Data-Driven System for Perioperative Acuity Prediction

By

Linda Zhang

Thesis
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE
in

Biomedical Informatics
December, 2016
Nashville, Tennessee

Approved:
Daniel Fabbri, Ph.D.
Thomas A. Lasko, M.D., Ph.D.
Jonathan P. Wanderer, M.D.

ACKNOWLEDGEMENTS

I would like to thank my mentor, Daniel Fabbri for all of his guidance and hard work. He’s been a huge source of help and encouragement since my first year in the department, and has always made himself available to talk to even though he is so busy. I would also like to give great thanks to my committee members, Tom Lasko and Jonathan Wanderer. They’ve given great feedback and advice throughout my master’s project.

I’m extremely grateful for all the students, professors and staff of the Vanderbilt Department of Biomedical Informatics, who have been a great source of support and help. This would not have been possible without the National Library of Medicine Training Grant 4T15LM007450-14 to support my education and research.

Lastly, I would like to thank all my friends and family that have been supportive and helpful over the years.

TABLE OF CONTENTS

Page

Appendix

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

The American Society of Anesthesiologist's' (ASA) Physical Status (PS) classification is a subjective assessment of a patient's overall health. It is a widely used grading system for preoperative health in surgical patients, and consists of five classes of increasing severity with a sixth class for organ donors. Associations with ASA scores and surgical outcomes, rate of postoperative complications, operating times, hospital length of stay, morbidity rates and other postoperative outcomes have been reported in literature (Daabiss 2011).

Even though ASA score has been shown to be a predictor of postoperative outcomes, there is considerable variation in the ASA classification allocation (Daabiss 2011). It is up to the anesthesiologists to subjectively rate the patients' preoperative health. They do this by analyzing information from the patient's history and physical examination and medications. This process can take anywhere from 2-10 minutes per patient. In a day, an anesthesiologist can see on average anywhere from 5-30 patients, so in total this process can a considerable amount of time.

Though there are definitions for the class allocations in the ASA scale, the definitions do not detail the diseases and conditions for patients in each class. Despite that, the ASA scale is considered reliable. The ASA scale has been shown to have a moderate inter-rater reliability in clinical practice, with studies finding Cohen's Kappa scores of 0.40 (Riley et al., 2014), and 0.64 (Ihejirika et al., 2015). In addition, there is good agreement with how different anesthesiologists rate a patient's ASA classification, and the classification is a valid indicator of preoperative health status (Sankar et al., 2014). Since ASA score is an important indicator for outcomes, it would introduce less variability if the method of evaluating ASA score was not so subjective.

In this paper, we attempt to solve the problem of predicting the ASA score of patients. We split the problem into two parts: 1) develop a model that uses data from the electronic medical record (EMR) to predict ASA score and 2) evaluate the model and compare its agreement with anesthesiologists to literature. While the multiple ASA scores suggest multi-class classification, we find that the highest distribution of scores as well as the hardest scores to separate fall into ASA classes 2 and 3. Because of this, we reframe the problem to be a binary classification problem, classifying patients into either ASA 1/2 or ASA 3/4/5. Supervised machine learning is used to develop a system that mines the underlying factors to determine which class a patient should belong to. In this system, both numerical and categorical features are used, and we attempt to incorporate time into the ICD9 features. In contrast to previous works, we train our model using a data set of 400,000 samples and evaluate our model by looking at both performance and agreement. The model provides a systematic way of determining ASA score, which saves anesthesiologists time in ASA determination. The results of the study demonstrate that this model can accurately predict the ASA score, providing a comparable level of agreement to anesthesiologists.

CHAPTER 2

BACKGROUND

ASA score

Preoperative assessment is the evaluation of a patient’s health before an operation with the goal of reducing the patient’s surgical and anesthetic perioperative morbidity or mortality, and to return the patient to desirable functioning as quickly as possible (Zambouri, 2007). It is important in not only ensuring patient safety, but also in managing patient workflow and allocating resources for the operation. The most widely used method for preoperative assessment is the ASA classification system.

The current ASA PS classification system was proposed by Dripps et al. in 1961 and adopted by the ASA in 1962. The system consists of five classes:

I. Patient is a completely healthy fit patient.
II. Patient has mild systemic disease.
III. Patient has severe systemic disease that is not incapacitating.
IV. Patient has incapacitating disease that is a constant threat to life.
V. A moribund patient who is not expected to live 24 hour with or without surgery.

The sixth class, later added, denotes an organ donor.

The ASA PS classification was originally created for statistical data collection and reporting in anesthesia (Saklad, 1941). It is now used in many other areas: allocation of resources (Vogt & Henson, 1997), reimbursement for anesthesia services, (Australian Government Department of Health and Aging, 2005) and predicting perioperative risk (Bjorgul et al., 2010, Cullen et al., 1994, Dalton et al., 2011, Davenport et al., 2006, Glance et al., 2012, Han et al., 2004, Hightower et al., 2010, Lee et al., 1999, Malviya et al., 2011, Skaga et al., 2007, Vacanti et al., 1970., Wolters et al., 1996).

ASA score reliability

Since it is up to anesthesiologists to interpret the definitions of the ASA classification system with respect to their patients, ASA score is subjective. For example, there is inherent subjectivity when trying to distinguish cases belonging to ASA classes 2, 3 and 4, because the difference depends on differentiating between “mild systemic”, “severe systemic” and “incapacitating” diseases (Sankar et al., 2014). Despite the fact that ASA score is a subjective assessment, studies have found that it is reliable. The ASA scale has moderate to substantial inter-rater reliability in clinical practice (Riley et al., 2014, Sankar et al., 2014, Ihejirika et al., 2015).

ASA score correlation to outcomes

The ASA score is used in many areas including research, evaluation, and prediction. One reason that it is so widely used is because ASA score is correlated with outcomes (Daabiss 2011). ASA score has specific correlation with medical factors such operating time, hospital length of stay, postoperative infection rates, and morbidity rate following various types of surgery (Ridgeway et al., 2005, Tang et al., 2001, Sauvanet et al., 2005, Prause et al., 1997, Carey et al., 2006). It has significant predictive impact on blood loss during surgery (Grosflam et al., 1995) and perioperative myocardial infarction (House et al., 2016). There is evidence of significant correlation between ASA class and perioperative variables, postoperative complications, and mortality rate (Wolters et al. 1996). Due to the ASA score’s patterns of association with patient characteristics and postoperative outcomes, it is considered to be a valid measure of preoperative health status (Sankar et al., 2014).

ASA prediction using machine learning

In the past, machine learning has been used to attempt to classify patients into the ASA classes. The decision tree classifier, multi-layer perceptron (or single layer neural network), Naïve Bayes classifier, and support vector machines have all been tested and used to classify ASA score (Karpagavalli et al., 2009, Lazouni et al., 2013).

Karpagavalli et al. (2009) use a decision tree classifier, multi-layer perceptron, and Naïve Bayes classifier to categorize 362 cases with 37 features collected from private hospitals into ASA classes 1, 2 and 3. The distribution of cases was 35.64% ASA 1, 28.45% ASA 2, and 35.91% ASA 3. The study found the decision tree classifier, Naïve Bayes, and multi-layer perceptron to achieve 10-fold cross validated accuracies of 96.13%, 76.24%, and 97.79% respectively on the data set.

The study by Lazouni et al. (2013) test the decision tree classifier, support vector machines, and multi-layer perceptron classifier on a set of 898 cases with 17 features classified into ASA 1-4. The distribution of cases in this data set was 24.38% ASA 1, 43.95% ASA 2, 25.84% ASA 3, and 5.80% ASA 4. This study found a 10-fold cross validated accuracy of 88.01%, 91.43%, and 93.25% for the decision tree, multi-layer perceptron, and SVM classifiers respectively.

While both studies produce high accuracies in their results, accuracy may not be the best metric to measure performance because of bias introduced by class imbalance. In addition, both studies used a fairly small sample size to train and test their classifiers. For a model like the deep neural network, this amount of data would likely be insufficient to train with. The models that were used may not generalize well to larger data sets; for example, the SVM classifier that Lazouni et al. (2013) find to perform the best does not scale well with large amounts of data.

