Prediction of American Society of Anesthesiologists Physical Status Classification from Preoperative Clinical Text Narratives Using Natural Language Processing

Philip Chung, MD, MS[1] Christine T. Fong, MS[1] Andrew M. Walters, MD[1]

Meliha Yetisgen, PhD[2] Vikas N. O'Reilly-Shah, MD, PhD[1]

[1]University of Washington, Department of Anesthesiology & Pain Medicine
[2]University of Washington, Department of Biomedical & Health Informatics and Department of Linguistics

Abstract

Importance: Large volumes of unstructured text notes exist for patients in electronic health records (EHR) that describe their state of health. Natural language processing (NLP) can leverage this information for perioperative risk prediction.
Objective: Predict a modified American Society of Anesthesiologists Physical Status Classification (ASA-PS) score using preoperative note text, identify which model architecture and note sections are most useful, and interpret model predictions with Shapley values.
Design: Retrospective cohort analysis from an EHR.
Setting: Two-hospital integrated care system comprising a tertiary/quaternary academic medical center and a level 1 trauma center with a 5-state referral catchment area.
Participants: Patients undergoing procedures requiring anesthesia care spanning across all procedural specialties from January 1, 2016 to March 29, 2021 who were not assigned ASA VI and also had a preoperative evaluation note filed within 90 days prior to the procedure.
Exposures: Each procedural case paired with the most recent anesthesia preoperative evaluation note preceding the procedure.
Main Outcomes and Measures: Prediction of a modified ASA-PS from preoperative note text. We compared 4 different text classification models for 8 different input text snippets. Performance was compared using area under the receiver operating characteristic curve (AUROC) and area under the precision recall curve (AUPRC). Shapley values were used to explain model predictions.
Results: Final dataset includes 38566 patients undergoing 61503 procedures. Prevalence of ASA-PS was 8.81% for ASA I, 31.4% for ASA II, 43.25% for ASA III, and 16.54% for ASA IV-V. The best performing models were the BioClinicalBERT model on the truncated note task (macro-average AUROC 0.845) and the fastText model on the full note task (macro-average AUROC 0.865). Shapley values reveal human-interpretable model predictions.
Conclusions and Relevance: Text classification models can accurately predict a patient's illness severity using only free-form text descriptions of patients without any manual data extraction. They can be an additional patient safety tool in the perioperative setting and reduce manual chart review for medical billing. Shapley feature attributions produce explanations that logically support model predictions and are understandable to clinicians.

Introduction

Models to assess adverse event risk are indispensable tools in the arsenal of the perioperative clinician; guiding decision-making with respect to prehabilitation, preoperative testing, intraoperative management strategy, postoperative disposition, and more. These models base risk assessments on a limited number of discrete predictor variables. Examples include the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP) Surgical Risk Calculator,[1,2] the Revised Cardiac Risk Index (RCRI),[3,4] and the Gupta Perioperative Risk for Myocardial Infarction or Cardiac Arrest (MICA).[5] While well validated, classification of these predictor variables in many cases require expert clinician chart review and patient assessment. The use of these manually abstracted discrete data elements works well in the context of individual patient assessment. However, without *a priori* discretization of these elements, usage of these risk assessment tools for other purposes are limited. These purposes might include automation of risk assessment for patient safety purposes, perioperative population health assessment, benchmarking within and across health systems, or simply for triage of patients in assessing the need for a preoperative clinic visit.

Machine learning and natural language processing (NLP) techniques, coupled with adoption of electronic health records (EHR), and widespread availability of high-performance computational resources offer new avenues for perioperative risk stratification whereby unstructured data sources such as medical note free-form text may be directly input into prediction models without abstracting data elements. Unlike historical keyword-based approaches, modern NLP techniques using large pretrained language models are able to account for inter-word dependencies across the entire text sequence and have been shown to achieve state of the art performance on a variety of NLP tasks[6–9] including text classification.[10,11] However it is unknown whether these techniques can be successfully applied to perioperative risk prediction. In particular, we investigate risk prediction using only unstructured text notes written by clinicians drawn from the EHR, which often contain narratives that richly and concisely describe a nuanced clinical picture of the patient while simultaneously prioritizing the clinician's pertinent concerns.

The American Society of Anesthesiologists Physical Status (ASA-PS) score [12,13] is a categorical clinician-driven assessment of patient periprocedural risk. ASA-PS has been shown to be an independent predictor of mortality and patient outcomes[14–19] despite well-described interrater variability in ASA-PS classification.[20,21] In this study, we investigated prediction of ASA-PS directly from free-form text taken from an anesthesia preoperative evaluation note using four different text classification approaches that span the spectrum of historical and modern techniques: (1) random forest[22] with n-gram and term-frequency inverse document frequency (TFIDF) transform,[23] (2) support vector machine[24] with n-gram and TFIDF transform, (3) fastText[25,26] word vector model, and (4) BioClinicalBERT deep neural network language model. We compared the model's prediction with the ASA-PS assigned by the anesthesiologist on the day of surgery and hypothesized that advanced NLP modeling techniques would provide improved predictions of ASA-PS score as compared to simpler models.

Methods

This retrospective study of routinely collected health records data was approved by the University of Washington Human Subjects Division with a waiver of consent. This study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guideline[27] and other guidelines specific to machine learning projects.[28–30] eFigure 1 depicts a flow diagram of study design.

Study Cohort

Inclusion criteria were patients who had a procedure requiring anesthesia at the University of Washington Medical Center or Harborview Medical Center from January 1, 2016 – March 29, 2021 where the patient also had an anesthesia preoperative evaluation note filed up to 6 hours after the anesthesia end time. This 6-hour grace period reflects the reality that in some urgent or emergency situations or due to EHR behavior, text documentation may be time stamped out of order.

The note must have contained the following sections: History of Present Illness (HPI), Past Medical and Surgical History (PMSH), Review of Systems (ROS), and Medications; notes missing at least one of these sections were excluded. Cases must have had a recorded value for ASA-PS assigned by the anesthesiologist of record, a free-form text Procedure description, and a free-form text Diagnosis description; cases missing at least one of these values are excluded.

A unit of analysis is defined as a single case with an anesthesia preoperative evaluation note filed within 90 days of the procedure. This unit was chosen because ASA-PS is typically recorded on a per-case basis by the anesthesiologist to reflect the patient's pre-anesthesia medical comorbidities at the time of the procedure. Likewise, preoperative evaluation notes filed >90 days before the case are not considered to reflect the patient's state of health so are excluded. Data was randomly split 70%-10%-20% into training, validation, and test datasets respectively. Patients with multiple cases were randomized into a single data split to avoid information leakage between the three datasets. New case number identifiers were generated for this study and used to refer to each case.

Outcomes

The outcome variable is a modified ASA-PS with valid values of ASA I, ASA II, ASA III, ASA IV-V. ASA V cases are extremely rare, resulting in class imbalances that affect model training and performance. Thus ASA IV and V were combined into a compound class "IV-V". ASA VI organ procurement cases are excluded. The final categories retain the spirit of the ASA-PS for perioperative risk stratification and resembles the original ASA-PS devised by Saklad in 1941.[12,31] The emergency surgery modifier "E" was discarded.

Predictors and Data Preparation

Free-form text from the anesthesia preoperative evaluation note is organized into many sections. Regular expressions are used to extract HPI, PMSH, ROS, and medications from the note. While diagnosis and procedure sections exist within the note, they were less frequently documented than in the procedural case booking data from the surgeon. Therefore, free-form text for these sections were taken from the case booking. Newline characters and whitespaces were removed from the text. Note section headers were excluded so that only the body of text from each section is included. We used text from each section to train models for ASA-PS prediction, resulting in 8 prediction tasks: Diagnosis, Procedure, HPI, PMSH, ROS, Medications (Meds), Note, Truncated Note (Note512). “Note” refers to using the whole note text as the predictor to train a model. When BioClinicalBERT is applied to the “Note” task, the WordPiece tokenizer[32–34] truncates input text to 512 tokens. This truncation does not occur for other models. For equitable comparison across models, we define the “Note512” task, which truncates the note text to the first 512 tokens used by the BioClinicalBERT model.

Statistical Analysis and Modeling

Four model architectures with different conceptual underpinnings were trained: (1) Random forest (RF),[22] (2) Support vector machine (SVM),[24] (3) fastText,[25,26], and (4) BioClinicalBERT.[35] Each model architecture was trained on each of the 8 prediction tasks for a total of 32 final models.

Each model was trained on the training dataset. Model hyperparameters were tuned using Tune[36] with the BlendSearch[37,38] algorithm to maximize Matthew’s Correlation Coefficient (MCC) computed on the validation dataset. The number of hyperparameter tuning trials was selected to be 20 times the number of model hyperparameters with early stopping if the MCC of the last 3 trials reaches a plateau with standard deviation <0.001. The best model was then evaluated on the held-out test dataset. Details on the approach taken for each of the four model architectures is available in supplemental methods.

Baseline Models

Two baseline models were created for comparison: a random classifier model and an age classifier model. The random classifier model generates a random prediction without using any features, thus serving as a negative control baseline. The age classifier model is a simple multiclass logistic regression model with cross-entropy loss and L2 penalty that uses age to directly predict the modified ASA-PS outcome variable. Defaults were used for all other model parameters. Both baselines were implemented using Scikit-learn.

Evaluation Metrics

Final models were evaluated on the held-out test dataset by computing both class-specific and class-aggregate performance metrics. Class-specific metrics include: receiver operator characteristic (ROC) curve, area under receiver operator curve (AUROC), precision-recall curve, area under precision-recall curve (AUPRC), precision (positive predictive value), recall (sensitivity), and F1. Class-aggregate performance metrics include MCC and AUCmu,[39] a

multiclass generalization of the binary AUROC. Additionally, macro-average AUROC, AUPRC, precision, recall and F1 were also computed.

Model Interpretability and Error Analysis

4-by-4 contingency tables were generated to visualize the distribution of model errors. Catastrophic errors were defined as cases where the model predicts ASA IV-V but the anesthesiologist assigned ASA I, or vice versa. For catastrophic errors made by the BioClinicalBERT model with the Note512 task, three new anesthesiologist raters independently assigned an ASA-PS based on only the input text from the Note512 task. These new ASA-PS ratings were compared against the original anesthesiologist's ASA-PS as well as the model prediction's ASA-PS.

The SHAP[40] python package was used to train a Shapley values feature attribution model on the test dataset to understand which words support prediction of each modified ASA-PS outcome variable. An analysis of model errors with Shapley value feature attributions was reviewed for each of the catastrophic error examples with representative examples included in the manuscript. Shapley values for predicting each ASA-PS are visualized as a heatmap over text examples. Text examples are de-identified by replacing ages, dates, names, locations, and entities with pseudonyms to achieve data obfuscation while preserving structural similarity to the original passage.

Results

Our study comprised 38,566 patients undergoing 61,503 procedures with 46,275 notes. Baseline patient, procedure, and note characteristics are described in Table 1. A flow diagram describing dataset creation is shown in eFigure 2.

AUROC for each model architecture and task is shown in Table 2; AUPRC is shown in eTable 1; AUCμ and MCC is shown in eTable 2. RF, SVM, and fastText perform best using the entire note compared to note sections. Tasks with longer text snippets yielded better performance–HPI, ROS and Meds sections result in better model performance as compared to Diagnosis, Procedure, and PMSH. On the Note task, fastText performs the best. On the Note512 task, BioCinicalBERT performs the best.

Direct comparison of models is most appropriate using the Note512 task since all models are given the same information content. For this task, BioClinicalBERT has better class-aggregate performance across AUROC, AUPRC, AUCmu, MCC, F1 (eTable 3), recall (sensitivity) (eTable 5) metrics while the fastText model has better precision (positive predictive value) (eTable 4). Class-specific metrics also reflect this finding and show that fastText has high recall for ASA II and III, the most prevalent classes, but recall for ASA I and IV-V is considerably lower. BioClinicalBERT has similar or better AUROC and AUPRC across all the ASA-PS classes. This is also seen in the ROC curves (eFigure 4) and the precision-recall curves (eFigure 5), in which the BioClinicalBERT model generally shows better performance across most thresholds.

