine), reduction of bias *via* weighting of low sample observations (extreme gradient boosting), and ease of handling text and language inputs (long short-term memory and the label-embedding attentive model). Random forest was implemented using R, whereas long short-term memory, extreme gradient boosting, support vector machine, and the label-embedding attentive model were implemented in Python using TensorFlow and trained on an Amazon Web Services graphics processing units. After initial hyper-parameter tuning, all models were trained and tested 20 times using fivefold cross validation: 80% of data for training and the remaining 20% for testing. Further details of the machine learning packages used and their hyper-parameter tuning can be found in Supplemental Digital Content 2 (http://links.lww.com/ALN/C203).

The deep learning methods in this study were the label-embedding attentive model[18] and long short-term memory. Procedure texts for these models were encoded into vectors using word2vec embedding[31] as input. The label-embedding attentive model encoded the descriptions for each anesthesia CPT from the CPT Professional Edition medical code set maintained by the American Medical Association (Chicago, Illinois).[2] Most deep learning models for text classification only embed input (feature) text.[36] A compatibility matrix was computed between embedded words and labels *via* cosine similarity. From this matrix, an attention score was calculated for each word and the entire procedural text sequence was then derived as the average of embedded words, weighted by the attention scores. This score was used for CPT classification.

Feature Importance

Within the support vector machine model, linear coefficients were used to investigate which features were most important for machine learning decisions. The higher the weight of the input feature, the more important the feature is to CPT classification. Within procedure text, weights were used to compare feature importance of individual words as well as the overall importance of the entire procedure text as the sum of the weights of individual words.

Confidence Parameter

To identify specific cases for which the machine learning models demonstrated a prespecified level of accuracy, an adjustable confidence parameter was created as a model output for each case using methods similar to previous statistical studies such as density ratio estimations.[37,38] Importantly, after machine learning model training, the confidence parameter is calculable for each case, before accessing the institution-assigned CPT code to ascertain classification accuracy. Support vector machine and the label-embedding attentive model were the selected machine learning methods to calculate the confidence parameter, given their relative amenability to handling procedure text, compared with other machine learning methods studied.

The confidence parameter was created by comparing the top two primary anesthesia CPT codes for each case using CPT probabilities in the support vector machine model and CPT scores in the label-embedding attentive model. The confidence parameter is calculated for each case as follows. For the support vector machine model, confidence parameter is calculated as:

$$\mathit{confidence\ parameter} = \frac{P_{CPT1}}{P_{CPT2}} - 1$$

Where P_{CPT1} and P_{CPT2} are the highest and second highest probabilities of all CPTs for that case. For the label-embedding attentive model, confidence parameter was calculated as:

$$\mathit{confidence\ parameter} = \mathit{score}_{CPT1} - \mathit{score}_{CPT2}$$

Where score_{CPT1} and score_{CPT2} are the highest and second highest scores of all CPTs for that case. Cases were stratified into three confidence parameter ranges to differentiate cases with high *versus* low classification confidence: High (confidence parameter at or above 1.6), Medium (less than 1.6 and at or above 1.2), and Low (less than 1.2; figs. 2 and 3). The High category was targeted to return at least 95% accuracy (*i.e.*, more than 5.0% misclassification rate), as was the goal for this study. The Medium and Low categories were targeted to achieve balanced classes. Although these strata were chosen for reporting purposes, it is worth noting that any confidence parameter threshold can be selected based on the desired accuracy.

Testing Generalizability, Calibration, and Model Processing Speed

To determine the generalized ability of the models to classify anesthesia CPT codes, select models were tested on the Holdout data set (data from a distinct institution unseen by the Train/Test data set). For ease of assessing model calibration (as described further in the statistical analysis), CPT codes were transformed to a continuous variable *via* CPT-specific anesthesia base unit values, as are currently used for anesthesiology reimbursement.[39] The higher the assigned base unit value, the higher the reimbursement. Of note, each CPT code has a single base unit value, but multiple CPT codes may have the same base unit value. Finally, to assess the feasibility of an automated CPT classification model to be deployed in real time, potentially embedded into the perioperative electronic health record, the Holdout data set was processed 10 times on the support vector machine model, measuring processing time (in seconds).

Statistical Analysis

Exploratory data analysis techniques such as histograms, QQ-Plots, box-plots, scatterplots, and basic descriptive (means, medians, interquartile range) were used to assess the distribution of measures, to explore the most informative transformations, extreme values of the covariates, confounders, and relevant predictors considered in the analysis. These analyses were performed within the Train/Test and Holdout data sets separately. Standardized differences were used to compare summary statistics across these two data sets. To reduce the dimensionality of the classification model and to facilitate comparisons across clusters of CPT codes, a clinical approach was adopted in which CPT codes were grouped by anatomical region of the surgical procedure.[40] These are referred to as CPT categories in the text. Model performance was analyzed by assessing accuracy—defined as a first choice CPT classification matching the institutional-assigned CPT code existing in the Multicenter Perioperative Outcomes Group database. Accuracy within the top three is defined as one of the top three CPT classifications from the model matching the institution-assigned CPT code; narrowing a billing specialist's classification task from 285 possible CPT codes to only three may yield efficiency gains. In response to peer review, other metrics of classification such as the net reclassification index and calibration were also used to assess the quality of the classification models. Both of these metrics were appropriately modified from the classical binary classification to the multiclass classification case. To assess, we considered the following statistics:

$$\mathit{net\ reclassification\ index} = \frac{\left(\hat{p}_{up} - \hat{p}_{down}\right)}{\sqrt{\dfrac{\hat{p}_{up} + \hat{p}_{down}}{n}}}$$

Where and $\hat{p}_{up}$ and $\hat{p}_{down}$ are the average of the probability estimates for CPT codes for which the base unit value of the model-classified CPT codes went up or down with respect to the original CPT code, and n is the total number of CPT codes classified. The net reclassification index is then interpreted as the net change in base unit value of CPT codes reclassified by both models. Calibration plots were constructed using z scores for base unit values from reference standard CPT codes as well as base unit values from model-classified CPT code for both models.