CHAPTER 3

METHODS

Overview

This study presents a system that builds a supervised machine learning model to predict the ASA scores of pre-operative patients. The problem is reframed so that the model classifies ASA score in a binary fashion. The system uses retrospective data with ASA scores given by anesthesiologists for pre-operative patients as labels to train a classifier. Features for the classifier were chosen from discussion with an anesthesiologist and observations made from the data. We selected features such as age, BMI and ICD9 codes as a proxy for patient history because these are all factors that anesthesiologists look at in their history and physical examination. Medications were also selected as features, because pre-operation medications play a large role in an anesthesiologist's decision. Different structures were tested for the ICD9 features in order to attempt to capture temporal aspects of ICD9 history. The trained classifier can predict the ASA score for future surgical patients.

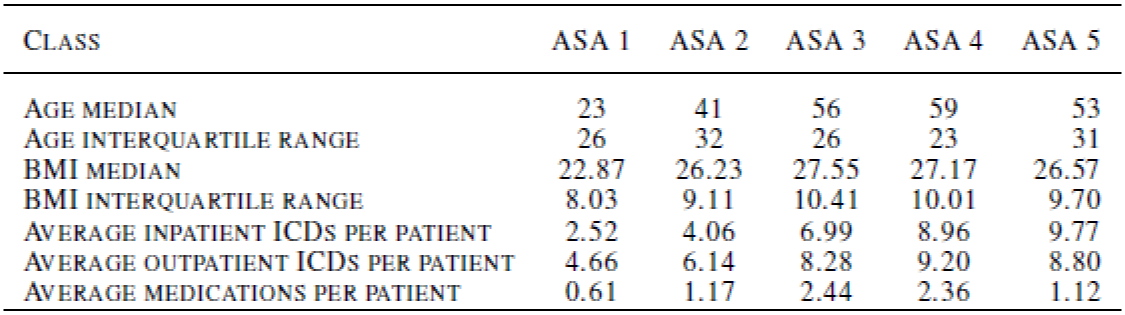
Setting

The study was conducted with data from the Vanderbilt University Medical Center (VUMC) in middle Tennessee. The subjects of the study are surgery patients from the years 1999-2014 for whom anesthesiologists recorded an ASA score.

Data

The data used to train and test our classifiers comes from the Vanderbilt Perioperative Data Warehouse (PDW), which records ASA scores for all patients who have received a surgery at Vanderbilt. A cohort of 415,000 samples from 216,000 patients that have an ICD9 history is analyzed. Of the data, a 10% holdout dataset is set aside to evaluate the final classifiers. The scores that were given by anesthesiologists are used as the label, or true score (noting that it may not always be accurate/may have variability). The ASA score 6 is excluded from the predictions and dataset, because that score indicates a brain-dead patient whose organs are being removed for donor purposes. A summary of the data for each ASA class is shown (Table 1).

Table 1. Summary statistics for data set separated by ASA class.



CLASS	ASA 1	ASA 2	ASA 3	ASA 4	ASA 5
AGE MEDIAN	23	41	56	59	53
AGE INTERQUARTILE RANGE	26	32	26	23	31
BMI MEDIAN	22.87	26.23	27.55	27.17	26.57
BMI INTERQUARTILE RANGE	8.03	9.11	10.41	10.01	9.70
AVERAGE INPATIENT ICDS PER PATIENT	2.52	4.06	6.99	8.96	9.77
AVERAGE OUTPATIENT ICDS PER PATIENT	4.66	6.14	8.28	9.20	8.80
AVERAGE MEDICATIONS PER PATIENT	0.61	1.17	2.44	2.36	1.12

With labels, we can apply supervised machine learning algorithms to the data, and attempt to predict ASA score. This approach is feasible because, though there is variability in ASA scores given by anesthesiologists, studies have shown that there is moderate inter-rater reliability and good agreement between anesthesiologists (Sankar et al., 2013), and over a large amount of data, the variability will even out. From the distribution of the data (Figure 1) and the knowledge that the hardest scores to distinguish were ASA scores 2 and 3, we reframed the problem as a binary classification problem.

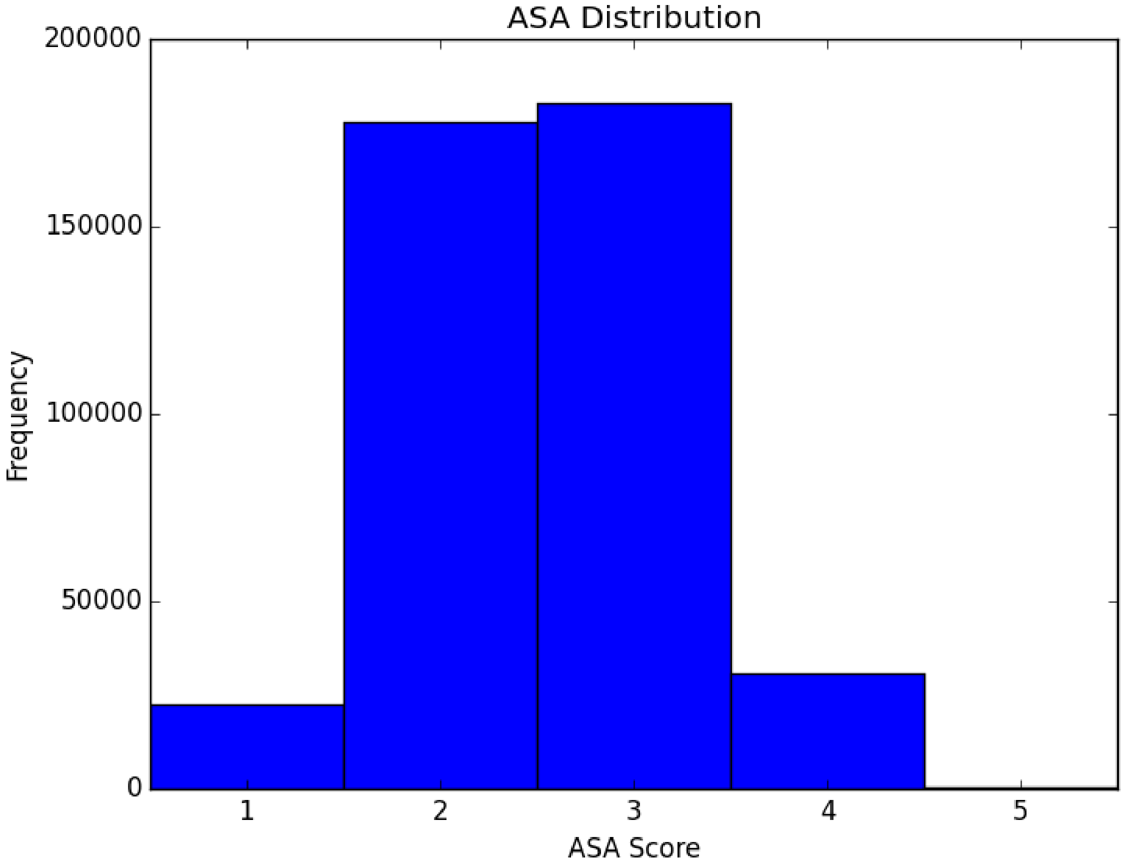


Figure 1. Distribution of ASA scores in the data. Cases with ASA scores 2 and 3 dominate the distribution.

Besides ASA score, the PDW provides age, BMI, surgery type, and inpatient ICD9 codes. In addition to the ASA scores and patient data from the PDW, we have access to the ICD9 outpatient history from the Vanderbilt Enterprise Data Warehouse (EDW), which is associated with patient healthcare encounters. The co-occurrence of inpatient and outpatient ICD9 chapters with ASA can be observed in Figure 2 and Figure 3. The co-occurrence trends indicate that the value of ICD9 codes for predicting ASA score. All data that is collected is pre-operation and normalized by frequency of both chapter and ASA score.