Figure 1 depicts 4-by-4 contingency tables to visualize distribution of model errors on the Note512 task. When erroneous predictions occur, they are typically adjacent to the ASA-PS assigned by the original anesthesiologist. For catastrophic errors made by the BioClinicalBERT model on the Note512 task, ASA-PS ratings from the three new anesthesiologist raters show greater concordance with the model's predictions than the original anesthesiologist's assignment (Figure 2).

Shapley values in Figure 3 provide clinically plausible explanations for model explanations, highlighting the directional probability of how specific input text contributes to predicting a specific ASA-PS. These feature attributions often provide clinically plausible explanations for why a model is making a wrong prediction and allows the clinician to evaluate the evidence the model is considering. Additional examples shown in eFigure 6, eFigure 7, eFigure 8, eFigure 9.

Discussion

Text classification techniques have undergone substantial evolution over the past decade. RF and SVM represent more rudimentary approaches that utilize bag-of-words and n-grams. These techniques are sensitive to word misspellings, cannot easily account for word order, have difficulty in capturing long-range references within sentences, and have difficulty in representing different meanings of a word when the same word appears in different contexts.[41–46] Modern NLP techniques have overcome many of these challenges with: vector space representation of words[25,26,47–49] and subword components[26,32,33,50] as seen in the fastText model, attention mechanism[51,52], and pretrained deep autoregressive neural networks[53–55] such as transformer neural networks[56]. This has resulted in successful large language models such as BERT[34,57] and the domain-specific BioClinicalBERT[35].

Longer text length provides more information for the model to make an accurate prediction. Even though text snippets such as Diagnosis or Procedure may have high relevance for the illness severity of the patient, the better performance on longer input text sequences indicate that more information is generally better. This is similar to what is observed in the multifaceted practice of clinical medicine–where a patient's overall clinical status is often better understood as the sum of many weaker but synergistic signals rather than a single descriptor. The limited input sequence length for BioClinicalBERT creates a performance ceiling as it limits the amount of information available to the model. Comparing Note and Note512 tasks, all other models that can utilize the full note have better performance when this input length is lifted with fastText being the top performer. These findings suggest that future development of a large language model similar to BioClinicalBERT capable of accepting a longer input context would likely have superior performance characteristics. fastText requires significantly less compute resources for model training and inference compared to BioClinicalBERT and remains a good option in lower resource settings. RF and SVM were our worst performing models, confirming that modern word vector and language model-based approaches are superior.

There is significant variability on the length and quality of clinical free-form text narrative written in the note, especially in the HPI section which is typically a clinician's narrative of the patient's

medical status and need for the procedure. In some cases, the HPI section contains one or two words in length (eFigure 8), whereas in other cases it is a rich narrative (eFigure 6, eFigure 9). We believe that relatively poor performance in the ASA-PS prediction using HPI alone is a consequence of variability in documentation, as the model may have limited information for prediction if the note text does not richly capture the clinical scenario.

Models rarely make catastrophic errors. Erroneous predictions are typically adjacent to the ASA-PS assigned by the anesthesiologist, suggesting the model is making appropriate associations between freeform text predictors and the outcome variable (Figure 1). Examination of catastrophic errors from the BioClinicalBERT model on the Note512 task reveal that for both types of catastrophic errors–model predicts ASA I and original anesthesiologist assigned ASA IV-V, as well as the converse–we find that new anesthesiologist raters show greater concordance with the model predictions rather than the original anesthesiologist (Figure 2). Many of the catastrophic errors occurred with emergency cases. Shapley feature attributions for one of these catastrophic errors in Figure 3 reveal that in some cases the original anesthesiologist may have made the wrong assignment, or may have written a note that does not reflect the true clinical scenario. In this example, the original anesthesiologist assigned the case ASA IV-V, but the model predicted I. Feature attributions show the BioClinicalBERT model correctly identifies pertinent negatives on trauma exam, normal hematocrit of 33, and normal Glasgow Coma Scale (GCS) of 15 to all support a prediction for ASA I and against ASA IV-V. [58] In fact, all new anesthesiologist raters agree with the model rather than the original anesthesiologist. Examples like this suggest that the model performance may be underestimated by our evaluation metrics since our ground truth test set contains imperfect ASA-PS assignments. It illustrates how the model is robust against potentially faulty labels and has learned to make clinically appropriate ASA-PS predictions based on the input text

Shapley feature attributions reveal that the model is able to identify indirect indicators of a patient's illness severity. For example, subcutaneous heparin is often administered for bed-bound inpatients to prevent the development of deep vein thrombosis. eFigure 8 depicts an example where the model learns to associate mention of subcutaneous heparin in the medication list with a higher ASA-PS, likely because hospitalized patients are generally more ill than outpatients who present to the hospital for same-day surgery. Similarly, the model learns the association between the broad spectrum antibiotic ertapenem with a higher ASA-PS as compared to narrow spectrum or prophylactic antibiotics such as metronidazole or cefazolin. These observations show that the model is able to identify and link these subtle indicators to a patient's illness severity. Shapley value feature attributions prove to be an effective tool that enables clinicians to understand how a model makes its prediction from text predictors.

Limitations

Our dataset is derived from a real-world EHR used to provide clinical care and includes human and computer generated errors. These issues include data entry and spelling, the use of abbreviations, references to other notes and test results not available to the model, and automatically generated/inserted text as part of a note template. The BioClinicalBERT model is limited to an input sequence of 512 tokens; future investigation is needed to understand if

long-context large language models can achieve better performance. We also did not explore more advanced NLP models such as those that perform entity and relation extraction, which may further enhance the prediction performance. Finally, the ASA-PS is known to have only moderate interrater agreement among human anesthesiologists.[20,21] Consequently, a perfect classification on this task is not possible since the ground truth labels derived from the EHR encapsulate this interrater variability. Further investigation is needed to explore the prediction of other outcome variables which may be less subject to interrater variability.

Conclusions

NLP models can accurately predict a patient's illness severity using only free-form text descriptions of patients without any manual data extraction. They can be automatically applied to entire panels of patients and serve as a perioperative risk stratification and clinical decision support tool to ensure patient safety. Illness severity predictions may also be used to reduce manual chart review overhead for medical billing. Shapley feature attributions produce explanations that logically support model predictions and are understandable to clinicians.

Other Information

Funding Support

Computational resources for this project were funded by the Azure Cloud Compute Credits grant program from the University of Washington eScience Institute and Microsoft Azure. The University of Washington Department of Anesthesiology and Pain Medicine Bonica Scholars program provided financial support for this work.

Data Access, Responsibility, and Analysis

Philip Chung had full access to all the data in the study and takes responsibility for the integrity of the data and accuracy of the data analysis.

Data Sharing Statement

Data will not be shared because the text dataset derived from electronic health records comprises personal identifiable information (PII) and protected health information (PHI). Code for experiments and results is publicly available at https://github.com/philipchung/nlp-asa-prediction.

Acknowledgement Section

The authors would like to acknowledge: University of Washington Anesthesia Department's Perioperative & Pain initiatives in Quality Safety Outcome group for assistance on data extraction and initial compute resources for data exploration, University of Washington Department of Medicine for computational environment support, Roland Lai and Robert Fabiano from University of Washington Research IT for creating a digital research environment within the Microsoft Azure Cloud where model development and experiments were performed, and the University of Washington Biomedical Natural Language Processing group for providing early feedback on experimental design and results.

Author Contributions

Philip Chung: conception, experimental design, data acquisition, model development, data analysis, data interpretation
Christine Fong: data acquisition, data interpretation
Andrew Walters: data acquisition, data interpretation
Meliha Yetisgen: experimental design, data analysis, data interpretation
Vikas O'Reilly-Shah: experimental design, data analysis, data interpretation

Supplemental Methods

Details on the approach taken for each of the four model architectures.

Random Forest

Text input was preprocessed into a unigram and bigram count matrix followed by TFIDF transform.[23] Random forest classifier from the Scikit-learn[59] python library was used with minimization of gini impurity objective function and weighting each outcome class by inverse frequency to adjust for class imbalance. Hyperparameters tuned include: number of trees and number of features when looking for best split. Defaults were used for all other model parameters.

Support Vector Machine

Text input was preprocessed into a unigram and bigram count matrix followed by TFIDF transform.[23] LinearSVC[60] from the Scikit-learn[59] python library was used with minimization of squared hinge loss with L2 penalty[61] and weighting each outcome class by inverse frequency to adjust for class imbalance. Crammer-Singer approach was used for the multiclass strategy.[62] The “C” regularization strength parameter was tuned as a hyperparameter. Defaults were used for all other model parameters.

fastText

Text was directly input into the fastText classification model, which internally combines word and sub-word vector representations using continuous bag-of-words[47] and softmax with negative sampling loss[48] objective function. Hyperparameters tuned include: learning rate, learning rate update rate, word vector dimension size, context window size, number of negatives sampled, number of epochs. Defaults were used for all other model parameters.

BioClinicalBERT

Text was tokenized using WordPiece tokenizer[32–34] and then used as input to a pre-trained BioClinicalBERT model[35] with the addition of ASA-PS and Emergency modifier prediction heads, each consisting of a linear and softmax layer, for our specific ASA-PS prediction task (eFigure 3). These prediction heads were jointly optimized with AdamW optimizer[63] during training using a weighted average of the cross-entropy loss from each prediction head; the weight of ASA-PS was held constant at 1.0 and the weight of the emergency modifier head was tuned as a hyperparameter. Cross-entropy loss is weighed by inverse class frequency to adjust for class imbalance. Both tokenizer and model are based on the Hugging Face[64] python implementation with GPU acceleration enabled by PyTorch[65] and PyTorch Lightning[66]. Tokenizer and model sequence length were set to the maximum of 512 tokens for the pretrained model. Longer input text sequences were truncated to this length. Hyperparameters tuned include: emergency head weight, batch size, learning rate, weight decay, dropout, gradient clipping, and number of epochs. ASHA[67] with a reduction factor of 3 was used to tune up to 4 instances of the same model with different hyperparameters in parallel.

Tables & Figures

Table 1

			Train	Validation	Test
Patient Characteristics	**Patient Count, no. (% across dataset splits)**		26994 (70.0%)	3858 (10.0%)	7714 (20.0%)
	Number of Surgeries per Patient, no. (% within dataset split)	**1**	19107 (70.78%)	2741 (71.05%)	5475 (70.97%)
		2	4528 (16.77%)	608 (15.76%)	1330 (17.24%)
		3	1635 (6.06%)	249 (6.45%)	425 (5.51%)
		4	715 (2.65%)	124 (3.21%)	224 (2.9%)
		>=5	1009 (3.74%)	136 (3.53%)	260 (3.37%)
	Age, mean (SD)		50.59 (18.16)	51.51 (18.09)	50.66 (18.0)
	Gender, no. (% within dataset split)	**Female**	18419 (70.62%)	2534 (9.72%)	5130 (19.67%)
		Male	24720 (69.79%)	3646 (10.29%)	7053 (19.91%)
		Unknown	0 (0.0%)	0 (0.0%)	1 (100.0%)
Procedural Case Characteristics	**Case Count, no. (% across dataset splits)**		43139 (70.14%)	6180 (10.05%)	12184 (19.81%)
	Anesthesia Type, no. (% within dataset split)	**General**	34901 (81.07%)	4961 (80.51%)	9927 (81.64%)
		MAC	7063 (16.41%)	1005 (16.31%)	1905 (15.67%)
		Regional	1089 (2.53%)	196 (3.18%)	327 (2.69%)
	ASA Physical Status Classification Score, no. (% within dataset split)	**I**	3734 (8.66%)	555 (8.98%)	1127 (9.25%)
		II	13631 (31.6%)	1875 (30.34%)	3806 (31.24%)
		III	18626 (43.18%)	2649 (42.86%)	5327 (43.72%)
		IV-V	7148 (16.57%)	1101 (17.82%)	1924 (15.79%)
	Time Between Pre-Anesthesia Note and Surgery, median days HH:MM:SS (IQR)		0 days 17:11:48 (0 days 00:17:00, 4 days 06:04:05)	0 days 17:28:55 (0 days 00:18:00, 4 days 05:04:10)	0 days 17:29:55 (0 days 00:17:05, 4 days 01:52:53)
Note Characteristics	**Notes Count, no. (% across dataset splits)**		32444 (70.11%)	4649 (10.05%)	9182 (19.84%)
	Text Word-Level Length, median (IQR)	**Full Note**	727 (514, 999)	723 (514, 1010)	722 (511, 997)
		Procedure	5 (4, 8)	5 (4, 8)	5 (4, 8)
		Diagnosis	3 (2, 5)	3 (2, 5)	3 (2, 5)
		HPI	86 (35, 162)	87 (35, 161)	88 (35, 163)
		PMSH	28 (18, 42)	28 (19, 44)	28 (18, 42)
		ROS	87 (53, 154)	87 (54, 155)	87 (54, 153)
		Medications	145 (59, 264)	143 (59, 264)	146 (57, 262)

Table 1: Baseline patient, procedure, and note characteristics for Train, Validation, Test datasets.