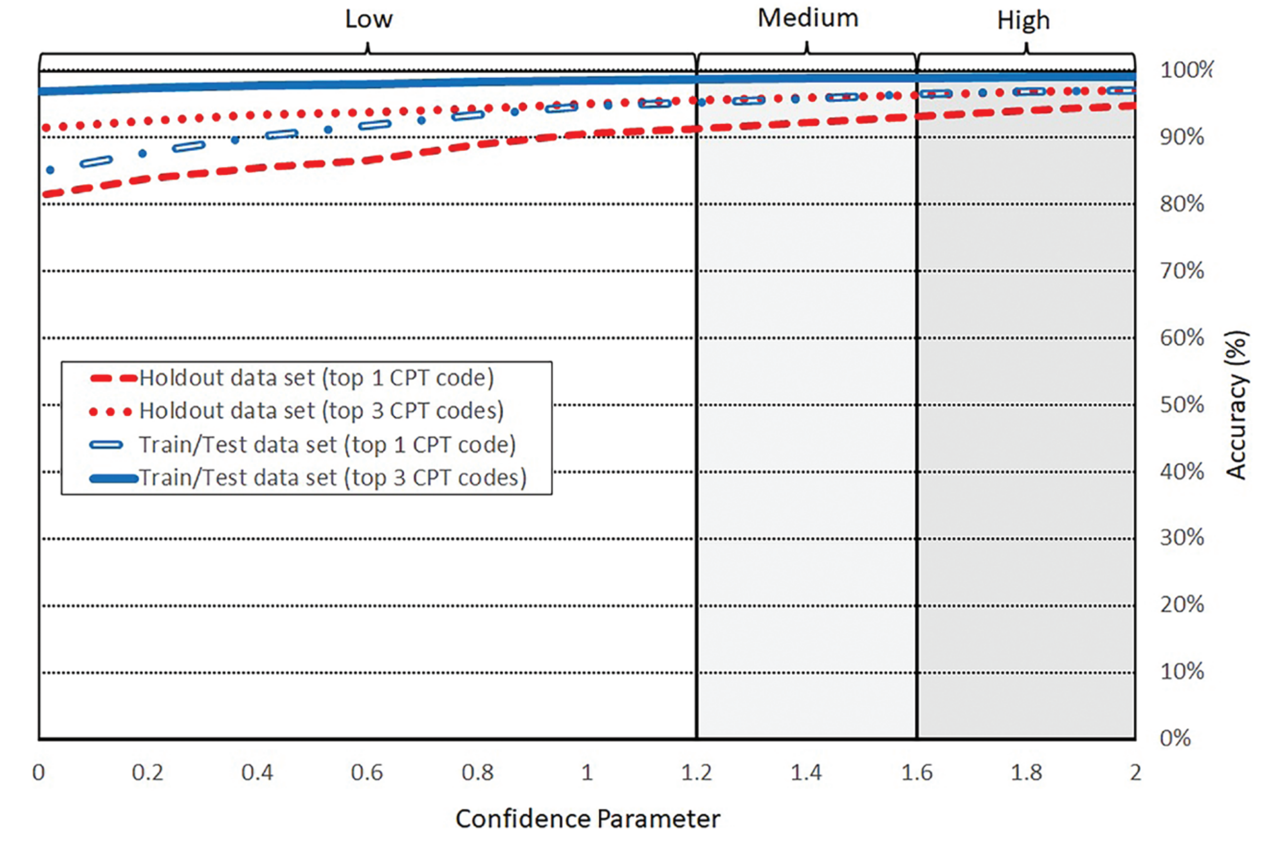

Fig. 2. Accuracy of current procedural terminology (CPT) code assignment as a function of confidence parameter. This graph shows the percentage accuracy of model CPT classification (*y* axis) for a given cutoff of confidence parameter (*x* axis) for the support vector machine model. The Train/Test and Holdout data set accuracies are plotted for both the first assigned CPT code (top 1 CPT code) and top three assigned CPT codes (top 3 CPT codes). High (≥ 1.6), Medium (1.6 > confidence parameter ≥ 1.2), and Low (< 1.2) areas are labeled above the figure. Confidence parameter is a derived measure of relative probability between best-fit and second-best-fit CPT classifications.

Results

The Train/Test data comprised 1,164,343 unique cases across 16 institutions and spanning 262 anesthesia CPT codes (table 1). The 2018 anesthesia CPT catalog consists of 285 unique codes.[1,39] The Holdout data set comprised 58,510 cases from a single institution and spanned 232 anesthesia CPT codes. In the Train/Test data set 36,356 cases were missing procedure text, representing 0.1% of the data. The Holdout data set had 17 such cases (less than 0.1% of the data). The Train/Test data set included 227 of the 232 codes contained in the Holdout data set; the five anesthesia CPT codes unique to the Holdout data set are described in Supplemental Digital Content 3 (http://links.lww.com/ALN/C204). Fifty-seven percent of patients were female. The mean age was 50 yr, and 8.5% were pediatric (age younger than 18 yr). Cases were primarily ASA II (46.5%) and ASA III (37.1%), and 4.5% were emergent.