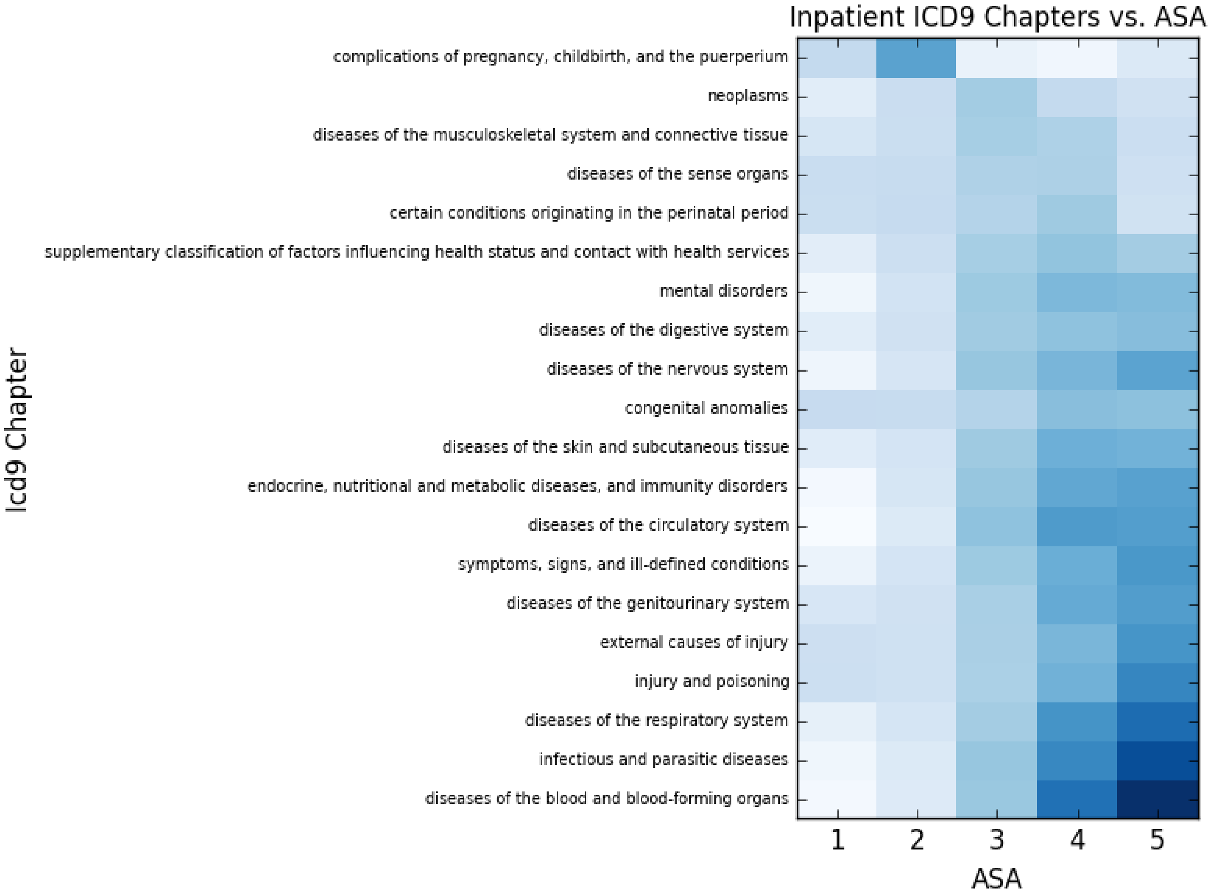


Figure 2. Co-occurrence of inpatient ICD9 chapters with ASA score normalized by frequency of ASA score and chapter. A darker color indicates greater presence of the chapter in those cases.

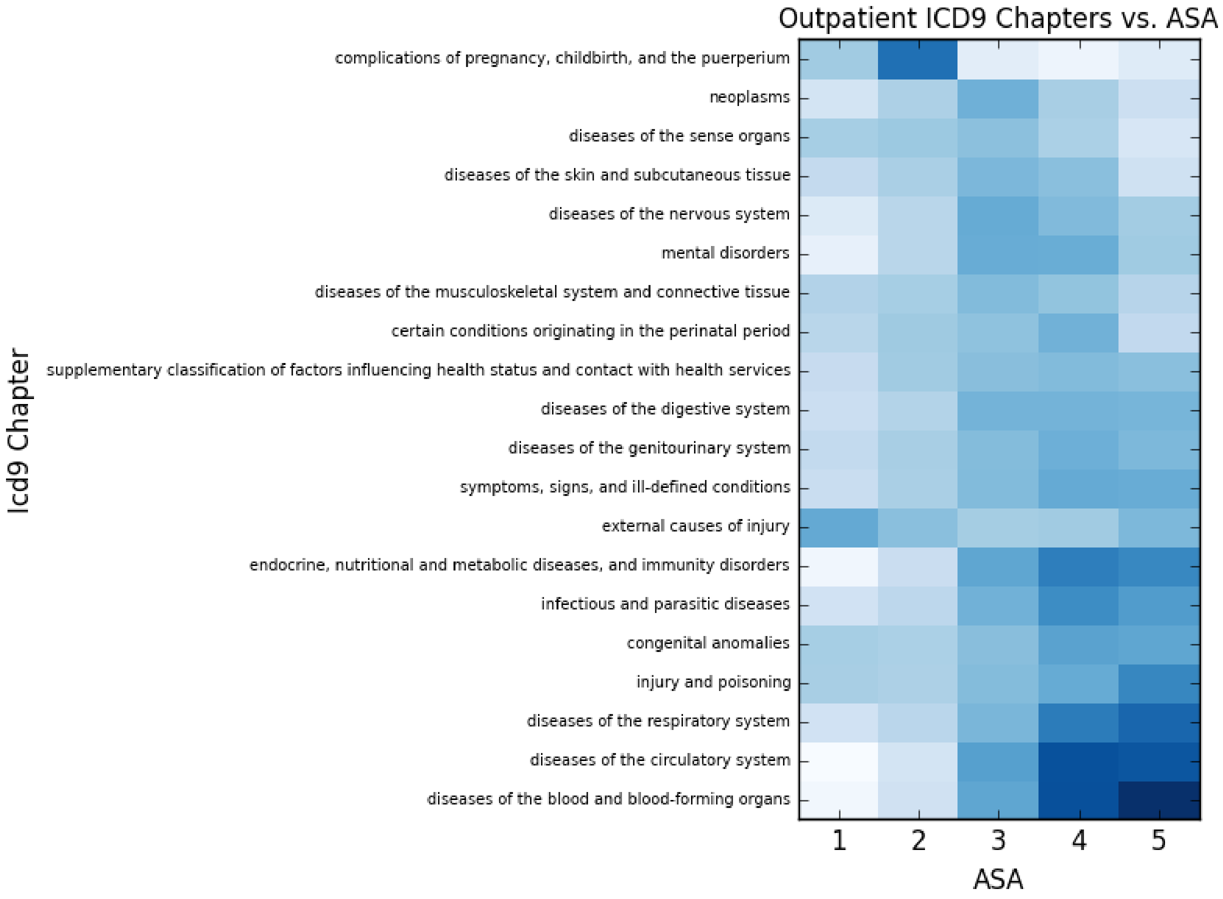


Figure 3. Co-occurrence of outpatient ICD9 chapters with ASA score normalized by frequency of ASA score and chapter. A darker color indicates greater presence of the chapter in those cases.

Supervised machine learning

Supervised machine learning is the task of building a model or function from a set of labeled training data. The four machine learning methods that we use in this study are: logistic regression, k-nearest neighbors (KNN), random forests, and deep neural networks.

Logistic regression

Logistic regression is a method that measures the relationship between the categorical dependent variable and any number of independent variables by estimating probabilities using a logistic function (McCullagh et al., 1989).

K-nearest neighbors

The k-nearest neighbors method is a machine learning model that classifies new data points based on which previously seen data points it is most similar to (Cover & Hart, 1967). In a classification task, the model finds the k nearest data points to the new case and assigns the majority class.