Table 2

A. Macro-average AUROC								
	Baseline	Diagnosis	Procedure	HPI	PMSH	ROS	Meds	Note	Note512
Random Classifier	0.500	---	---	---	---	---	---	---	---
Age Classifier	0.677	---	---	---	---	---	---	---	---
Random Forest	---	0.741	0.751	0.788	0.695	0.778	0.781	0.820	0.802
Support Vector Machine	---	0.714	0.717	0.789	0.697	0.787	0.768	0.850	0.829
fastText	---	0.757	0.758	0.791	0.720	0.793	0.789	0.865	0.844
BioClinicalBERT	---	0.767	0.755	0.814	0.737	0.806	0.784	0.843	0.845

B. Class-specific AUROC										
		Baseline	Diagnosis	Procedure	HPI	PMSH	ROS	Meds	Note	Note512
Random Classifier	I	0.500	---	---	---	---	---	---	---	---
	II	0.500	---	---	---	---	---	---	---	---
	III	0.500	---	---	---	---	---	---	---	---
	IV-V	0.500	---	---	---	---	---	---	---	---
Age Classifier	I	0.842	---	---	---	---	---	---	---	---
	II	0.600	---	---	---	---	---	---	---	---
	III	0.656	---	---	---	---	---	---	---	---
	IV-V	0.611	---	---	---	---	---	---	---	---
Random Forest	I	---	0.790	0.810	0.864	0.810	0.869	0.861	0.898	0.886
	II	---	0.708	0.713	0.744	0.636	0.729	0.738	0.783	0.759
	III	---	0.660	0.674	0.708	0.644	0.708	0.718	0.747	0.719
	IV-V	---	0.804	0.806	0.835	0.691	0.803	0.807	0.854	0.844
Support Vector Machine	I	---	0.776	0.793	0.874	0.827	0.904	0.869	0.938	0.924
	II	---	0.653	0.633	0.738	0.592	0.691	0.680	0.806	0.775
	III	---	0.639	0.650	0.709	0.655	0.728	0.702	0.775	0.750
	IV-V	---	0.789	0.794	0.836	0.714	0.826	0.821	0.881	0.865
fastText	I	---	0.815	0.820	0.870	0.833	0.889	0.863	0.943	0.930
	II	---	0.724	0.718	0.755	0.675	0.771	0.755	0.833	0.809
	III	---	0.684	0.685	0.720	0.668	0.729	0.724	0.798	0.771
	IV-V	---	0.805	0.811	0.819	0.702	0.782	0.815	0.884	0.867
BioClinicalBERT	I	---	0.838	0.816	0.901	0.851	0.902	0.861	0.917	0.922
	II	---	0.711	0.707	0.768	0.674	0.748	0.737	0.806	0.804
	III	---	0.688	0.681	0.741	0.682	0.752	0.719	0.776	0.779
	IV-V	---	0.830	0.818	0.848	0.741	0.823	0.818	0.874	0.874

Table 2: (A) Macro-average AUROC and (B) class-specific AUROC for each model architecture and task on the held-out test set compared to baseline models. Random Classifier serves as a negative control baseline. Age classifier serves as a simple clinical baseline since ASA-PS typically increases as a patient ages and has increased medical comorbidities.

Figure 1

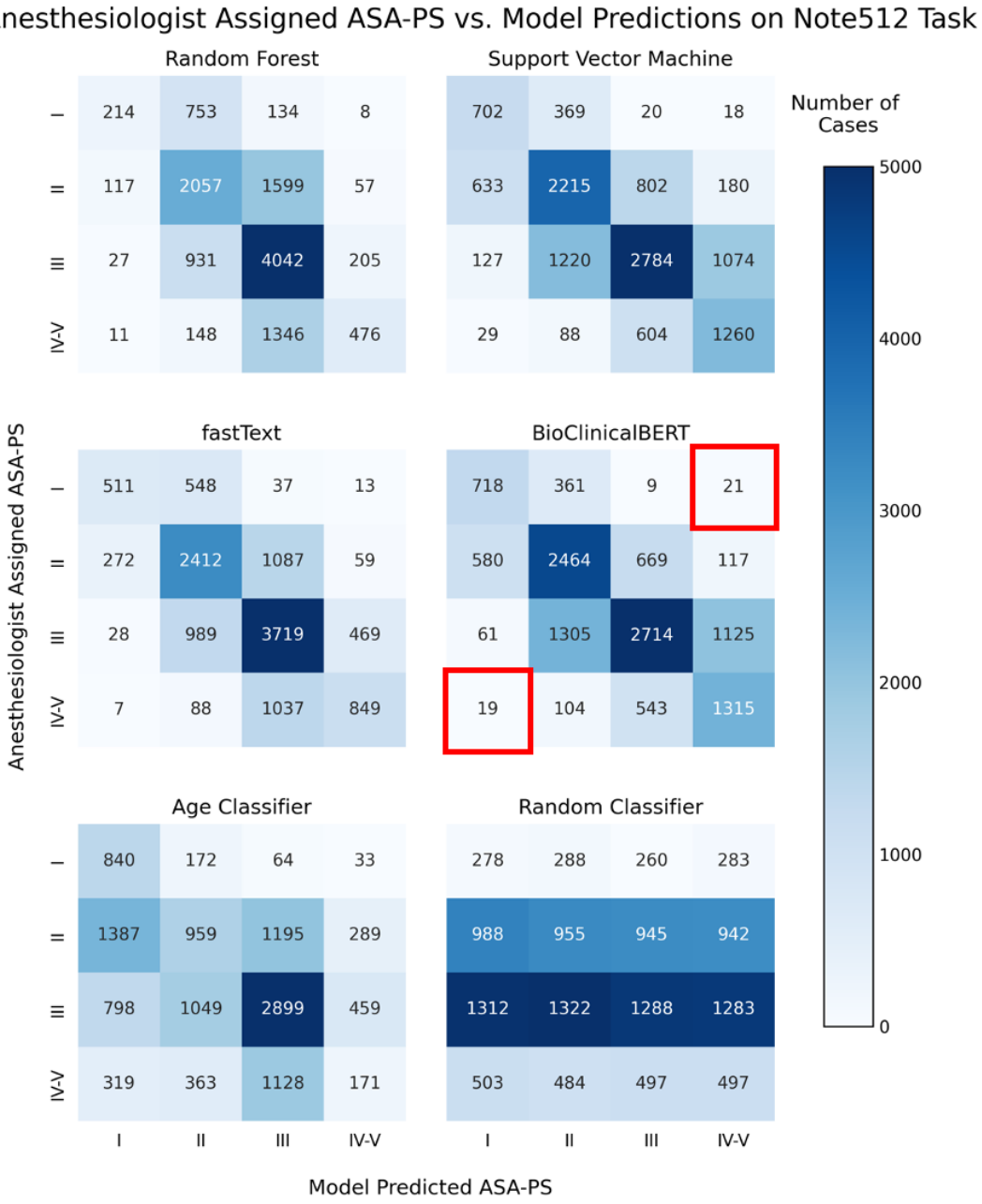


Figure 1: 4-by-4 contingency tables for each model architecture on the Note512 task. The vertical axis corresponds to modified ASA-PS recorded in the anesthetic record by the anesthesiologist. The horizontal axis corresponds to the model predicted modified ASA-PS. Numbers in the table represent case count from the test set and show how these cases are distributed based on model prediction and actual ASA-PS recorded in the anesthetic record. Cells outlined in red in the BioClinicalBERT contingency table correspond to our definition of catastrophic errors. The 21 cases where anesthesiologist assigned ASA I and BioClinicalBERT model predicted ASA IV-V comprise 1.7% of cases. The 19 cases where anesthesiologist assigned ASA IV-V and BioClinicalBERT model predicted ASA I comprise 1.6% of cases.