Using CPT categories, codes were unevenly distributed between the data sets. Case distributions into each CPT grouping varied between individual institutions, but the distribution reflects the content of the overall Multicenter Perioperative Outcomes Group database. The Holdout data set was similar to the Train/Test data set (table 1). Because sample sizes are large, statistically significant differences were observed between data sets. Two body regions showed a relative sparsity: Burn Debridement (1,054 cases *vs.* 1 between the Train/Test and the Holdout data sets, respectively) and Other (0 cases *vs.* 97).

Primary Outcome Adjudication

Institution-assigned primary anesthesia CPTs were used as the reference standard labels when developing the models. Among the 500 cases from the Train/Test data set adjudicated by anesthesiologist manual review, 25 of 500 (5.0%) cases were found to be misclassified by primary anesthesia CPT in the source data set. Nine of 25 errors would have been correctly classified by the support vector machine model. A sample of 50 cases, including all 25 for which

Table 1. Key Metrics and Comparisons of the Two Data Sets Used in This Study (Train/Test and Holdout)

Category		Train/Test	Holdout	OR	*P* Value
Case demographics					
Unique anesthesia cases		1,164,343	58,510		
Unique anesthesia CPTs		262	232		
CPT categories					
Head	00100-00222	156,017 (13.4%)	14,934 (25.5%)	0.5	< 0.0001
Neck	00300-00352	53,302 (4.6%)	3,238 (5.5%)	0.8	< 0.0001
Thorax (chest, shoulder)	00400-00474	58,001 (5.0%)	2,746 (4.7%)	1.1	0.0061
Intrathoracic	00500-00580	57908 (5.0%)	3,778 (6.5%)	0.8	< 0.0001
Spine and spinal cord	00600-00670	36,520 (3.1%)	1,291 (2.2%)	1.4	< 0.0001
Upper abdomen	00700-00797	170,005 (14.6%)	7,646 (13.1%)	1.1	< 0.0001
Lower abdomen	00800-00882	227,202 (19.5%)	7,658 (13.1%)	1.6	< 0.0001
Perineum	00902-00952	105,208 (9.0%)	3,584 (6.1%)	1.5	< 0.0001
Pelvis (except hip)	01112-01190	4,904 (0.4%)	227 (0.4%)	1.0	0.9999
Upper leg (except knee)	01200-01274	35,094 (3.0%)	1,162 (2.0%)	1.5	< 0.0001
Knee and popliteal area	01320-01444	45,967 (3.9%)	1,502 (2.6%)	1.5	< 0.0001
Lower leg (below knee)	01462-01522	37,350 (3.2%)	1,217 (2.1%)	1.5	< 0.0001
Shoulder and axilla	01610-01682	24,076 (2.1%)	907 (1.6%)	1.3	< 0.0001
Upper arm and elbow	01710-01782	7,110 (0.6%)	408 (0.7%)	0.9	0.0129
Forearm, wrist, and hand	01810-01860	38,149 (3.3%)	1,269 (2.2%)	1.5	< 0.0001
Radiological procedure	01916-01936	45,378 (3.9%)	3,329 (5.7%)	0.7	< 0.0001
Burn debridement	01951-01953	1,054 (0.1%)	1 (<0.1%)	1.0	< 0.0001
Obstetric	01958-01969	61,098 (5.2%)	3,516 (6.0%)	0.9	< 0.0001
Other procedure	01990-01999	0 (0.0%)	97 (0.2%)	0.0	N/A
Patient demographics					
Female		659,272 (56.6%)	32,078 (54.8%)	1.1	< 0.0001
Age, yr		51 (22)	50 (23)	0.09*	
Pediatric (age <18 yr)		98,778 (8.5%)	6,549 (5.6%)	1.6	< 0.0001
ASA I		111,269 (9.6%)	6,307 (10.8%)	0.9	< 0.0001
ASA II		536,752 (46.5%)	25,998 (44.4%)	1.1	< 0.0001
ASA III		428,397 (37.1%)	23,095 (39.5%)	0.9	< 0.0001
ASA IV		75,230 (6.5%)	2,969 (5.1%)	1.3	< 0.0001
ASA V		1,600 (0.1%)	132 (0.2%)	0.5	< 0.0001
ASA VI		16 (<0.1%)	9 (<0.1%)	1.1	< 0.0001

Frequencies are displayed as percentages or means with SD, as appropriate. *P* values are calculated to evaluate differences between groups using chi-squared test for categorical features, Student *t* test for continuous features. CPT categories are defined by body region. Odds ratio (OR) thresholds for determining the effect size: small (OR ≤ 1.5), medium (1.5 < OR ≤ 2), and large (3 < OR). For SD: small (≤ 0.2), medium (0.2 < SD ≤ 0.5), large (0.5 < SD ≤ 0.8), and very large (0.8 < SD).

*SD. ASA, American Society of Anesthesiologists Physical Status classification; CPT, Current Procedural Terminology.

the institution-assigned CPT code was in error (per anesthesiologist review) and a random 25 for which the institution-assigned CPT code was correct, were validated by the University of Michigan Anesthesiology Department billing manager. The review by the anesthesia billing manager showed agreement in 22 of the 25 cases found to be incorrect and 25 of the 25 for cases found to be correct, for an overall 88% concordance with the anesthesia attending review.

Procedure Text and Natural Language Processing

Feature importance was used to gain insight into model classifications and potential improvements, but not used to evaluate model error. Procedure text was the most important feature used to classify anesthesia CPT codes. This text had an average word count of 10 words per case. The vocabulary size across all cases was 25,098 unique words. Most individual words were rare, occurring in less than 10 cases across both data sets, accounting for 19,159 (76.3%) of the vocabulary size. Unique medical word misspellings totaled 8,353. The top misspelled medical terms included "discectomy," "dilatation," "curettage," and "excision," along with longer terms such as "esophagogastroduodenoscopy" and "cholangiopancreatography." In all, 21.3% of cases contained at least one misspelled word that was subsequently corrected.