Random forests

Random forests is a machine learning method that uses an ensemble of decision tree predictors. The method builds a number of trees each from a random, independently sampled subset of variables from the original feature set. It classifies new cases by taking a majority vote of the decision trees it has built (Breiman, 2001).

Deep neural networks

In the past decade, a machine learning model called neural networks have become very popular. A standard neural network consists of a number of simple, connected processors called neurons. These neurons are activated through inputs or from other neurons, and the network learns the weights that make a desired output (Schmidhuber, 2015). Deep neural networks use feed forward neural networks with many layers. They consist of an input feeding into a number of fully connected hidden layers which feed into an output. Convolutional neural networks are deep neural networks with limited connections that are localized based on the input data, and use convolutional filters. An example is an image, in which pixels are locally connected to surrounding pixels. Deep learning models scale well to large data sets, but require a large training data set.

Feature analysis and selection

We selected classes (or categories) of features that we believed, from discussion with an anesthesiologist, are factors that anesthesiologists look for in pre-operation patients. The different classes of features selected were age, body mass index (BMI), surgery service, if the patient previously had a surgery, preoperative medications, prior inpatient ICD9 codes and history of outpatient ICD9 codes. Temporality was incorporated into the ICD9 features by sectioning codes into stretches of time (we use months) for which they were placed, up to a year before the surgery. The various feature classes are:

Age: Age of the patient, as a continuous numeric value.

BMI: Body mass index of the patient, as a continuous decimal value.

Surgery service: Category of surgery the patient is about to receive. Each category (70 total) is a binary feature.

Previous surgery: If the patient has previously had a surgery, a single binary feature.

Preoperative medications: Medications classified in hierarchy, using the top-level hierarchy yielding 21 categories. Each category is a feature, with numerical counts for each instance of medication in that category.

Inpatient ICD9 codes: ICD9 codes received while in the hospital. In creating these features, we tested numerical counts for: raw codes, parent codes, ICD9 chapters, PheWAS (Denny et al., 2013) classification (Vanderbilt-developed Phenome-Wide Association Study classification), ICD9 hierarchy and temporally structured ICD9 chapters.

Outpatient ICD9 codes: ICD9 codes received while not admitted to the hospital. In creating these features, we tested numerical counts for: raw codes, parent codes, ICD9 chapters, PheWAS classification, ICD9 hierarchy and temporally structured ICD9 chapters.

We tested different combinations of feature classes and created a final combined classifier. The combined classifier and classifiers using a single feature class (i.e. ASA prediction with only age) are compared to analyze how predictive each classifier was on ASA score. Performance was measured with 5-folds cross validation and the AUC was used to select the features for the final combined classifier.

Hierarchy

The feature classes ICD9 codes and medications both have a hierarchical structure. A recent study has shown that leveraging hierarchy in ICD9 features improves prediction (Singh et al., 2014). The basic idea behind leveraging hierarchy is that there are relationships between the parents and children in the ICD9 coding structure. Children under the same parent code have correlated relationships, which may have a relationship with some outcome of interest. The approach developed in the study consists of two parts: 1) map each ICD9 code to a conditional probability of outcome using the training data and 2) use the probabilities to construct features. The first part estimates the probability of the positive prediction given the subtree of the hierarchy at each node (ICD9 code). The second part constructs features using top level codes, and setting values by taking the maximum conditional probability from the subtree nodes. We attempt to leverage the hierarchy in both our ICD9 code and medication feature classes.

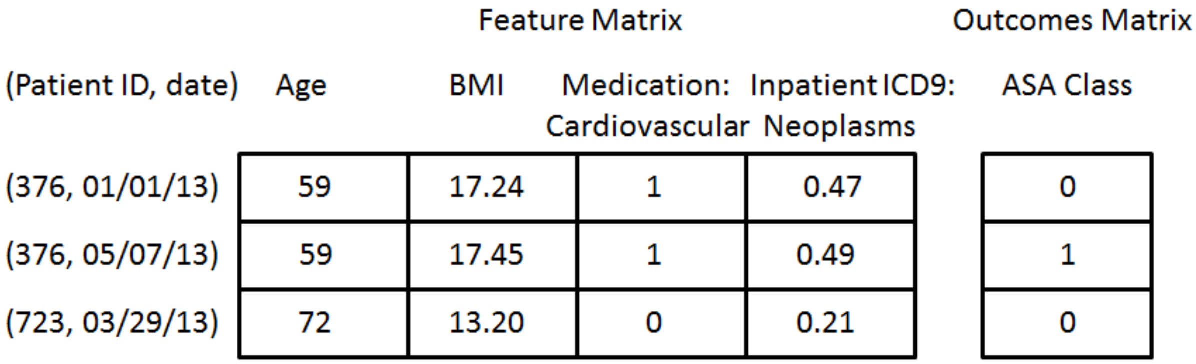
ICD features

We realized that there was more information in the ICD9 codes that we could incorporate in features. ICD9 codes have a hierarchical structure, and could be generalized to higher level features. In addition, ICD9 codes are not all given at the same point in time. They accumulate over time as a patient comes into the hospital; because of this, we hypothesized that ICD9 codes given closer to the surgery date would have a greater impact on their physical status. In testing different structures for the ICD9 code features, two different approaches were used: ICD9 hierarchy and temporally structured ICD9 chapters. We created our hierarchy-leveraging features by using the approach outlined in the study by Singh et al. (2014). For our temporally structured ICD9 chapters, we took the past year of surgeries and counted the number of each ICD9 chapter found in each month (12 total months). In short, each chapter with each month was turned into a separate feature.

For the features from ICD9 codes, we tested: raw codes, parent codes, ICD9 chapters, PheWAS classification, and ICD9 hierarchy and temporally structured ICD9 chapters. The raw ICD9 codes resulted in about 4,000 features, while both the parent codes and PheWAS codes resulted in about 1,000 features. The ICD9 chapters result in 20 features, the ICD9 hierarchy results in 20 features, and the temporally structured ICD9 chapters result in 240 features.

Feature matrix

From the collected data, feature matrices were created to be used as the input for the machine learning algorithms (Figure 4).



	Feature Matrix				Outcomes Matrix
(Patient ID, date)	Age	BMI	Medication: Cardiovascular	Inpatient ICD9: Neoplasms	ASA Class
(376, 01/01/13)	59	17.24	1	0.47	0
(376, 05/07/13)	59	17.45	1	0.49	1
(723, 03/29/13)	72	13.20	0	0.21	0

Figure 4. Example feature matrix for the ASA classifier. Each row represents one patient case, and each column in the feature matrix represents a feature. If a patient has multiple surgeries, one row is included per surgery.

All features are collected from data that is available from before the patient has their surgery.

Classification

Python’s scikit-learn (Buitinck et al., 2013) and lasagne (Dielman et al., 2015) packages were used to develop the classifiers (lasagne for deep learning) for predicting ASA score. We tune the hyperparameters of our machine learning models using hyperopt (Bergstra et al., 2013) to search over a specified space. To simplify the problem, we remove the trivial ASA 6 score (which indicates dead), and separate the scores into two classes: ASA 1/2, and ASA 3/4/5. This split turns the problem into a binary classification problem. The supervised machine learning techniques that we test in our study include logistic regression, k-nearest neighbors, random forest, deep neural networks and a hybrid network consisting of both deep and convolutional neural networks.

Deep learning for temporal data

We attempt to use convolutional neural networks to find the relationships in the temporal ICD9 data. In this data set, the localized features are the co-occurrence of ICD9 chapters in time, one month connected to the next month. The filters attempt to capture information from temporal proximity.

The hybrid neural network uses both networks to learn from the data. The deep neural network operates on the non-temporal data (age, BMI, etc) while the convolutional neural network takes temporal data as its input, or the temporal outpatient ICD9 features. We used a convolutional neural network in order to find features within the temporal data. The output hidden units from the convolutional neural network are combined with the hidden units from the last hidden layer of the deep neural network, and concatenated to form a layer of hidden units from both. The resulting layer is fed into an output layer, which uses a softmax as the activation function.