Figure 2

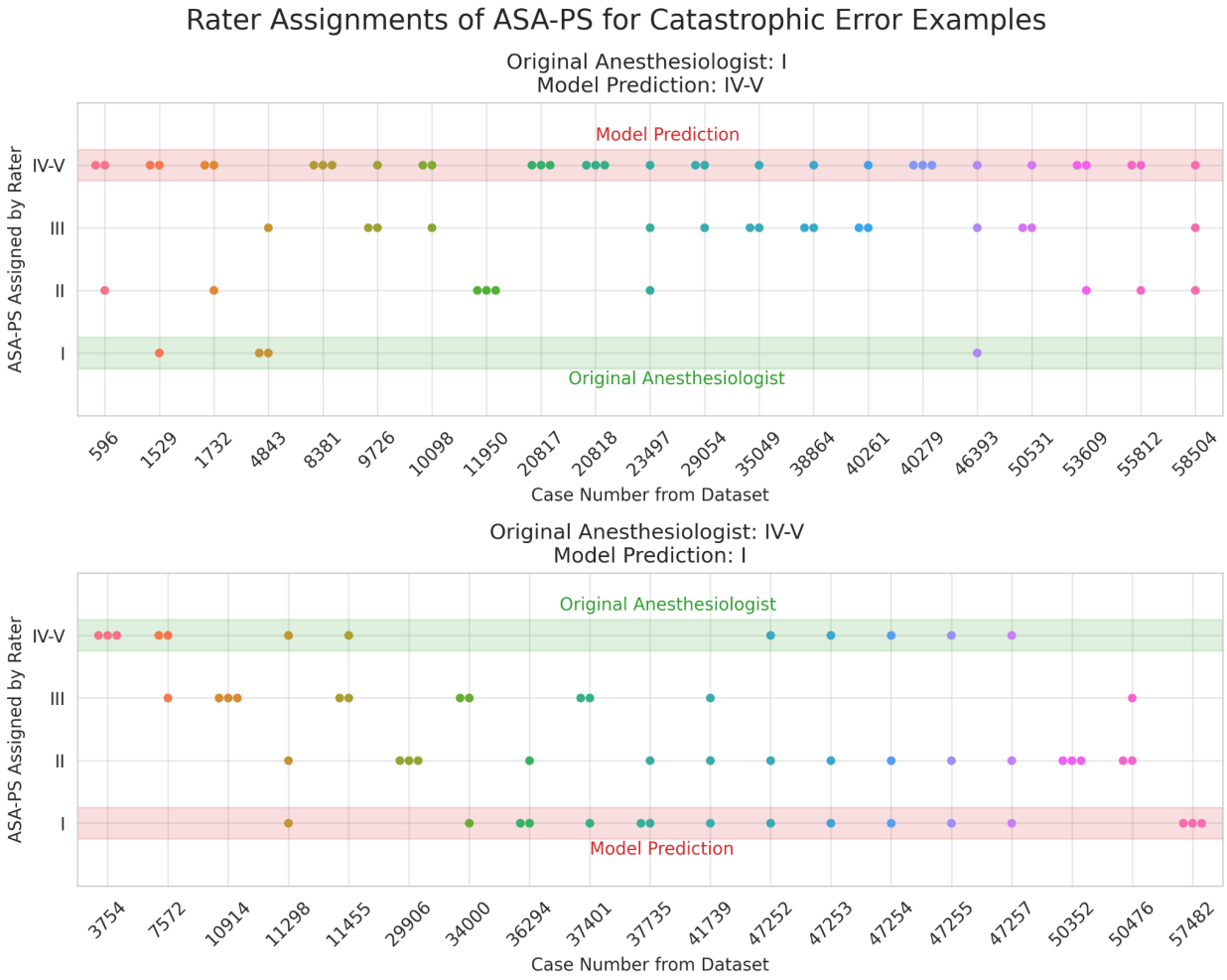


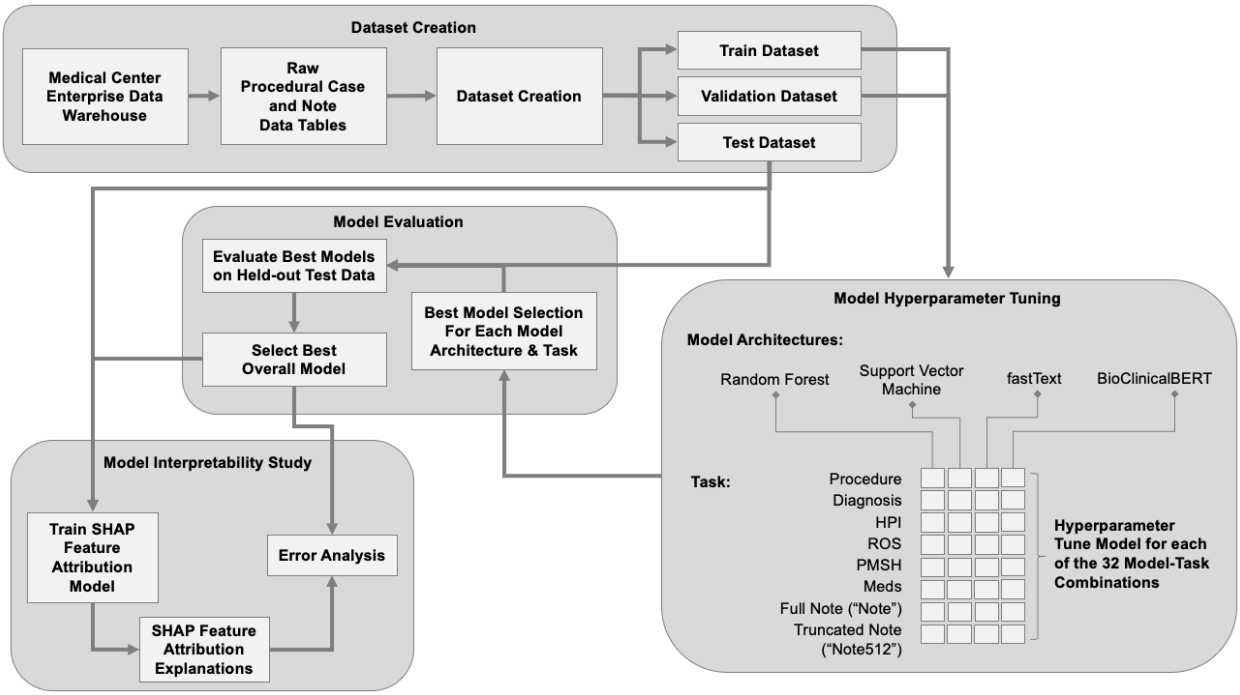
Figure 2: Rater assignments of ASA-PS for catastrophic error examples from the BioClinicalBERT model on Note512 task. Top plot shows scenario where model prediction is ASA IV-V, but original anesthesiologist assigned case ASA I. Bottom plot shows scenario where model prediction is ASA I, but original anesthesiologist assigned case ASA IV-V. Three anesthesiologist raters were asked to read the input text from the Note512 task and assign an ASA-PS for each of the catastrophic error examples. For each case, a dot marks a rater's ASA-PS assignment. The model's prediction and original anesthesiologist ASA-PS is shown as a highlighted region overlaid on the plots. Shapley feature attribution visualizations are shown for cases #57482 (Figure 3, eFigure 6), #41739 (eFigure 7), #11950 (eFigure 8), #29054 (eFigure 9).

Figure 3

Figure 3: Attribution of input text features to predicting modified ASA-PS for the BioClinicalBERT model on Note512 task. Shapley values for each text token is shown to compare feature attributions to ASA I (top) and feature attributions to ASA IV-V (bottom). Red tokens positively support predicting the target ASA-PS whereas blue tokens do not support predicting the target ASA-PS. The magnitude and direction of support is overlaid on a force plot above the text. The baseline probability of predicting each class in the test set is shown as the "base value" on the force plot. The base value + sum of Shapley values from each token corresponds to the probability of predicting the ASA-PS and is shown as the bolded number. For simplicity, feature attributions to ASA II and III are omitted in this figure, but a full-visualization with all outcome ASA-PS for this text snippet is available in eFigure 6. Text examples are de-identified by replacing ages, dates, names, locations, and entities with pseudonyms to achieve data obfuscation while preserving structural similarity to the original passage.

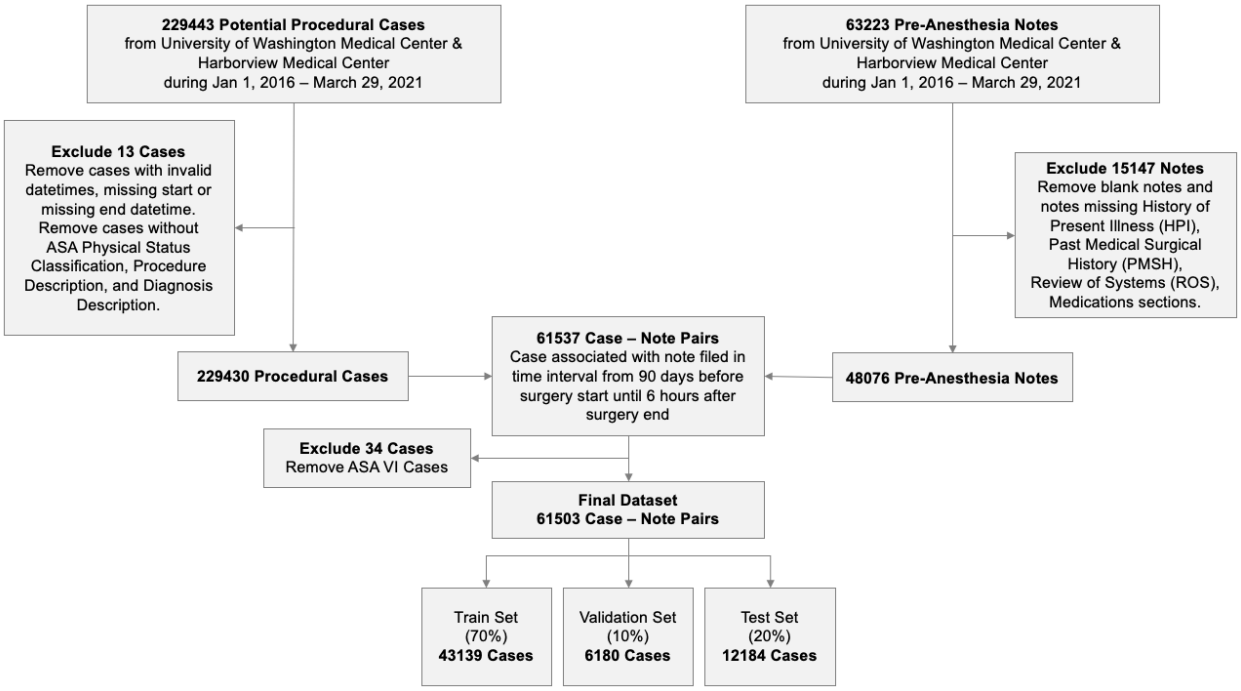
Supplemental Figures

eFigure 1



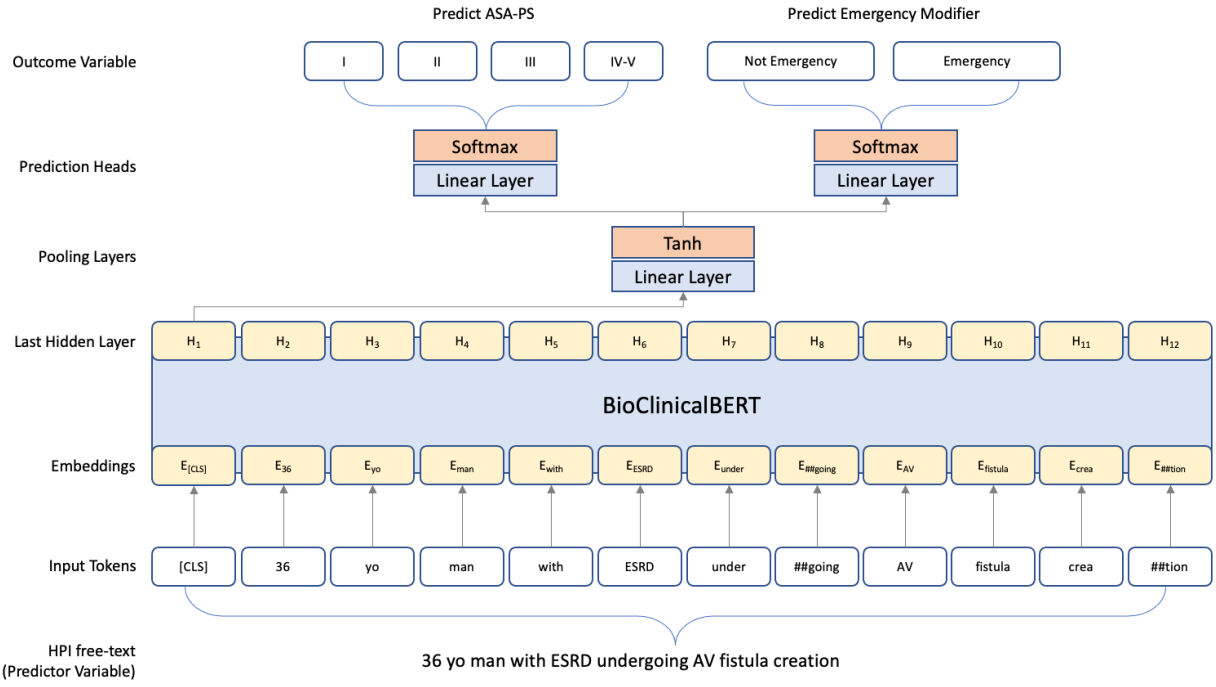
eFigure 1: Flowchart of study design: dataset creation, model development, evaluation, and interpretation.