Machine Learning Model Parameters

In the support vector machine model, the average weight for each individual word in the procedure text was 7.9 whereas the average combined weight of all words within the procedure text was 337.5. Weights for other features were considerably lower than the combined procedural text weight: ASA Physical Status classification (6.1), text length (4.3), age (3.2), sex (2.1), emergent status (1.5), and case duration (1.5).

Table 2. Results of Five Machine Learning Models on the Train/Test Data Set

Machine Learning Model	Average Accuracy (95% CI) Top CPT Code
Random forest	82.0% (68.1% to 95.9%)
Support vector machine	87.9% (87.6% to 88.2%)
Extreme gradient boosting	87.9% (87.5% to 88.3%)
Long short-term memory	86.4% (83.5% to 89.3%)
Label-embedding attentive model	84.2% (84.1% to 84.3%)

Accuracies of the five machine learning models calculated from the Train/Test data set training tested 20 times using fivefold cross validation, shown with 95% CI. CPT, Current Procedural Terminology.

Train/Test Data Set

The highest overall accuracy was found with the support vector machine model (87.9%; 95% CI, 87.6% to 88.2%; table 2). Extreme gradient boosting (87.9%; 95% CI, 87.5% to 88.3%), and long short-term memory (86.4%; 95% CI, 83.5% to 89.3%), and the label-embedding attentive model (84.2%; 95% CI, 84.1% to 84.3%) were all more accurate than random forest modeling (82.0%; 95% CI, 68.1% to 95.9%). Using CPT categories to identify cases for which the random forest model demonstrated differential performance, there was a low of 70.7% for radiology procedures and a high of 92.0% for shoulder procedures. There was an observed positive relationship between the number of cases comprising a specific CPT code and the accuracy of the models for the CPT code, with a Pearson correlation of 0.72. Overall accuracy within the top three was 96.8% for support vector machine model and 94.0% for the label-embedding attentive model.

Confidence Parameters

The best performing model in the testing was the support vector machine model at 87.9% (95% CI, 87.6% to 88.2%), or a misclassification rate of 12%. However, through the use of confidence parameters assigned along with CPT code output, results were partitioned into identifiable groups and those with higher confidence parameters correlated with accuracy of CPT classification (Pearson correlations greater than 0.97; fig. 2). Cases within the High (confidence parameter at or above 1.6) category represented 47% of the data in testing (fig. 3) and yielded a 96.4% accuracy (fig. 2). At a more stringent confidence parameter of at least 2.0, first CPT classification accuracy increased to 97.1%, encompassed 39.3% of the cases, and accuracy at this confidence within the top three was 99.1%. For the label-embedding attentive model, there was a 94.4% accuracy in 62.2% of cases with a 98.2% top three accuracy.

Holdout Data Set Performance Metrics

Accuracy. The best performing machine learning model by overall accuracy in the Holdout data set was the support vector machine model (81.2%). When stratifying by confidence parameter metrics there was a 93.1% accuracy (fig. 2) for high confidence parameter (at least 1.6) encompassing 58.0% of the data (fig. 3). At the more stringent confidence, very high confidence parameter (at least 2.0) demonstrated a 94.7% accuracy and 48.0% data set coverage. Accuracy within the support vector machine model top three was 96.3% for the Holdout data set. The overall accuracy of the label-embedding attentive model was 82.1% for the Holdout data set, and accuracy within the top three was 94.6%. The label-embedding attentive model accuracy is improved to 95.0% for cases with confidence parameter at or above 1.6, encompassing 62.0% of the data set. This means that, using the label-embedding attentive model, CPTs were classified within the study's desired threshold (at least 95% accuracy) on 62% of the data. At the more stringent confidence (confidence parameter at or above 2.0) accuracy improved to 96.9% with a data set coverage of 48.3%.

When CPT codes were grouped by body region, we found that the label-embedding attentive model correctly identified the proper body region in 91.4% of its first-choice CPT classifications, whereas the support vector machine model correctly identified 93.1%. Furthermore, the label-embedding attentive model and support vector machine models correctly identified the proper body region in 97.5% and 97.7% of top-three choices, respectively.

Net Reclassification Index. After transformation of CPT codes to anesthesiology base unit values, the support vector machine model net reclassification index was 0.294 (95% CI, 0.270–0.318), indicating that the support vector machine model led to a 29.4% excess proportion of increased anesthesiology base unit values compared with original CPT code base unit values. Using a similar approach, the label-embedding attentive model net reclassification index was 0.343 (95% CI, 0.342% to 0.344%).

Calibration. After transformation of CPT codes to anesthesiology base unit values, calibration plot intercepts were −0.00 ($P = 0.855$) and −0.00 ($P = 0.995$) whereas calibration plot slopes were 0.849 ($P < 0.001$) and 0.845 ($P < 0.001$) for the support vector machine and the label-embedding attentive models, respectively. Details can be found in Supplemental Digital Content 4 (http://links.lww.com/ALN/C205).

Processing Time

The processing speed of the support vector machine model on the Holdout data set (58,510 cases) was 1.09 ± 0.05 s. Processing speeds were equivalent across all models.