Optimized parameters

The final parameters from hyperopt that we found were logistic regression with an elastic net penalty, k-nearest neighbors with k=4, random forests using 79 estimators and a depth of 19, deep neural networks with 0.1 hidden drop, a depth of 3 and width of 400, and the hybrid network with the same deep network parameters plus convolutional network parameters with 20 filters and a window size of 3.

Classifier evaluation

We evaluated the performance of our classifiers using 5-fold cross validation and calculating the average ROC AUC scores, as well as calculating the AUC on our holdout dataset. To gain insight into the amount of data required for training and the plateau of performance, we analyzed the AUC at different sized training sets of the random forest and deep neural network. Since AUC correlates with Kappa (Ben-David, 2008), we calculated a theoretical human achievable AUC corresponding to the best unweighted Kappa statistic using the following equation from the paper:

$$TP = Q(1- K)(1-2N) + KP + FP \tag{1}$$

We assign the ASA class 3/4/5 to the positive class; the variables are TP = true positives, FP = false positives, Q = predicted positives, P = positives, N = negatives, K = Kappa. We know P = 0.517 and N = 0.483 from the proportion of ASA class 1/2 and 3/4/5, but make the assumption that Q = 0.517 in the best case scenario that all positives are true positives for calculations. We then create an AUC curve corresponding to K = 0.64, the highest unweighted Kappa in literature (Ihejirika et al., 2015). We plot this theoretical human AUC against the performance AUCs of our classifiers.

Distribution of model predictions

We calculate the prediction probability of our best random classifier for all cases, and observe the correct/misclassified ratio of cases for different predicted probabilities. Analyzing the distribution of these predictions allows us to determine which/how many cases are being classified correctly or incorrectly, and how certain the model thinks those classifications are.

Cohen’s Kappa

In order to analyze the effect of the moderate variability between raters, we treated our model as a rater, and the raw data as another rater. We calculated the unweighted Cohen’s Kappa for inter-rater agreement between our model and the raw scores. This statistic measures the agreement between the two “raters”. We use the unweighted Kappa because our model uses only two classes. We reason that since the raw data is a representation of a number of raters which should have good agreement with each other, the model should agree with the raw data, as it acts as another rater. If our calculated Cohen’s Kappa is comparable to the ones found in literature, the model should be as agreeable as the opinion of another anesthesiologist.

Manual review

We evaluated the performance of our model on a subset of 117 cases that had been reviewed by a single anesthesiologist. With these cases, we compared the ASA score determined by the model and the anesthesiologist, and calculated both AUC and Cohen’s Kappa.

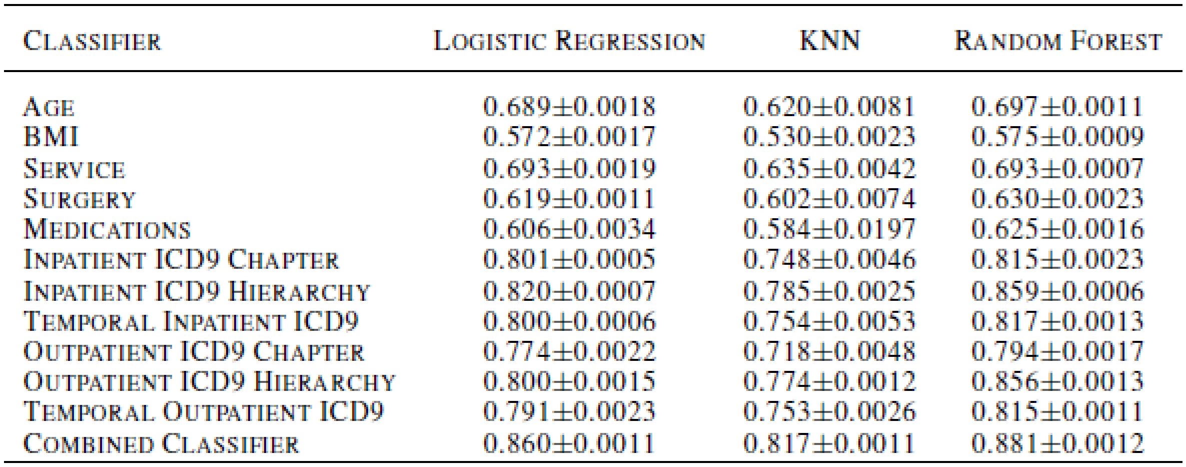
CHAPTER 4

RESULTS

Individual feature class results

The results for the individual classifiers and combined classifier for the logistic regression, k-nearest neighbors and random forest methods can be seen in the following table.

Table 2. The average ROC AUC scores for each individual feature class, and combined classifier measured using 5-fold cross validation.



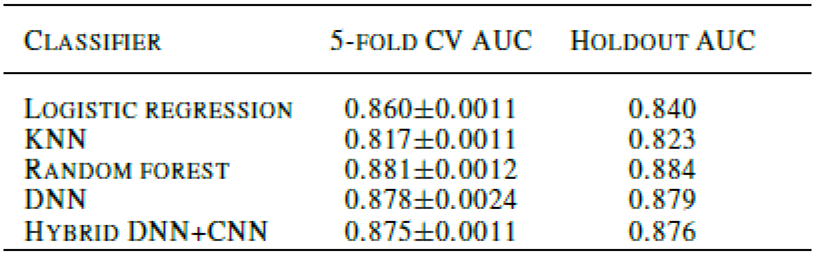
Classifier	Logistic Regression	KNN	Random Forest
Age	0.689±0.0018	0.620±0.0081	0.697±0.0011
BMI	0.572±0.0017	0.530±0.0023	0.575±0.0009
Service	0.693±0.0019	0.635±0.0042	0.693±0.0007
Surgery	0.619±0.0011	0.602±0.0074	0.630±0.0023
Medications	0.606±0.0034	0.584±0.0197	0.625±0.0016
Inpatient ICD9 Chapter	0.801±0.0005	0.748±0.0046	0.815±0.0023
Inpatient ICD9 Hierarchy	0.820±0.0007	0.785±0.0025	0.859±0.0006
Temporal Inpatient ICD9	0.800±0.0006	0.754±0.0053	0.817±0.0013
Outpatient ICD9 Chapter	0.774±0.0022	0.718±0.0048	0.794±0.0017
Outpatient ICD9 Hierarchy	0.800±0.0015	0.774±0.0012	0.856±0.0013
Temporal Outpatient ICD9	0.791±0.0023	0.753±0.0026	0.815±0.0011
Combined Classifier	0.860±0.0011	0.817±0.0011	0.881±0.0012

The ICD9 chapter counts, ICD9 hierarchy, and temporal ICD9 chapter counts outperformed the other structures for ICD9 codes significantly (as a result, they are not shown Table 2; the full table of all the differently structured feature classes can be seen in Appendix A). We found that the best combination of feature classes was: age, BMI, service, medications, inpatient ICD9 hierarchy, outpatient ICD9 chapters. This combination resulted in the highest AUCs for each combined classifier.

Combined classifier results

The results for the combined classifiers under each machine learning method are presented in the following Table 3.

Table 3. Classification AUC for combined classifiers using logistic regression, k-nearest neighbors, random forests, deep neural network and combination deep and convolutional neural network.



Classifier	5-fold CV AUC	Holdout AUC
Logistic regression	0.860±0.0011	0.840
KNN	0.817±0.0011	0.823
Random forest	0.881±0.0012	0.884
DNN	0.878±0.0024	0.879
Hybrid DNN+CNN	0.875±0.0011	0.876

The best machine learning classifier is the random forest, which achieved an AUC of **0.884** on the holdout dataset. We found that the best deep neural network performs as well as the best hybrid deep/convolutional neural network. In addition, logistic regression performs well with an AUC of 0.84.