eFigure 2



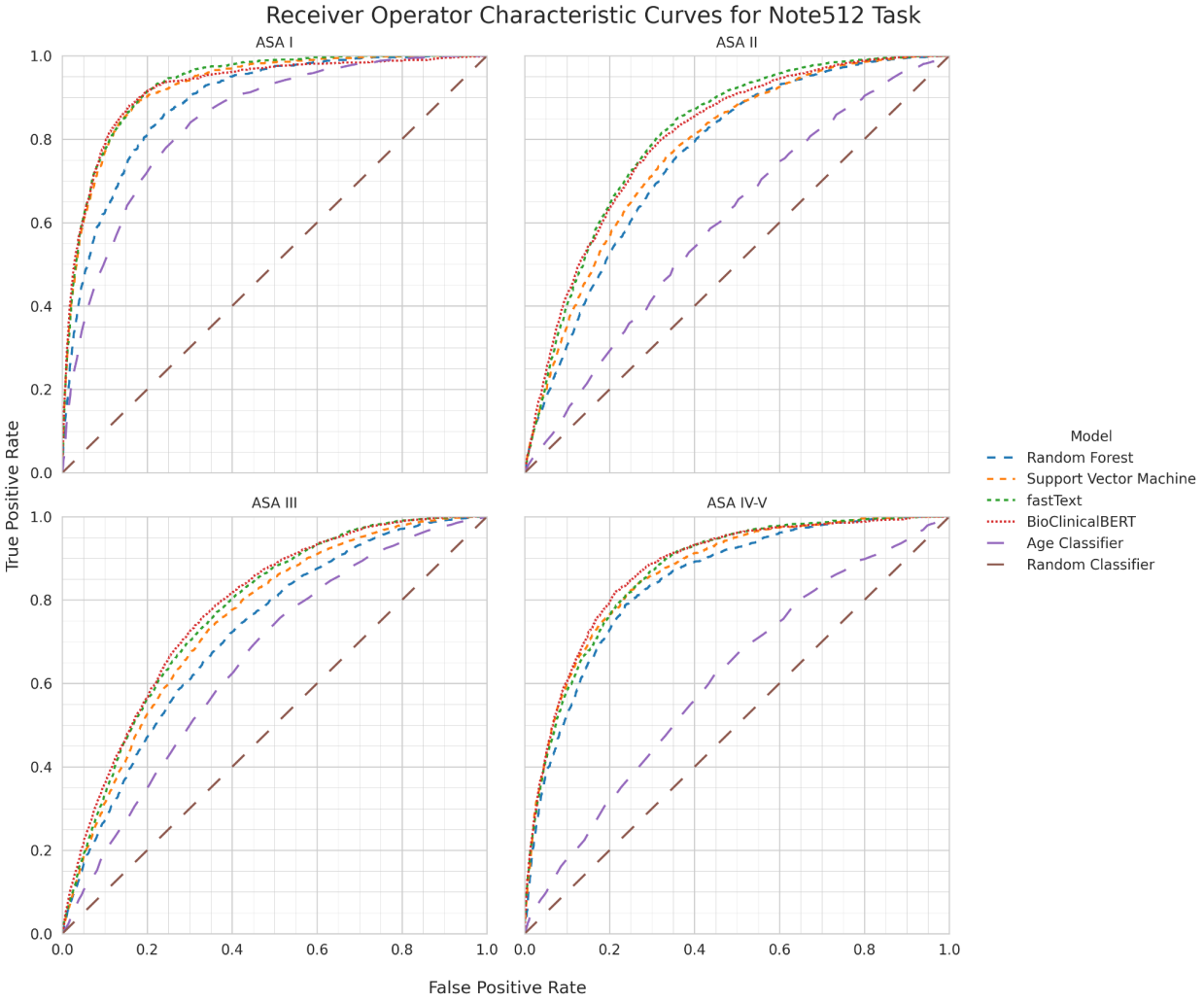
eFigure 2: CONSORT Flow Diagram for Dataset Creation. If a patient has multiple procedural cases and pre-anesthesia notes, all of a patient's cases and notes are allocated to the same data split.

eFigure 3



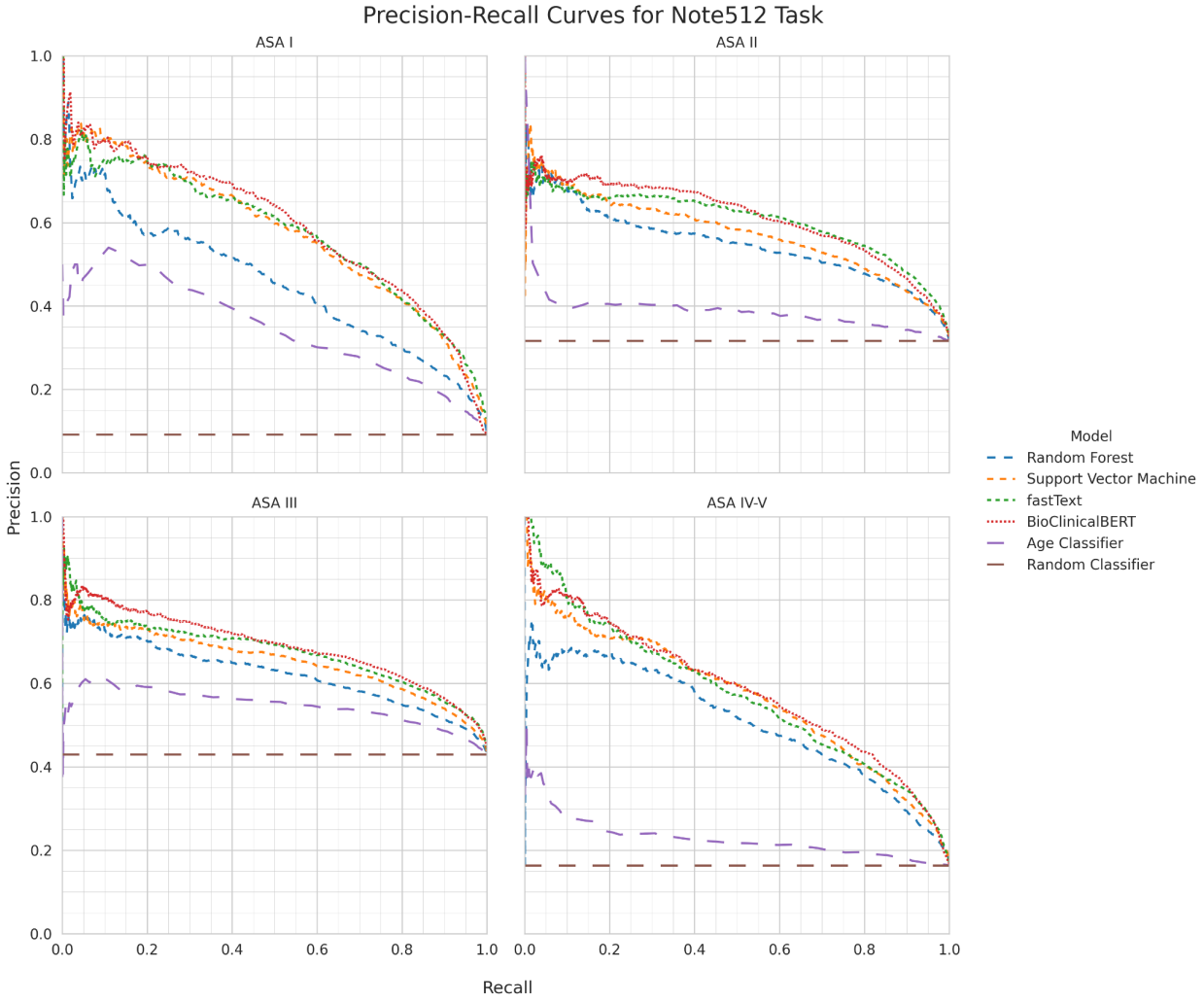
eFigure 3: BioClinicalBERT Model Architecture with additional prediction heads for fine-tuning and prediction of modified ASA-PS

eFigure4



eFigure 4: ROC performance of each model architecture on the Note512 task compared to baseline models. Each plot depicts model performance for predicting a specific ASA-PS.

eFigure 5

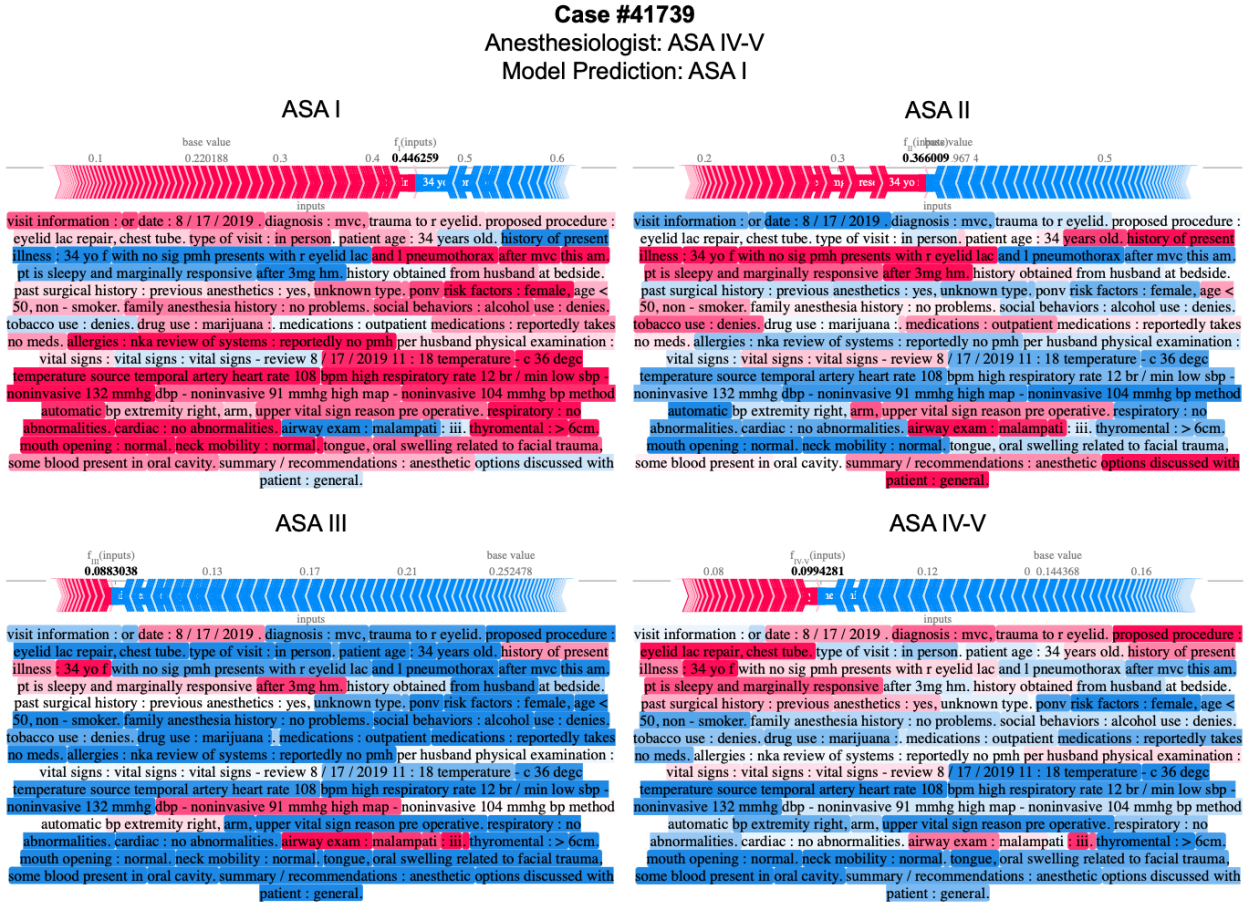


eFigure 5: Precision-recall curve performance of each model architecture on the Note512 task compared to baseline models. Each plot depicts model performance for predicting a specific ASA-PS.