Discussion

In this retrospective multicenter study, a machine learning–based approach to CPT code classification is described using commonly available perioperative electronic health

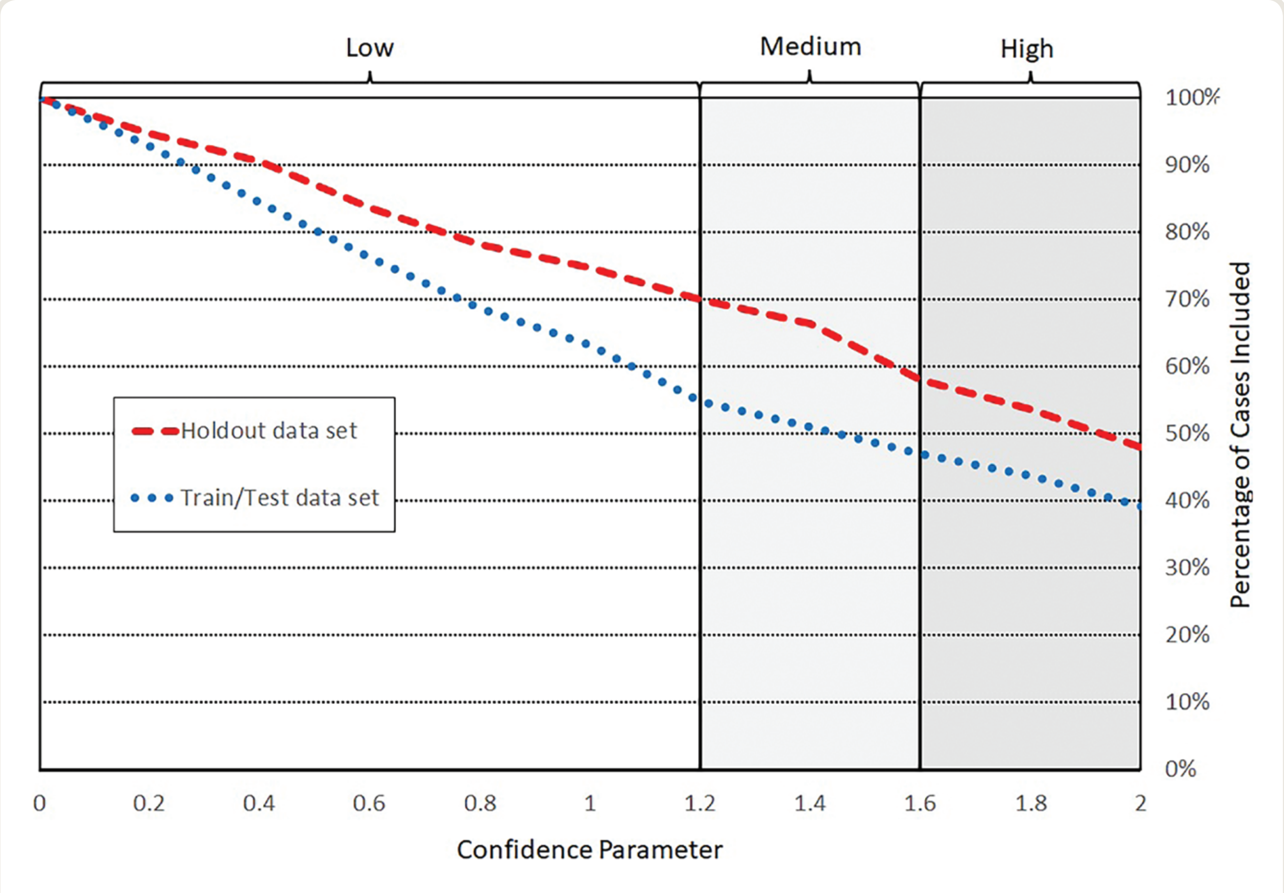

Fig. 3. Percentage case inclusion as a function of confidence parameter. This graph shows the percentage of cases included for model current procedural terminology (CPT) classification (*y* axis) for a given cutoff of confidence parameter (*x* axis) for the support vector machine model. The Train/Test and Holdout data set accuracies are plotted. High (≥ 1.6), Medium (1.6 > confidence parameter ≥ 1.2), and Low (< 1.2) areas are labeled above the figure. Confidence parameter is a derived measure of relative probability between best-fit and second best-fit CPT classifications.

record data. This study found important differences in accuracy between five machine learning techniques. Within training, the models studied showed a range of classification accuracy from 82% to 88%, a 50% difference in misclassification rate between the worst and best performing models. Within validation, an overall accuracy of 82.1% in the Holdout data set of the best performing model (label-embedding attentive model) was observed. When restricting to high confidence cases, (confidence parameter at or above 1.6) comprising 62% of cases within the data set, there was an augmented accuracy of 95.0%, the quality target for this study and eclipsing the most recently reported accuracy for fee-for-service payment within Centers for Medicare and Medicaid Services (91.9%).[7] The models developed in this study may offer a reduction in processing time and personnel resources required to perform these administrative tasks.

This approach is different from traditional computer-aided coding in the medical space: Whereas traditional approaches focus on automating transcription tasks, this study focuses on classification capabilities. The confidence parameter was created to stratify cases into groups to improve model utility. This allowed identification of cases with high classification confidence, which may enable reallocation of administrative or auditing resources to review cases for which ambiguity exists.

To investigate external validation, the support vector machine and label-embedding attentive models were tested on the Holdout data. Both models yielded lower overall accuracies (81.2% to 82.1%) for the Holdout data set relative to Train/Test, yet through the use of a confidence parameter, an identifiable 58.0% to 62.0% of cases with a confidence parameter at or above 1.6 demonstrated overall accuracies of 93.1 to 95.0%. These results are encouraging for the generalizability of the models—potentially owing to use of data from both academic and private hospitals across 16 medical centers. The machine learning models developed proved robust to unseen data at the holdout institution, with a broadly similar case mix, yet with some site-specific, idiosyncratic documentation and practice patterns.