Learning curve and theoretical human AUC

We also compared the learning rate of the random forest to the deep neural network (Figure 5). The random forest achieves a high AUC with much less data than the deep neural network requires. Both classifiers plateau approximately 0.5 below the theoretical maximum AUC (**0.935**).

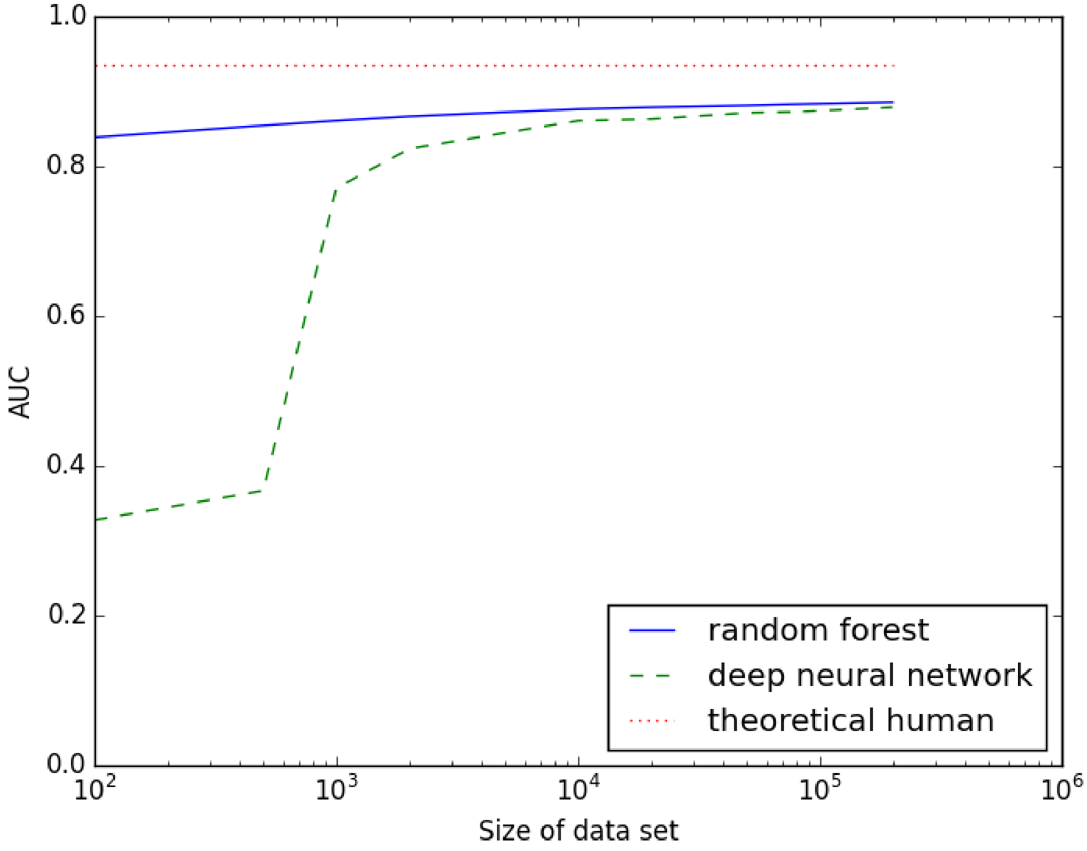


Figure 5. Learning rate of deep neural network and random forest plotted with the theoretical human AUC (0.935) calculated with Kappa from literature.

Distribution of predicted probabilities

We graphed the predicted probabilities of the model from 0 to 1. The results can be seen in figure 6.

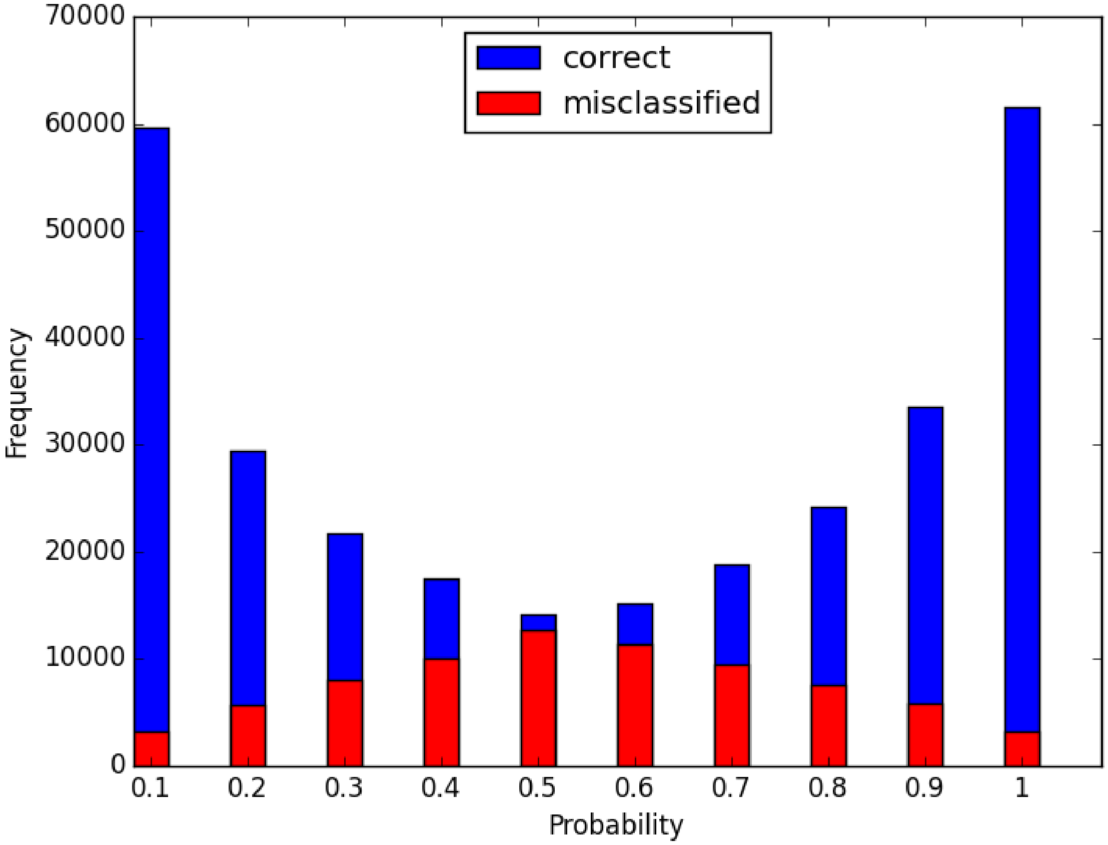


Figure 6. Predicted probabilities of the random forest model for every case in the holdout set using a threshold of 0.5: <0.5 predicts ASA 1/2 and >0.5 predicts ASA 3/4/5. The classifier predicts "easy" cases with high accuracy (near p=0 and p=1).

Model agreement

The Cohen's Kappa of the model compared to the raw data was **0.63**. A model trained on all the data and tested on the dataset reviewed by the single anesthesiologist resulted in an AUC of **0.770** and Cohen's Kappa of **0.541**. These Kappa scores are comparable to scores found in literature, which came from experiments that compare the inter-rater reliability of different anesthesiologists. Because the scores are comparable, we infer that our model can act as a separate anesthesiologist that will predict scores that agree with those of real anesthesiologists.

CHAPTER 5

DISCUSSION

In this study, we designed a system that predicts ASA score given EMR data. Through testing, we found that a random forest classifier with features classes of age, BMI, service, medications, inpatient ICD9 hierarchy, and outpatient ICD9 chapters produced the highest AUC. Because of pseudo gold standard labels that we used, we tested the inter-rater reliability of our model by calculating the Cohen's Kappa statistic. The model has moderate inter-rater reliability with the anesthesiologists in the data set, comparable to the Cohen's Kappa found in literature.

To find the best way to represent the medications and ICD9 code feature classes, we tested various feature granularities. We found that the individual medication and ICD9 code features did not improve performance, and that the top level categories in the hierarchy for both performed the best. In fact, we observed that more granular categories (for example, each ICD9 code as a feature compared to the ICD9 chapters) caused overfitting in the models, and actually produced a lower AUC in general.