eFigure 6

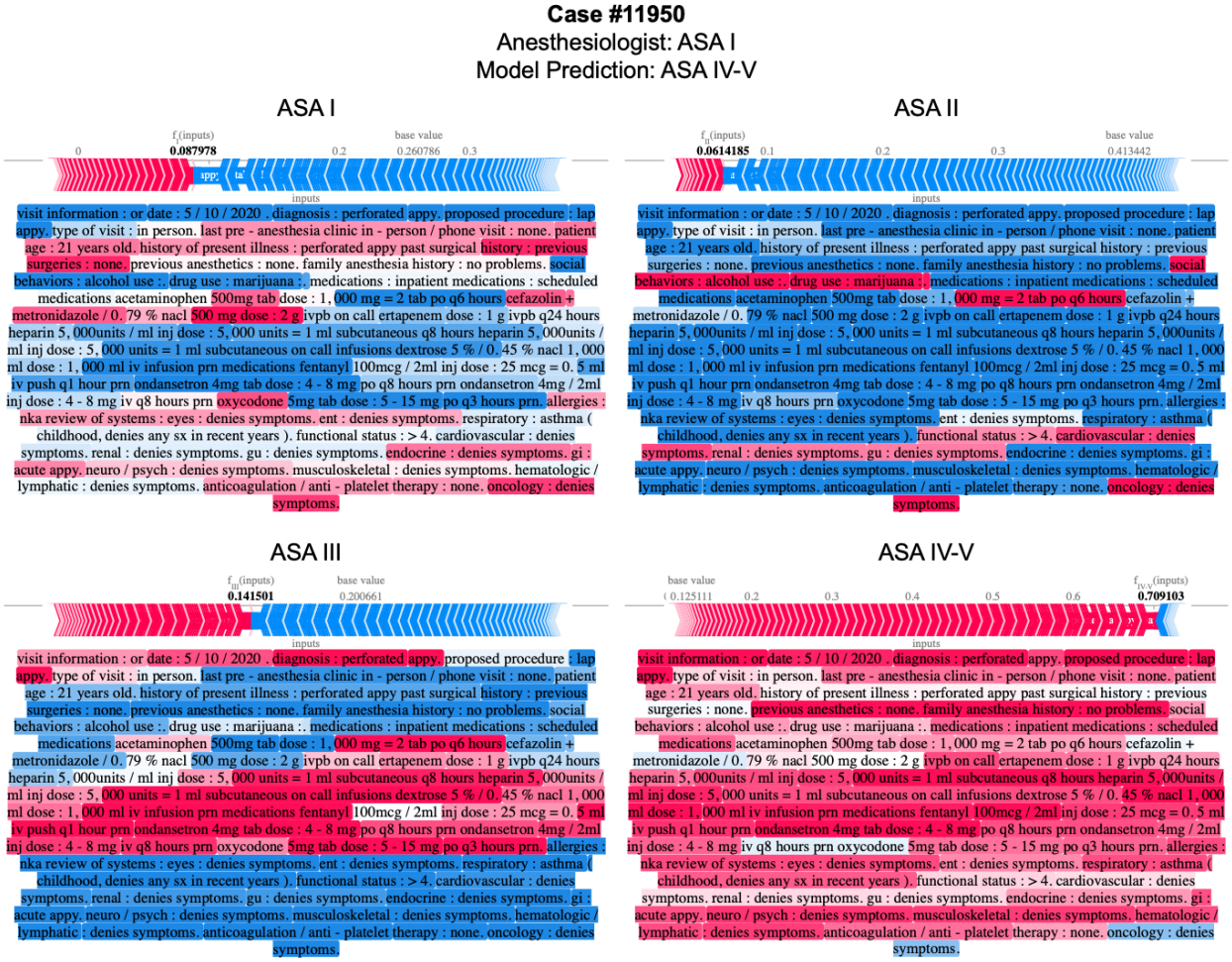
eFigure 6: Attribution of input text features to predicting modified ASA-PS for the BioClinicalBERT model on Note512 task. Model prediction is ASA I, Anesthesiologist assigned case ASA IV-V. Notable findings include the model focusing on pertinent negatives on trauma exam and imaging findings and a normal hematocrit of 33 all of which support predicting a ASA-PS I. The same pertinent negatives as well as a Glasgow Coma Scale (GCS) of 15 are negatively Shapley values for ASA-PS IV-V, which reduce the probability of predicting ASA IV-V. Despite the anesthesiologist's assignment of ASA IV-V, the text description does not suggest the patient has severe systemic disease with constant threat to life (ASA IV) or is moribund and requires the operation to survive (ASA V). Text examples are de-identified by replacing ages, dates, names, locations, and entities with pseudonyms to achieve data obfuscation while preserving structural similarity to the original passage.

eFigure 7



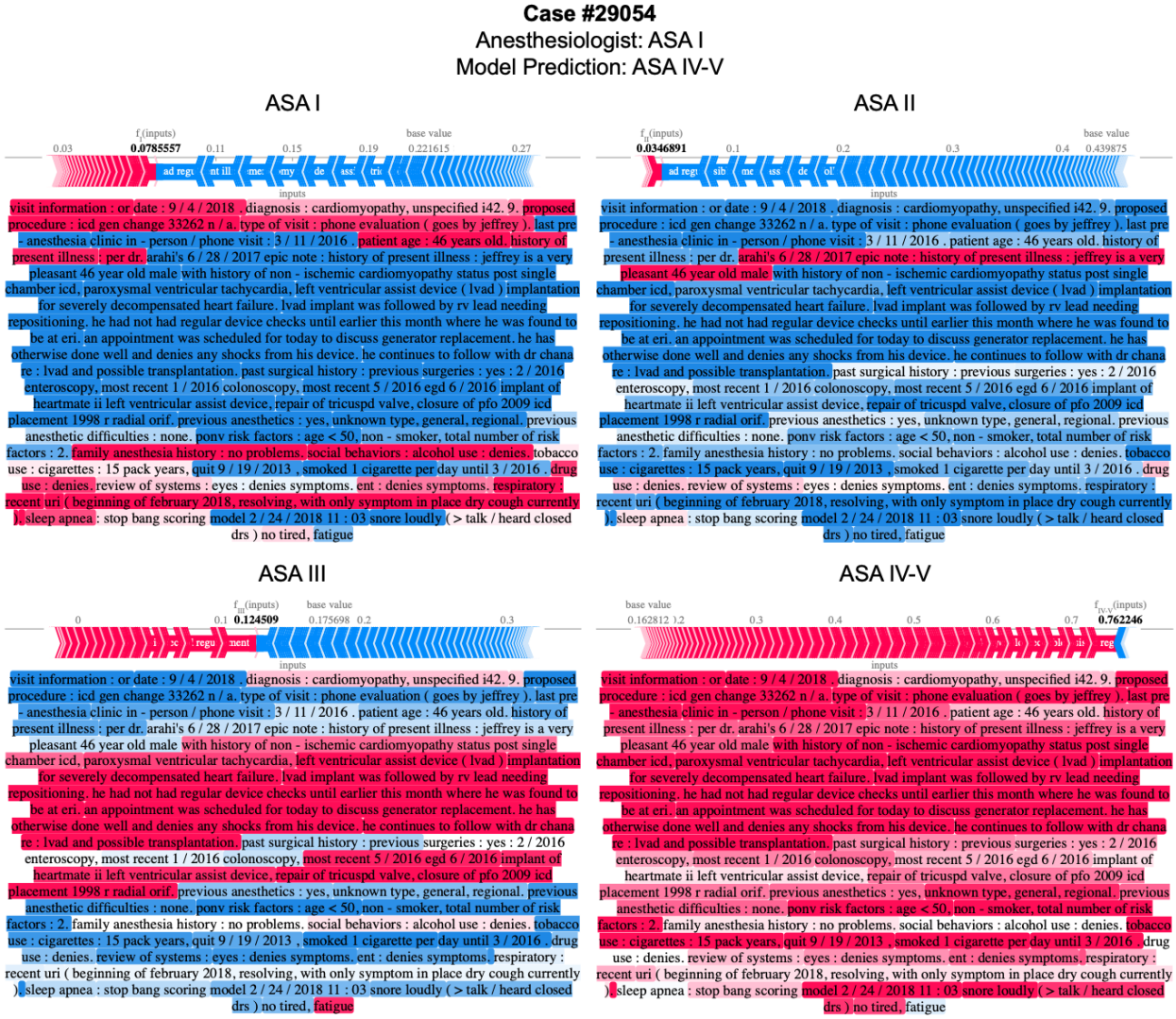
eFigure 7: Attribution of input text features to predicting modified ASA-PS for the BioClinicalBERT model on Note512 task. Model prediction is ASA I, Anesthesiologist assigned case ASA IV-V. Notable findings include the model associating chest tube with ASA IV-V. The model has trouble with consistently attributing the multiple mentions of eyelid laceration with a specific ASA-PS. The model may be inappropriately assigning mention of left pneumothorax to ASAI. This example depicts a challenge for the model in which a relatively minor injury (eyelid laceration) is simultaneously present with a potentially severe injury (pneumothorax), though the severity of the pneumothorax is not mentioned and thus the text predominantly supports ASA I (healthy) or ASA II (mild systemic disease). This kind of mixed illness/injury example coupled with a narrative that does not clearly describe disease severity may be a struggle for the model. Text examples are de-identified by replacing ages, dates, names, locations, and entities with pseudonyms to achieve data obfuscation while preserving structural similarity to the original passage.

eFigure 8



eFigure 8: Attribution of input text features to predicting modified ASA-PS for the BioClinicalBERT model on Note512 task. Model prediction is ASA IV-V, Anesthesiologist assigned case ASA I. Notable findings include: young age associated with ASA I and ASA IV-V, but negatively associated with ASA II and III; diagnosis of perforated appendix and procedure of laparoscopic appendectomy negatively associated with ASA I and positively associated with higher ASA-PS; model identifying broad-spectrum antibiotics such as ertapenem to be associated with ASA IV-V, but narrower-spectrum antibiotics such as metronidazole, cefazolin to be heavily associated with ASA I; inpatient medications such as subcutaneous heparin and ondansetron negatively associated with lower ASA-PS and positively associated with higher ASA-PS. Text examples are de-identified by replacing ages, dates, names, locations, and entities with pseudonyms to achieve data obfuscation while preserving structural similarity to the original passage.

eFigure 9



eFigure 9: Attribution of input text features to predicting modified ASA-PS for the BioClinicalBERT model on Note512 task. Model prediction is ASA IV-V, Anesthesiologist assigned case ASA I. Notable findings include medical conditions and interventions associated with higher ASA-PS such as cardiomyopathy, internal cardiac defibrillator (ICD) generator change, paroxysmal ventricular tachycardia, left ventricular assist device (LVAD), heart failure, possible transplantation, tricuspid valve repair, and patent foramen ovale (PFO) closure; history of chronic cigarette smoking and snoring associated with ASA IV-V. The text description is at least ASA III (severe systemic illness), and can be argued to be ASA IV (severe systemic disease with constant threat to life) if heart failure is progressively worsening. In this example the model appears to make a more appropriate ASA-PS prediction than the anesthesiologist. Text examples are de-identified by replacing ages, dates, names, locations, and entities with pseudonyms to achieve data obfuscation while preserving structural similarity to the original passage.

eTable 1

A. Macro-average AUPRC									
	Baseline	Diagnosis	Procedure	HPI	PMSH	ROS	Meds	Note	Note512
Random Classifier	0.250	---	---	---	---	---	---	---	---
Age Classifier	0.375	---	---	---	---	---	---	---	---
Random Forest	---	0.457	0.462	0.510	0.392	0.484	0.489	0.567	0.534
Support Vector Machine	---	0.443	0.451	0.525	0.413	0.514	0.490	0.627	0.593
fastText	---	0.478	0.473	0.518	0.421	0.512	0.495	0.642	0.607
BioClinicalBERT	---	0.486	0.473	0.570	0.446	0.536	0.499	0.616	0.619

B. Class-specific AUPRC										
		Baseline	Diagnosis	Procedure	HPI	PMSH	ROS	Meds	Note	Note512
Random Classifier	I	0.091	---	---	---	---	---	---	---	---
	II	0.316	---	---	---	---	---	---	---	---
	III	0.429	---	---	---	---	---	---	---	---
	IV-V	0.163	---	---	---	---	---	---	---	---
Age Classifier	I	0.343	---	---	---	---	---	---	---	---
	II	0.383	---	---	---	---	---	---	---	---
	III	0.546	---	---	---	---	---	---	---	---
	IV-V	0.227	---	---	---	---	---	---	---	---
Random Forest	I	---	0.285	0.285	0.394	0.295	0.374	0.327	0.488	0.455
	II	---	0.490	0.487	0.518	0.425	0.515	0.498	0.580	0.550
	III	---	0.565	0.576	0.614	0.551	0.610	0.621	0.650	0.625
	IV-V	---	0.488	0.500	0.514	0.299	0.437	0.510	0.550	0.508
Support Vector Machine	I	---	0.272	0.305	0.436	0.323	0.433	0.345	0.606	0.575
	II	---	0.460	0.441	0.519	0.392	0.493	0.477	0.614	0.574
	III	---	0.568	0.567	0.618	0.570	0.639	0.618	0.684	0.655
	IV-V	---	0.473	0.492	0.527	0.367	0.491	0.519	0.605	0.568
fastText	I	---	0.317	0.308	0.428	0.316	0.429	0.340	0.617	0.575
	II	---	0.507	0.491	0.531	0.453	0.559	0.517	0.645	0.605
	III	---	0.590	0.583	0.620	0.568	0.617	0.622	0.705	0.675
	IV-V	---	0.495	0.510	0.491	0.349	0.444	0.502	0.601	0.575
BioClinicalBERT	I	---	0.330	0.301	0.529	0.354	0.445	0.337	0.582	0.591
	II	---	0.499	0.487	0.562	0.454	0.553	0.521	0.616	0.612
	III	---	0.599	0.585	0.641	0.588	0.655	0.628	0.679	0.690
	IV-V	---	0.517	0.519	0.546	0.388	0.492	0.509	0.588	0.585

eTable 1: (A) Macro-average AUPRC and (B) class-specific AUPRC for each model architecture and task on the held-out test set compared to baseline models. Random Classifier serves as a

negative control baseline. Age classifier serves as a simple clinical baseline since ASA-PS typically increases as a patient ages and has increased medical comorbidities.