For the remaining lower-confidence cases, the models can narrow assignment choices for medical billing specialists to selection between the top three choices. Top three choice narrows the classification task of billing specialists from 285 to 3, providing a shorter and prioritized list from which to choose. Top three accuracies were 94.6% to 96.8%. Thus, these models could aid coding personnel by providing a smaller subset of CPT assignment choices. In other instances, the models created in this study could be used for postassignment analysis as auditing tools used to identify discrepancies and potential coding errors by comparing manual and automated assignments. There is commonly a window for resubmission of CPT assignments, during which automation could help target efforts to reclaim lost revenues. Through an assessment of base unit values corresponding to the CPT codes, such automated models may identify cases commonly over- and under-billed, and may aid such auditing processes.

Given the promising results of this study, models developed from this work have been directly incorporated into the billing workflow at the University of Michigan for auditing and resubmission purposes. Beyond use at our single center, the CPT classification tool developed in this study has substantial applicability in the broader business practices of anesthesia care. Billing departments and vendors spend a considerable amount of time processing information for reimbursement and are slowed in an environment in which documentation errors are common. An estimated 15.7% of anesthesia cases contain at least one documentation error after first billing attempt, and the median time to correct documentation errors was 33 days.[41] Furthermore, 1.3% of all anesthetic cases went without reimbursement because of improper documentation and failure to correct errors. Within this study, medical misspellings accounted for 33.3% of the procedure text vocabulary, and 21.3% of cases contained at least one misspelled term. These tools could be used to refocus resources away from routine, high-confidence, CPT assignment and toward areas of more complex processing and auditing to further improve the speed and accuracy of the overall billing process. When deployed as a web application, the models are able to process more than one million cases in under 10 min. In the context of studies demonstrating anesthesiology practices gaining revenue *via* decreasing charge lag,[8] and reports demonstrating hospital operating margins between 2% and 3%, a machine learning classification approach represents an opportunity to reduce costs without compromising patient care.[42–44]

Additionally, the methods developed in this study may expedite CPT assignment for use in research and quality improvement projects. The classification models created enable near real-time anesthesia CPT assignment upon upload of core electronic health record data to a research or quality improvement coordinating center, freeing researchers and quality improvement champions from a dependency on billing data which may not be available in a timely manner.

Work remains to develop the full potential of billing aides like the CPT classification models created in this study. These tools require continued retraining as new information becomes available and updating when medical coding changes occur. Without updating over time, the tools will perform with gradual lower accuracy. One implementation concept is where the models can be trained with historical data when a new center begins to use them but retrained periodically when new data become available. Existing and new centers would benefit from novel data inclusion.

Study Limitations

This study has several important limitations which must be further explored.

1. Surgical CPTs were frequently unavailable from the contributing institutions, and of those providing surgical CPTs there was a similar or longer delay in availability from procedure date, compared with anesthesia CPT codes. Such lag times in surgical CPT coding preclude early crosswalking to anesthesia CPT codes and justify the approach based on procedure text used in this study.
2. Training sets derived from manual CPT assignment contain errors,[6] and a model trained on errors will invariably reproduce similar errors. In this study, through physician validation, there was a manual CPT assignment error rate of 5.0%; thus, models created in this study would benefit from audited and validated data sets to increase model assignment accuracy.
3. Bias from overfitting to individual or institution-specific procedure text assignments and billing practices may have existed within this study. To alleviate this bias, data from multiple centers were used in training, and an external validation was conducted on a Holdout data set.
4. Although natural language processing was used to correct many of the spelling and formatting errors in procedure text, creating this feature required manual physician review and there remained several additional instances that went uncorrected. Further text processing and expansion of acronyms can help align similar cases, improving model accuracy.
5. Among CPT codes for which the machine learning models demonstrate low or medium confidence, accuracy is not yet comparable with current standards.
6. Although some level of site-specific text characterization was required because of local site acronyms, standard lexicon tools for natural language processing were not used in this study, such as those available through the Unified Medical Language System,[45] potentially limiting the reproducibility of the models.
7. Because base unit values were not unique to each CPT code, it was possible for model outputs to yield an incorrect CPT yet correct base unit value, thus limiting the fidelity of net reclassification index and model calibration assessments.

8. Although this study demonstrates rapid data processing and has potential for real-time classification of anesthesia CPT codes, these models have not been thoroughly analyzed in practice. The group plans to test these capabilities through prospective application of CPT classification to anesthesia quality improvement measures reliant on CPT codes.
9. Sparsity remains an issue with large data predictive modeling. In cases that were not well represented in Train/Test data set, the models demonstrated decreased accuracy. The data sets used to create these models contained sparse procedural information, and it is likely that accuracy would improve with inclusion of additional data, such as operative notes.

Conclusion and Future Directions

In summary, this study describes a rapid automated classification model for anesthesia CPT codes, with an accuracy comparable with current standards in a high-confidence subset of cases, and processing time far eclipsing current billing practices. These findings may serve to reduce the burden of manual coding of more common cases, and may increase efficiency within the billing cycle and aid processes that rely on billing data. These results broadly demonstrate the potential for machine learning and natural language processing–based classification models in healthcare operations.

Future applications include automation of high-confidence CPT assignment to enable redistribution of manual efforts, and workflow integration for classification decision support. Because similar difficulties in reimbursement processes exist throughout the hospital, the methods to create these models could be used for classification of other medical billing codes such as surgical CPT and International Classification of Diseases. In classifying surgical and anesthesia CPT as well as International Classification of Diseases, it is conceivable to create a system automating the majority of the procedural billing process.

Acknowledgments

The authors gratefully acknowledge the valuable contribution to protocol development and final article review by the Multicenter Perioperative Outcomes Group Perioperative Clinical Research Committee.