Adding temporality to the ICD9 features did not increase the AUC of any of the models significantly. This implies that either temporality is not important, only the occurrence or co-occurrence of certain diseases/problems matters, or the temporality could not sufficiently be represented. Because temporality did not improve prediction, the hybrid deep and convolutional neural network did not perform better than the deep neural network.

We found that medications did not have as high of an impact on prediction as our anesthesiologist had originally stated. We hypothesize that this is due to the fact that the pre-op medications and previous diagnosis via ICD9 codes give overlapping insight into the same information on the patient's status.

Results show that on the holdout set, random forests performed the best, closely followed by deep neural networks. Random forests also require much less data to achieve a high AUC than deep neural networks. The high performance of logistic regression implies that there are simple, linear relationships between features and the ASA score. The structure of the best-performing DNN is a network with 3 layers, using 400 units. Using more layers decreased the AUC of the model, implying that the increased number of layers and connections did not add any value to the model prediction, and instead increased overfitting. The theoretical maximum AUC calculation demonstrates that there is likely a plateau in which AUC cannot improve due to the variability in ASA data. All models fall short of the theoretical maximum AUC, which implies that there may be room to improve.

Since ASA assignments may be varied between anesthesiologists, we find interrater reliability of our model versus the raw data using the Cohen's Kappa statistic. We decided to use this method after observing that Cohen's Kappa for ASA has been calculated in literature by averaging rater versus reference Kappa values (Riley et al., 2014, Ihejirika et al., 2015) or calculating group rater versus rater Kappa values (Sankar et al., 2014). The Cohen's Kappa for our model compared to the raw data was found to be 0.63. The model compared to data of a specific anesthesiologist resulted in a Cohen's Kappa of 0.54. In literature, unweighted Kappa scores of 0.40 (Riley et al., 2014), and 0.64 (Ihejirika et al., 2015) are found. Our score is comparable to these scores, implying that our model can be thought of as another reliable anesthesiologist.

The graph of predicted probabilities from the random forest classifier shows that the cases in which the model predicts a probability near 0.5 are the hardest cases (high percentage misclassified), while the cases with probabilities near 0 and 1 contain the most correct cases. In addition, more of the total cases fall into the ends (probability near 0 or 1). From a practical standpoint, the model can be used to accurately evaluate a majority of cases (easy cases) and require manual review only for the more difficult cases, reducing the workload for anesthesiologists.

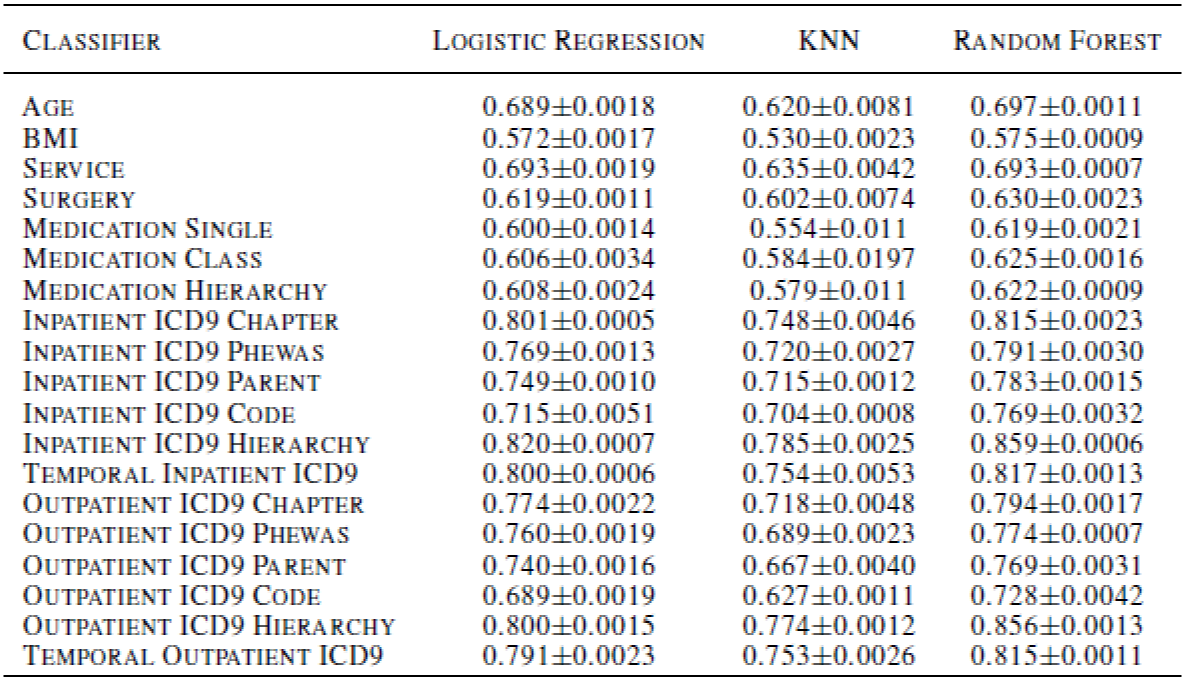
There are several limitations in our study. One is that our labels, or golden standard, are the ASA scores designated by one physician. Some of these cases can be misclassified. We hope this is mitigated by the size of the data set and the fact that anesthesiologists' inter-rater reliability is moderate. We address this limitation by providing the Cohen's Kappa statistic for the predictions against all the data and against a single anesthesiologist. Even though our calculated Cohen’s Kappa doesn’t directly compare to those in literature, it acts as a comparable estimate. The second major limitation is that we only predict the ASA classes in two groups: ASA 1/2 and ASA 3/4/5. We argue, however, that this is sufficient, because the distinction between the ASA classes 2 and 3 is the hardest to determine, cases generally fall into ASA 2 or 3, and because we predict a high number of “easy” cases with high accuracy. The third limitation is that the data is only limited to Vanderbilt, and has not been tested on a broader, multi-institutional corpus. Future directions could include training and testing the model on EMR data from other institutions, and examining the differences or variance.

Conclusion

This study attempts to solve the problem of creating a reliable model to rate patients on the ASA scoring scale. This model differs from previous work in that it uses more modern models, a significantly greater amount of training data, and uses retrospective anesthesiologist's' scores on cases as the golden standard instead of manually reviewed cases and scores. The output classifier for two class prediction achieves an AUC of 0.884. Modeling the classifier and raw data as raters, we achieve a Cohen's Kappa of 0.63, with a Cohen's Kappa of 0.54 against our anesthesiologist, comparable to scores seen in literature that range 0.4-0.64. This work demonstrates the feasibility of using this model as a standardized ASA scorer.

APPENDIX A

The following table includes the 5-fold cross validated AUC for all single feature class classifiers that we tested.



Classifier	Logistic Regression	KNN	Random Forest
Age	0.689±0.0018	0.620±0.0081	0.697±0.0011
BMI	0.572±0.0017	0.530±0.0023	0.575±0.0009
Service	0.693±0.0019	0.635±0.0042	0.693±0.0007
Surgery	0.619±0.0011	0.602±0.0074	0.630±0.0023
Medication Single	0.600±0.0014	0.554±0.011	0.619±0.0021
Medication Class	0.606±0.0034	0.584±0.0197	0.625±0.0016
Medication Hierarchy	0.608±0.0024	0.579±0.011	0.622±0.0009
Inpatient ICD9 Chapter	0.801±0.0005	0.748±0.0046	0.815±0.0023
Inpatient ICD9 PheWAS	0.769±0.0013	0.720±0.0027	0.791±0.0030
Inpatient ICD9 Parent	0.749±0.0010	0.715±0.0012	0.783±0.0015
Inpatient ICD9 Code	0.715±0.0051	0.704±0.0008	0.769±0.0032
Inpatient ICD9 Hierarchy	0.820±0.0007	0.785±0.0025	0.859±0.0006
Temporal Inpatient ICD9	0.800±0.0006	0.754±0.0053	0.817±0.0013
Outpatient ICD9 Chapter	0.774±0.0022	0.718±0.0048	0.794±0.0017
Outpatient ICD9 PheWAS	0.760±0.0019	0.689±0.0023	0.774±0.0007
Outpatient ICD9 Parent	0.740±0.0016	0.667±0.0040	0.769±0.0031
Outpatient ICD9 Code	0.689±0.0019	0.627±0.0011	0.728±0.0042
Outpatient ICD9 Hierarchy	0.800±0.0015	0.774±0.0012	0.856±0.0013
Temporal Outpatient ICD9	0.791±0.0023	0.753±0.0026	0.815±0.0011

REFERENCES

[1] Daabiss, M. (2011). American Society of Anesthesiologists physical status classification. Indian Journal of Anaesthesia, 55(2), 111–115. http://doi.org/10.4103/0019-5049.79879

[2] Riley R., Holman C., & Fletcher D. Inter-rater reliability of the ASA physical status classification in a sample of anaesthetists in Western Australia. Anaesth Inensive Care(2014), 42(5), 614-618.