eTable 2

A. Matthew's Correlation Coefficient (MCC)									
	Baseline	**Diagnosis**	**Procedure**	**HPI**	**PMSH**	**ROS**	**Meds**	**Note**	**Note512**
Random Classifier	0.000	---	---	---	---	---	---	---	---
Age Classifier	0.161	---	---	---	---	---	---	---	---
Random Forest	---	0.264	0.265	0.280	0.197	0.293	0.314	0.370	0.317
Support Vector Machine	---	0.252	0.247	0.332	0.194	0.326	0.299	0.431	0.398
fastText	---	0.280	0.278	0.336	0.230	0.360	0.324	0.461	0.425
BioClinicalBERT	---	0.280	0.267	0.369	0.226	0.364	0.321	0.439	0.430

B. AUCμ									
	Baseline	**Diagnosis**	**Procedure**	**HPI**	**PMSH**	**ROS**	**Meds**	**Note**	**Note512**
Random Classifier	0.500	---	---	---	---	---	---	---	---
Age Classifier	0.726	---	---	---	---	---	---	---	---
Random Forest	---	0.772	0.781	0.821	0.727	0.809	0.806	0.852	0.836
Support Vector Machine	---	0.767	0.778	0.830	0.755	0.849	0.827	0.889	0.872
fastText	---	0.777	0.776	0.812	0.745	0.825	0.809	0.882	0.865
BioClinicalBERT	---	0.830	0.816	0.871	0.798	0.865	0.838	0.884	0.891

eTable 2: (A) Matthew's correlation coefficient (MCC) and (B) AUCμ for each model architecture and task on the held-out test set compared to baseline models. MCC is a categorical analog of Pearson's correlation coefficient. AUCμ is a multiclass generalization of AUROC and U statistic and is more theoretically grounded than macro-average AUROC, but less commonly reported.

eTable 3

A. Macro-average F1									
	Baseline	Diagnosis	Procedure	HPI	PMSH	ROS	Meds	Note	Note512
Random Classifier	0.231	---	---	---	---	---	---	---	---
Age Classifier	0.337	---	---	---	---	---	---	---	---
Random Forest	---	0.456	0.445	0.394	0.391	0.448	0.474	0.509	0.457
Support Vector Machine	---	0.436	0.420	0.509	0.382	0.483	0.463	0.588	0.566
fastText	---	0.439	0.450	0.491	0.416	0.510	0.476	0.606	0.580
BioClinicalBERT	---	0.441	0.454	0.545	0.400	0.530	0.496	0.600	0.590

B. Class-specific F1										
		Baseline	Diagnosis	Procedure	HPI	PMSH	ROS	Meds	Note	Note512
Random Classifier	**I**	0.133	---	---	---	---	---	---	---	---
	II	0.278	---	---	---	---	---	---	---	---
	III	0.314	---	---	---	---	---	---	---	---
	IV-V	0.199	---	---	---	---	---	---	---	---
Age Classifier	**I**	0.377	---	---	---	---	---	---	---	---
	II	0.301	---	---	---	---	---	---	---	---
	III	0.553	---	---	---	---	---	---	---	---
	IV-V	0.117	---	---	---	---	---	---	---	---
Random Forest	**I**	---	0.293	0.242	0.160	0.335	0.341	0.303	0.353	0.290
	II	---	0.489	0.507	0.530	0.390	0.536	0.537	0.601	0.533
	III	---	0.560	0.581	0.644	0.593	0.629	0.628	0.647	0.656
	IV-V	---	0.480	0.451	0.244	0.247	0.285	0.428	0.437	0.349
Support Vector Machine	**I**	---	0.337	0.354	0.422	0.355	0.458	0.422	0.565	0.540
	II	---	0.405	0.390	0.537	0.343	0.445	0.427	0.608	0.574
	III	---	0.499	0.456	0.559	0.441	0.528	0.490	0.612	0.591
	IV-V	---	0.503	0.479	0.519	0.388	0.503	0.512	0.567	0.558
fastText	**I**	---	0.195	0.253	0.333	0.312	0.395	0.283	0.560	0.530
	II	---	0.525	0.509	0.558	0.473	0.584	0.543	0.633	0.613
	III	---	0.606	0.608	0.641	0.600	0.650	0.640	0.692	0.671
	IV-V	---	0.429	0.431	0.432	0.280	0.412	0.437	0.540	0.504
BioClinicalBERT	**I**	---	0.365	0.348	0.530	0.398	0.506	0.416	0.578	0.577
	II	---	0.422	0.475	0.561	0.384	0.532	0.507	0.607	0.611
	III	---	0.459	0.498	0.560	0.410	0.594	0.560	0.644	0.594
	IV-V	---	0.519	0.493	0.530	0.408	0.488	0.501	0.570	0.577

eTable 3: (A) Macro-average F1 and (B) class-specific F1 for each model architecture and task on the held-out test set compared to baseline models.

eTable 4

A. Macro-average Precision								
	Baseline	Diagnosis	Procedure	HPI	PMSH	ROS	Meds	Note	Note512
Random Classifier	0.250	---	---	---	---	---	---	---	---
Age Classifier	0.339	---	---	---	---	---	---	---	---
Random Forest	---	0.460	0.468	0.581	0.401	0.507	0.511	0.604	0.579
Support Vector Machine	---	0.430	0.426	0.500	0.389	0.483	0.462	0.573	0.550
fastText	---	0.524	0.509	0.558	0.478	0.557	0.516	0.631	0.616
BioClinicalBERT	---	0.451	0.444	0.531	0.412	0.516	0.484	0.591	0.576

B. Class-specific Precision									
		Baseline	Diagnosis	Procedure	HPI	PMSH	ROS	Meds	Note	Note512
Random Classifier	I	0.090	---	---	---	---	---	---	---	---
	II	0.313	---	---	---	---	---	---	---	---
	III	0.431	---	---	---	---	---	---	---	---
	IV-V	0.165	---	---	---	---	---	---	---	---
Age Classifier	I	0.251	---	---	---	---	---	---	---	---
	II	0.377	---	---	---	---	---	---	---	---
	III	0.548	---	---	---	---	---	---	---	---
	IV-V	0.180	---	---	---	---	---	---	---	---
Random Forest	I	---	0.318	0.300	0.550	0.299	0.414	0.355	0.643	0.580
	II	---	0.493	0.487	0.512	0.457	0.498	0.509	0.534	0.529
	III	---	0.550	0.553	0.548	0.530	0.581	0.592	0.619	0.568
	IV-V	---	0.478	0.532	0.713	0.319	0.533	0.586	0.619	0.638
Support Vector Machine	I	---	0.249	0.239	0.394	0.234	0.315	0.294	0.508	0.471
	II	---	0.485	0.475	0.517	0.401	0.545	0.511	0.602	0.569
	III	---	0.570	0.597	0.625	0.578	0.654	0.634	0.683	0.661
	IV-V	---	0.418	0.394	0.465	0.343	0.416	0.411	0.499	0.498
fastText	I	---	0.474	0.414	0.543	0.427	0.511	0.379	0.625	0.625
	II	---	0.489	0.492	0.535	0.452	0.550	0.526	0.633	0.597
	III	---	0.557	0.553	0.587	0.540	0.610	0.591	0.656	0.632
	IV-V	---	0.576	0.576	0.568	0.492	0.556	0.569	0.609	0.611
BioClinicalBERT	I	---	0.243	0.289	0.469	0.274	0.382	0.328	0.547	0.521
	II	---	0.488	0.472	0.546	0.462	0.556	0.520	0.613	0.582
	III	---	0.642	0.595	0.646	0.601	0.655	0.631	0.665	0.690
	IV-V	---	0.432	0.419	0.462	0.311	0.472	0.456	0.537	0.511

eTable 4: (A) Macro-average precision and (B) class-specific precision for each model architecture and task on the held-out test set compared to baseline models.

eTable 5

A. Macro-average Recall								
	Baseline	Diagnosis	Procedure	HPI	PMSH	ROS	Meds	Note	Note512
Random Classifier	0.250	---	---	---	---	---	---	---	---
Age Classifier	0.413	---	---	---	---	---	---	---	---
Random Forest	---	0.453	0.434	0.393	0.399	0.438	0.460	0.486	0.437
Support Vector Machine	---	0.486	0.498	0.526	0.459	0.572	0.548	0.615	0.596
fastText	---	0.424	0.432	0.469	0.403	0.491	0.460	0.590	0.558
BioClinicalBERT	---	0.527	0.485	0.575	0.489	0.577	0.530	0.611	0.619

B. Class-specific Recall									
		Baseline	Diagnosis	Procedure	HPI	PMSH	ROS	Meds	Note	Note512
Random Classifier	I	0.251	---	---	---	---	---	---	---	---
	II	0.249	---	---	---	---	---	---	---	---
	III	0.247	---	---	---	---	---	---	---	---
	IV-V	0.251	---	---	---	---	---	---	---	---
Age Classifier	I	0.757	---	---	---	---	---	---	---	---
	II	0.250	---	---	---	---	---	---	---	---
	III	0.557	---	---	---	---	---	---	---	---
	IV-V	0.086	---	---	---	---	---	---	---	---
Random Forest	I	---	0.271	0.203	0.094	0.380	0.290	0.264	0.243	0.193
	II	---	0.486	0.528	0.550	0.340	0.581	0.569	0.687	0.537
	III	---	0.571	0.612	0.780	0.674	0.687	0.669	0.677	0.777
	IV-V	---	0.483	0.391	0.147	0.201	0.195	0.337	0.337	0.240
Support Vector Machine	I	---	0.521	0.682	0.454	0.736	0.834	0.747	0.637	0.633
	II	---	0.348	0.331	0.557	0.299	0.375	0.367	0.613	0.578
	III	---	0.445	0.369	0.506	0.357	0.443	0.399	0.555	0.535
	IV-V	---	0.631	0.612	0.588	0.446	0.637	0.679	0.657	0.636
fastText	I	---	0.123	0.182	0.240	0.246	0.322	0.226	0.508	0.461
	II	---	0.567	0.527	0.582	0.495	0.622	0.561	0.634	0.630
	III	---	0.665	0.676	0.705	0.675	0.694	0.698	0.732	0.715
	IV-V	---	0.342	0.344	0.349	0.195	0.327	0.355	0.485	0.429
BioClinicalBERT	I	---	0.729	0.436	0.610	0.724	0.748	0.568	0.611	0.647
	II	---	0.372	0.479	0.577	0.328	0.510	0.493	0.601	0.643
	III	---	0.357	0.428	0.494	0.312	0.544	0.504	0.625	0.522
	IV-V	---	0.651	0.599	0.621	0.593	0.505	0.555	0.607	0.664

eTable 5: (A) Macro-average recall and (B) class-specific recall for each model architecture and task on the held-out test set compared to baseline models.