Research Support

All work and partial funding can be attributed to the Department of Anesthesiology, University of Michigan Medical School (Ann Arbor, Michigan). Research reported in this publication was supported by grant No. T32GM103730 from the National Institute for General Medical Sciences of the National Institutes of Health (Bethesda, Maryland; to Drs. Burns and Colquhoun) and by grant No. K01HL141701 from the National Heart, Lung, and Blood Institute of the National Institutes of Health (to Dr. Mathis). The content of this study is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This project received additional funding through the Michigan Translational Research and Commercialization (MTRAC) for Life Sciences Innovation Hub, a partnership between the University of Michigan's Fast Forward Medical Innovation and Office of Technology Transfer (Ann Arbor, Michigan), and the State of Michigan. Support for underlying electronic health record data collection was provided, in part, by Blue Cross Blue Shield of Michigan (BCBSM; Detroit, Michigan) and Blue Care Network as part of the BCBSM Value Partnerships program for contributing hospitals in the state of Michigan. Although BCBSM and the Multicenter Perioperative Outcomes Group work collaboratively, the opinions, beliefs, and viewpoints expressed by the authors do not necessarily reflect the opinions, beliefs, and viewpoints of BCBSM or any of its employees.

Competing Interests

This work has been declared through the University of Michigan Office of Tech Transfer, and a provisional patent (U.S. Provisional Application No. 62/791,257) has been filed related to the work presented in this study. Dr. Shah declares receiving consulting fees from Merck & Co. Inc. (Kenilworth, New Jersey), The 37 Company (The Netherlands), and Medtronic (Dublin, Ireland).

Correspondence

Address correspondence to Dr. Burns: Department of Anesthesiology, University of Michigan, 1500 East Medical Center Drive, 1H247 UH, SPC 5048, Ann Arbor, Michigan 48109-5048. mlburns@med.umich.edu. Information on purchasing reprints may be found at www.anesthesiology.org or on the masthead page at the beginning of this issue. Anesthesiology's articles are made freely accessible to all readers, for personal use only, 6 months from the cover date of the issue.

References

1. 2018 CROSSWALK Book: A Guide for Surgery/Anesthesia CPT Codes. Schaumburg, Illinois, American Society of Anesthesiologists, 2017
2. CPT 2018, Current procedural terminology 2018: Professional edition. Chicago, Illinois, American Medical Association, 2017
3. Polsky D, Candon M, Saloner B, Wissoker D, Hempstead K, Kenney GM, Rhodes K: Changes in primary care access between 2012 and 2016 for new patients with Medicaid and private coverage. JAMA Intern Med 2017; 177:588–90

4. Holt J, Warsy A, Wright P: Medical decision making: Guide to improved CPT coding. South Med J 2010; 103:316–22
5. 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; 319:691–7
6. Henderson R, Nielsen, KC, Klien, SM, Pietrobon, R: Miscoding rates for professional anesthesia billing: Trial results - software solution. Electron J Health Informatics 2010; 5(2)
7. (CERT) CERT: 2018 Medicare Fee-for-Service Supplemental Improper Payment Data. 2018. Available at: https://www.cms.gov/Research-Statistics-Data-and-Systems/Monitoring-Programs/Medicare-FFS-Compliance-Programs/CERT/Downloads/2018MedicareFFSSuplementalImproperPaymentData.pdf. Accessed March 14, 2019.
8. Reich DL, Kahn RA, Wax D, Palvia T, Galati M, Krol M: Development of a module for point-of-care charge capture and submission using an anesthesia information management system. ANESTHESIOLOGY 2006; 105:179–86; quiz 231–2
9. Liu JB, Liu Y, Cohen ME, Ko CY, Sweitzer BJ: Defining the intrinsic cardiac risks of operations to improve preoperative cardiac risk assessments. ANESTHESIOLOGY 2018; 128:283–92
10. Oh J, Makar M, Fusco C, McCaffrey R, Rao K, Ryan EE, Washer L, West LR, Young VB, Guttag J, Hooper DC, Shenoy ES, Wiens J: A Generalizable, data-driven approach to predict daily risk of clostridium difficile infection at two large academic health centers. Infect Control Hosp Epidemiol 2018; 39:425–33
11. Zhao S, Dong X, Shen W, Ye Z, Xiang R: Machine learning-based classification of diffuse large B-cell lymphoma patients by eight gene expression profiles. Cancer Med 2016; 5:837–52
12. Erickson BJ, Korfiatis P, Akkus Z, Kline TL: Machine learning for medical imaging. Radiographics 2017; 37:505–15
13. Connor CW. Artificial intelligence and machine learning in anesthesiology. ANESTHESIOLOGY 2019; 131:1346–59
14. Mathur P, Burns ML: Artificial intelligence in critical care. Int Anesthesiol Clin 2019; 57:89–102
15. Lee HC, Ryu HG, Chung EJ, Jung CW: Prediction of bispectral index during target-controlled infusion of propofol and remifentanil: A deep learning approach. ANESTHESIOLOGY 2018; 128:492–501
16. Hatib F, Jian Z, Buddi S, Lee C, Settels J, Sibert K, Rinehart J, Cannesson M: Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. ANESTHESIOLOGY 2018; 129:663–74
17. Lee CK, Hofer I, Gabel E, Baldi P, Cannesson M: Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. ANESTHESIOLOGY 2018; 129:649–62
18. Wang G, Li C, Wang W, Zhang Y, Shen D, Zhang X, Henao R, Carin L: Joint embedding of words and labels for text classification. ArXiv e-prints. 2018. Available at: https://ui.adsabs.harvard.edu/#abs/2018arXiv180504174W. Accessed May 1, 2018.
19. Shi H, Xie P, Hu Z, Zhang M, Xing EP: Towards automated ICD coding using deep learning. ArXiv e-prints. 2017. Available at: https://ui.adsabs.harvard.edu/#abs/2017arXiv171104075S. Accessed November 1, 2017.
20. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M: Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. J Med Internet Res 2016; 18:e323
21. (MPOG) MPOG: Perioperative Clinical Research Committee (PCRC). 2019. Available at: https://mpog.org/pcrc/. Accessed May 21, 2019.
22. Freundlich RE, Kheterpal S: Perioperative effectiveness research using large databases. Best Pract Res Clin Anaesthesiol 2011; 25:489–98
23. Kheterpal S: Clinical research using an information system: The multicenter perioperative outcomes group. Anesthesiol Clin 2011; 29:377–88
24. Sun E, Mello MM, Rishel CA, Vaughn MT, Kheterpal S, Saager L, Fleisher LA, Damrose EJ, Kadry B, Jena AB; Multicenter Perioperative Outcomes Group (MPOG): Association of overlapping surgery with perioperative outcomes. JAMA 2019; 321:762–72
25. Lee LO, Bateman BT, Kheterpal S, Klumpner TT, Housey M, Aziz MF, Hand KW, MacEachern M, Goodier CG, Bernstein J, Bauer ME; Multicenter Perioperative Outcomes Group Investigators: Risk of epidural hematoma after neuraxial techniques in thrombocytopenic parturients: A report from the Multicenter Perioperative Outcomes Group. ANESTHESIOLOGY 2017; 126:1053–63
26. Larach MG, Klumpner TT, Brandom BW, Vaughn MT, Belani KG, Herlich A, Kim TW, Limoncelli J, Riazi S, Sivak EL, Capacchione J, Mashman D, Kheterpal S, Kooij F, Wilczak J, Soto R, Berris J, Price Z, Lins S, Coles P, Harris JM, Cummings KC 3rd, Berman MF, Nanamori M, Adelman BT, Wedeven C, LaGorio J, McCormick PJ, Tom S, Aziz MF, Coffman T, Ellis TA 2nd, Molina S, Peterson W, Mackey SC, van Klei WA, Ginde AA, Biggs DA, Neuman MD, Craft RM, Pace NL, Paganelli WC, Durieux ME, Nair BJ, Wanderer JP, Miller SA, Helsten DL, Turnbull ZA, Schonberger RB; Multicenter Perioperative Outcomes Group: Succinylcholine use and dantrolene availability for malignant hyperthermia treatment: Database analyses and systematic review. ANESTHESIOLOGY 2019; 130:41–54
27. Cavnar WaT, JM: N-gram based text categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium

on Document Analysis and Information Retrieval (Las Vegas, NV, 1994). 1994:161–75

28. Wang S, Manning CD: Baselines and bigrams: Simple, good sentiment and topic classification. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2; 2012; Jeju Island, Korea
29. Mikolov T, Chen K, Corrado G, Dean J: Efficient Estimation of Word Representations in Vector Space. ArXiv e-prints. 2013. Available at: https://ui.adsabs.harvard.edu/#abs/2013arXiv1301.3781M. Accessed January 1, 2013.
30. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J: Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems, 2013, pp. 3111–9
31. Moen SP, Ananiadou TS: Distributional semantics resources for biomedical text processing. Proceedings of LBM 2013; Dec:39–44
32. Ho TK: Random decision forests. Paper presented at: Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, Canada, August 14–16, 1995, pp. 278–82
33. Hochreiter S, Schmidhuber J: Long short-term memory. Neural Comput 1997; 9:1735–80
34. Chen T, Guestrin C: Xgboost: A scalable tree boosting system. Paper presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, August 2016, pp. 785–94
35. Cortes C, Vapnik V: Support-vector networks. Machine learning 1995; 20(3):273–97
36. Kim Y: Convolutional neural networks for sentence classification. CoRR 2014;abs/1408.5882
37. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW: Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology 2010; 21:128–38
38. Sugiyama M, Suzuki T, Kanamori T: Density ratio estimation in machine learning. Cambridge, United Kingdom, Cambridge University Press, 2012
39. Relative Value Guide Book: A Guide for Anesthesia Values. Schaumburg, Illinois, American Society of Anesthesiologists, 2018
40. Consultants AB: Anesthesia CPT Code Ranges, by Area of the Body. 2016. Available at: http://www.anesthesiallc.com/publications/cpt-codes-for-anesthesia-procedures-services. Accessed December 17, 2018.
41. Spring SF, Sandberg WS, Anupama S, Walsh JL, Driscoll WD, Raines DE: Automated documentation error detection and notification improves anesthesia billing performance. Anesthesiology 2007; 106:157–63
42. Board A: Hospital profit margins declined from 2015 to 2016, Moody's finds. 2017. Available at: https://www.advisory.com/daily-briefing/2017/05/18/moodys-report. Accessed March 9, 2019.
43. Cohen AEaJK: Becker's Hospital Review: 230 hospital benchmarks 2017. Available at: https://www.beckershospitalreview.com/lists/230-hospital-benchmarks-2017. Accessed May 3, 2017.
44. Catalyst H: How Hospital Financial Transparency Drives Operational and Bottom Line Improvements. 2017. Available at: https://www.healthcatalyst.com/success_stories/improved-hospital-profit-margins. Accessed November 12, 2018.
45. (US) NLoM: UMLS® Reference Manual. SPECIALIST Lexicon and Lexical Tools. 2009. Available at: https://www.ncbi.nlm.nih.gov/books/NBK9676/ and https://www.nlm.nih.gov/research/umls/about_umls.html. Accessed July 16, 2019.