[3] Ihejirika R.C., Thakore R. V., Sathiyakumar V., Ehrenfeld J. M., Obremskey W.T., & Sethi M. K. An assessment of the inter-rater reliability of the ASA physical status score in the orthopaedic trauma population. Injury (2015), 46(4), 542-546.

[4] Sankar, A., Johnson, S. R., Beattie, W. S., Tait, G., & Wijeysundera, D. N. (2014). Reliability of the American Society of Anesthesiologists physical status scale in clinical practice. BJA: British Journal of Anaesthesia, 113(3), 424–432.

[5] Zambouri A. Preoperative evaluation and preparation for anesthesia and surgery. Hippokratia. 2007;11(1):13-21.

[6] Dripps RD. New classification of physical status. Anesthesiol. 1963;24:111.

[7] Saklad M. Grading of patients for surgical procedures. Anesthesiology. 1941;2:281–4.

[8] Bjorgul K, Novicoff WM, Saleh KJ. American Society of Anesthesiologist physical status score may be used as a comorbidity index in hip fracture surgery. J Arthroplasty. 2010;25:134–7.

[9] Cullen DJ, Apolone G, Greenfield S, Guadagnoli E, Cleary P. ASA Physical Status and age predict morbidity after three surgical procedures. Ann Surg. 1994;220:3–9.

[10] Dalton JE, Kurz A, Turan A, Mascha EJ, Sessler DI, Saager L. Development and validation of a risk quantification index for 30-day postoperative mortality and morbidity in noncardiac surgical patients. Anesthesiology. 2011;114:1336–44.

[11] Davenport DL, Bowe EA, Henderson WG, Khuri SF, Mentzer RM., Jr National Surgical Quality Improvement Program (NSQIP) risk factors can be used to validate American Society of Anesthesiologists Physical Status Classification (ASA PS) levels. Ann Surg. 2006;243:636–41.

[12] Glance LG, Lustik SJ, Hannan EL, et al. The surgical mortality probability model: derivation and validation of a simple risk prediction rule for noncardiac surgery. Ann Surg. 2012;255:696–702.

[13] Han K-R, Kim HL, Pantuck AJ, Dorey FJ, Figlin RA, Belldegrun AS. Use of American Society of Anesthesiologists physical status classification to assess perioperative risk in patients undergoing radical nephrectomy for renal cell carcinoma. Urology. 2004;63:841–6. [PubMed]

[14] Hightower CE, Riedel BJ, Feig BW, et al. A pilot study evaluating predictors of postoperative outcomes after major abdominal surgery: physiological capacity compared with the ASA physical status classification system. Br J Anaesth. 2010;104:465–71.

[15] 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;100:1043–9.

[16] Malviya S, Voepel-Lewis T, Chiravuri SD, et al. Does an objective system-based approach improve assessment of perioperative risk in children? A preliminary evaluation of the 'NARCO' Br J Anaesth. 2011;106:352–8.

[17] Skaga NO, Eken T, Søvik S, Jones JM, Steen PA. Pre-injury ASA physical status classification is an independent predictor of mortality after trauma. J Trauma Acute Care Surg. 2007;63:972–8.
[18] Vacanti CJ, VanHouten RJ, Hill RC. A statistical analysis of the relationship of physical status to postoperative mortality in 68,388 cases. Anesth Analg. 1970;49:564–6.
[19] Wolters U, Wolf T, Stützer H, Schröder T. ASA classification and perioperative variables as predictors of postoperative outcome. Br J Anaesth. 1996;77:217–22.
[20] Ridgeway S, Wilson J, Charlet A, Pearson A, Coello R. Infection of the surgical site after arthroplasty of the hip. J Bone Joint Surg Br. 2005;87:844–50.
[21] Tang R, Chen HH, Wang YL, Changchien CR, Chen JS, Hsu KC, et al. Risk factors for surgical site infection after elective resection of the colon and rectum: A single-center prospective study of 2,809 consecutive patients. Ann Surg. 2001;234:181–9.
[22] Sauvanet A, Mariette C, Thomas P, Lozac'h P, Segol P, Tiret E. Mortality and morbidity after resection for adenocarcinoma of the gastroesophageal junction: Predictive factors. J Am Coll Surg. 2005;201:253–62.
[23] Prause G, Offner A, Ratzenhofer-Komenda B, Vicenzi M, Smolle J, Smolle-Juttner F. Comparison of two preoperative indices to predict perioperative mortality in non-cardiac thoracic surgery. Eur J Cardiothorac Surg. 1997;11:670–5.
[24] Carey MS, Victory R, Stitt L, Tsang N. Factors that influence length of stay for in-patient gynecology surgery: Is the Case Mix Group (CMG) or type of procedure more important? J Obstet Gynaecol Can. 2006;28:149–55.
[25] Grosflam JM, Wright E, Cleary P, Katz J. Predictors of blood loss during total hip replacement surgery. Arthritis Care Res. 1995;8:167–73.
[26] House LM, Marolen KN, St. Jacques PJ, McEvoy MD, Ehrenfeld JM. Surgical Apgar score is associated with myocardial injury after noncardiac surgery. Journal of Clinical Anesthesia. 2016; 34: 395-402.
[27] Karpagavalli, S., Jamuna, K.S., & Vijaya, M.S. Machine learning approach for preoperative anaesthetic risk prediction. International Journal of Recent Trends in Engineering, 1(2), 19-22.
[28] Lazouni, M. Daho, M., Settouti, N., Chikh , M., & Mahmoudi, S. Machine Learning Tool for Automatic ASA Detection. Modeling Approaches and Algorithms for Advanced Computer Applications. Volume 488, 9-16.
[29] McCullagh, Peter, and John A. Nelder. Generalized linear models. Vol. 37. CRC press, 1989.
[30] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967;13(1):21-27.
[31] Breiman, L. Random forests. Machine Learning. 2001; 45(1): 5-32.
[32] Schmidhuber, J. Deep learning in neural networks: an overview. Neural Networks 2015; 61, 85-117.
[33] Denny JC, Bastarache L, Ritchie MD et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 2013 Dec;31(12):1102-10.
[34] Singh A., Nadkarni G., Guttag J., & Bottinger E. Leveraging hierarchy in medical codes for predictive modeling. Proceedings of the 5th ACM conference on bioinformatics, computational biology and health informatics(2014), 96-103.

[35] Buitinck et al. API design for machine learning software: experiences from the scikit-learn project. ECML PKDD Workshop: Languages for Data Mining and Machine Learning (2013), 108-122.
[36] Dielman S., et al. Lasagne: First release. http://dx.doi.org/10.5281/zenodo.27878. (2015)
[37] Bergstra, J., Yamins, D., Cox, D. D. (2013) Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. Proc. of the 30th International Conference on Machine Learning (ICML 2013).
[38] Ben-David, A. About the relationship between ROC curves and Cohen's kappa. Engineering applications of artificial intelligence(2008), 874-882.