References

1. Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *J Am Coll Surg*. 2013;217(5):833-842.e1-e3. doi:10.1016/j.jamcollsurg.2013.07.385

2. Cohen ME, Ko CY, Bilimoria KY, et al. Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. *J Am Coll Surg*. 2013;217(2):336-346.e1. doi:10.1016/j.jamcollsurg.2013.02.027

3. Lee TH, Marcantonio ER, Mangione CM, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation*. 1999;100(10):1043-1049. doi:10.1161/01.cir.100.10.1043

4. Goldman L, Caldera DL, Nussbaum SR, et al. Multifactorial index of cardiac risk in noncardiac surgical procedures. *N Engl J Med*. 1977;297(16):845-850. doi:10.1056/NEJM197710202971601

5. Gupta PK, Gupta H, Sundaram A, et al. Development and validation of a risk calculator for prediction of cardiac risk after surgery. *Circulation*. 2011;124(4):381-387. doi:10.1161/CIRCULATIONAHA.110.015701

6. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv [csCL]*. Published online June 16, 2016. http://arxiv.org/abs/1606.05250

7. Zellers R, Bisk Y, Schwartz R, Choi Y. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. *arXiv [csCL]*. Published online August 16, 2018. http://arxiv.org/abs/1808.05326

8. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv:180407461 [cs]*. Published online February 22, 2019. Accessed January 27, 2020. http://arxiv.org/abs/1804.07461

9. Wang A, Pruksachatkun Y, Nangia N, et al. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *arXiv:190500537 [cs]*. Published online July 12, 2019. Accessed January 27, 2020. http://arxiv.org/abs/1905.00537

10. Zhang Z, Liu J, Razavian N. BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2020:24-34. doi:10.18653/v1/2020.clinicalnlp-1.3

11. Liu L, Perez-Concha O, Nguyen A, Bennett V, Jorm L. Automated ICD Coding using Extreme Multi-label Long Text Transformer-based Models. *arXiv [csCL]*. Published online December 12, 2022. http://arxiv.org/abs/2212.05857

12. Mayhew D, Mendonca V, Murthy BVS. A review of ASA physical status - historical perspectives and modern developments. *Anaesthesia*. 2019;74(3):373-379. doi:10.1111/anae.14569

13. Horvath B, Kloesel B, Todd MM, Cole DJ, Prielipp RC. The Evolution, Current Value, and Future of the American Society of Anesthesiologists Physical Status Classification System. *Anesthesiology*. 2021;135(5):904-919. doi:10.1097/ALN.0000000000003947

14. Wolters U, Wolf T, Stützer H, Schröder T. ASA classification and perioperative variables as predictors of postoperative outcome. *Br J Anaesth*. 1996;77(2):217-222. doi:10.1093/bja/77.2.217

15. Tiret L, Hatton F, Desmonts JM, Vourc'h G. Prediction of outcome of anaesthesia in patients over 40 years: a multifactorial risk index. *Stat Med*. 1988;7(9):947-954. doi:10.1002/sim.4780070906

16. Hackett NJ, De Oliveira GS, Jain UK, Kim JYS. ASA class is a reliable independent predictor of medical complications and mortality following surgery. *Int J Surg*. 2015;18:184-190. doi:10.1016/j.ijsu.2015.04.079

17. Tran A, Mai T, El-Haddad J, et al. Preinjury ASA score as an independent predictor of readmission after major traumatic injury. *Trauma Surgery & Acute Care Open*. Published online 2017. doi:10.1136/tsaco-2017-000128

18. Konda SR, Parola R, Perskin C, Egol KA. ASA Physical Status Classification Improves Predictive Ability of a Validated Trauma Risk Score. *Geriatr Orthop Surg Rehabil*. 2021;12:2151459321989534. doi:10.1177/2151459321989534

19. Davenport DL, Ferraris VA, Hosokawa P, Henderson WG, Khuri SF, Mentzer RM Jr. Multivariable predictors of postoperative cardiac adverse events after general and vascular surgery: results from the patient safety in surgery study. *J Am Coll Surg*. 2007;204(6):1199-1210. doi:10.1016/j.jamcollsurg.2007.02.065

20. Cuvillon P, Nouvellon E, Marret E, et al. American Society of Anesthesiologists' physical status system: a multicentre Francophone study to analyse reasons for classification disagreement. *Eur J Anaesthesiol*. 2011;28(10):742-747. doi:10.1097/EJA.0b013e328348fc9d

21. Sankar A, Johnson SR, Beattie WS, Tait G, Wijeysundera DN. Reliability of the American Society of Anesthesiologists physical status scale in clinical practice. *Br J Anaesth*. 2014;113(3):424-432. doi:10.1093/bja/aeu100

22. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

23. Buckley C, Lewit AF. Optimization of inverted vector searches. In: *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '85. Association for Computing Machinery; 1985:97-110. doi:10.1145/253495.253515

24. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. COLT '92. Association for Computing Machinery; 1992:144-152. doi:10.1145/130385.130401

25. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of Tricks for Efficient Text Classification. *arXiv:160701759 [cs]*. Published online August 9, 2016. Accessed December 13, 2019. http://arxiv.org/abs/1607.01759

26. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. *arXiv [csCL]*. Published online July 15, 2016. http://arxiv.org/abs/1607.04606

27. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med*. 2015;13:1. doi:10.1186/s12916-014-0241-z

28. Doshi-Velez F, Perlis RH. Evaluating Machine Learning Articles. *JAMA*. 2019;322(18):1777-1779. doi:10.1001/jama.2019.17304

29. Liu Y, Chen PHC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA*. 2019;322(18):1806-1816. doi:10.1001/jama.2019.16489

30. Yusuf M, Atal I, Li J, et al. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ Open*. 2020;10(3):e034568. doi:10.1136/bmjopen-2019-034568

31. Saklad M. Grading of patients for surgical procedures. *Anesthesiology*. 1941;2(3):281-284. doi:10.1097/00000542-194105000-00004

32. Wu Y, Schuster M, Chen Z, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:160908144 [cs]*. Published online October 8, 2016. Accessed January 8, 2021. http://arxiv.org/abs/1609.08144

33. Schuster M, Nakajima K. Japanese and Korean voice search. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ; 2012:5149-5152. doi:10.1109/ICASSP.2012.6289079

34. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:181004805 [cs]*. Published online May 24, 2019. Accessed December 9, 2019. http://arxiv.org/abs/1810.04805

35. Alsentzer E, Murphy JR, Boag W, et al. Publicly Available Clinical BERT Embeddings. *arXiv [csCL]*. Published online April 6, 2019. Accessed January 27, 2020. http://arxiv.org/abs/1904.03323

36. Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: A Research Platform for Distributed Model Selection and Training. *arXiv [csLG]*. Published online July 13, 2018. http://arxiv.org/abs/1807.05118

37. Wang C, Wu Q, Weimer M, Zhu E (eric). FLAML: A Fast and Lightweight AutoML Library. In: *Fourth Conference on Machine Learning and Systems (MLSys 2021)*. ; 2021. Accessed August 10, 2022. https://www.microsoft.com/en-us/research/publication/2021/03/MLSys21FLAML.pdf

38. Wang C, Wu Q, Huang S, Saied A. Economical Hyperparameter Optimization with Blended Search Strategy. In: *The Ninth International Conference on Learning Representations (ICLR 2021)*. ; 2021. Accessed January 5, 2023. https://www.microsoft.com/en-us/research/publication/economical-hyperparameter-optimization-with-blended-search-strategy/

39. Kleiman R, Page D. AUC\textmu: A Performance Metric for Multi-Class Machine Learning Models. Chaudhuri K, Salakhutdinov R, eds. 09--15 Jun 2019;97:3439-3447. http://proceedings.mlr.press/v97/kleiman19a.html

40. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. *arXiv:170507874 [cs, stat]*. Published online November 24, 2017. Accessed March 30, 2021. http://arxiv.org/abs/1705.07874

41. Lewis DD. *Representation and Learning in Information Retrieval*. University of Massachusetts Amherst; 1992. Accessed January 3, 2023. https://scholarworks.umass.edu/dissertations/AAI9219460/

42. Lewis DD. Feature Selection and Feature Extraction for Text Categorization. In: *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. ; 1992. https://aclanthology.org/H92-1041/

43. Cavnar WB, Trenkle JM. *N-Gram-Based Text Categorization*. Nevada Univ., Las Vegas, NV (United States); 1994. Accessed January 4, 2023. https://www.osti.gov/biblio/68573

44. Damashek M. Gauging Similarity with *n*-Grams: Language-Independent Categorization of Text. *Science*. 1995;267(5199):843-848. doi:10.1126/science.267.5199.843

45. Yang Y, Pedersen JO. A Comparative Study on Feature Selection in Text Categorization. *International Conference on Machine Learning*. Published online 1997. Accessed January 3, 2023. https://www.semanticscholar.org/paper/c3ebcef26c22a373b6f26a67934213eb0582804e

46. Loper E, Bird S. NLTK: The Natural Language Toolkit. *arXiv [csCL]*. Published online May 17, 2002. http://arxiv.org/abs/cs/0205028

47. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. *arXiv [csCL]*. Published online January 16, 2013. http://arxiv.org/abs/1301.3781

48. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. *arXiv:13104546 [cs, stat]*. Published online October 16, 2013. Accessed December 9, 2019. http://arxiv.org/abs/1310.4546

49. Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics; 2014:1532-1543. doi:10.3115/v1/D14-1162

50. Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics; 2016:1715-1725. doi:10.18653/v1/P16-1162

51. Luong MT, Pham H, Manning CD. Effective Approaches to Attention-based Neural Machine Translation. *arXiv:150804025 [cs]*. Published online September 20, 2015. Accessed December 9, 2019. http://arxiv.org/abs/1508.04025

52. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:14090473 [cs, stat]*. Published online May 19, 2016. Accessed December 9, 2019. http://arxiv.org/abs/1409.0473

53. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539

54. Peters ME, Neumann M, Iyyer M, et al. Deep Contextualized Word Representations. In: *Proceedings of the 2018 Conference of the North AMerican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics; 2018:2227-2237. doi:10.18653/v1/N18-1202

55. Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics; 2018:328-339. doi:10.18653/v1/P18-1031

56. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. In: *NIPS*. ; 2017. https://www.semanticscholar.org/paper/Attention-is-All-you-Need-Vaswani-Shazeer/204e3073870fae3d05bcbc2f6a8e263d9b72e776

57. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:190711692 [cs]*. Published online July 26, 2019. Accessed January 27, 2020. http://arxiv.org/abs/1907.11692

58. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. *Lancet*. 1974;2(7872):81-84. doi:10.1016/s0140-6736(74)91639-0

59. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830. Accessed January 5, 2023. https://www.jmlr.org/papers/v12/pedregosa11a.html

60. Fan RE, Chang KW, Hsieh CJ, Lin CJ. LIBLINEAR: A Library for Large Linear Classification. Published 2008. Accessed January 5, 2023. https://www.jmlr.org/papers/volume9/fan08a/fan08a.pdf

61. Ng AY. Feature selection, L1 vs. L2 regularization, and rotational invariance. In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML '04. Association for Computing Machinery; 2004:78. doi:10.1145/1015330.1015435

62. Crammer K, Singer Y. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. *J Mach Learn Res*. 2001;2(Dec):265-292. Accessed January 5, 2023. https://www.jmlr.org/papers/v2/crammer01a.html

63. Ilya Loshchilov FH. Decoupled Weight Decay Regularization. doi:10.48550/arXiv.1711.05101

64. Wolf T, Debut L, Sanh V, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:191003771 [cs]*. Published online July 13, 2020. Accessed January 8, 2021. http://arxiv.org/abs/1910.03771

65. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep

learning library. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc.; 2019:8026-8037. Accessed January 5, 2023. https://dl.acm.org/doi/10.5555/3454287.3455008

66. Falcon W. Pytorch lightning. *GitHub Note: https://github com/PyTorchLightning*. Published online 2019. https://scholar.google.ca/scholar?cluster=8006153255328035 43&hl=en&as_sdt=0,5&sciodt=0,5

67. Li L, Jamieson K, Rostamizadeh A, et al. A System for Massively Parallel Hyperparameter Tuning. *arXiv [csLG]*. Published online October 13, 2018. https://arxiv.org/abs/1810.05934