Development and Validation of a Deep Neural Network Model for Prediction of Postoperative In-hospital Mortality

Christine K. Lee, M.S., Ph.D., Ira Hofer, M.D., Eilon Gabel, M.D., Pierre Baldi, Ph.D., Maxime Cannesson, M.D., Ph.D.

ABSTRACT

Background: The authors tested the hypothesis that deep neural networks trained on intraoperative features can predict postoperative in-hospital mortality.

Methods: The data used to train and validate the algorithm consists of 59,985 patients with 87 features extracted at the end of surgery. Feed-forward networks with a logistic output were trained using stochastic gradient descent with momentum. The deep neural networks were trained on 80% of the data, with 20% reserved for testing. The authors assessed improvement of the deep neural network by adding American Society of Anesthesiologists (ASA) Physical Status Classification and robustness of the deep neural network to a reduced feature set. The networks were then compared to ASA Physical Status, logistic regression, and other published clinical scores including the Surgical Apgar, Preoperative Score to Predict Postoperative Mortality, Risk Quantification Index, and the Risk Stratification Index.

Results: In-hospital mortality in the training and test sets were 0.81% and 0.73%. The deep neural network with a reduced feature set and ASA Physical Status classification had the highest area under the receiver operating characteristics curve, 0.91 (95% CI, 0.88 to 0.93). The highest logistic regression area under the curve was found with a reduced feature set and ASA Physical Status (0.90, 95% CI, 0.87 to 0.93). The Risk Stratification Index had the highest area under the receiver operating characteristics curve, at 0.97 (95% CI, 0.94 to 0.99).

Conclusions: Deep neural networks can predict in-hospital mortality based on automatically extractable intraoperative data, but are not (yet) superior to existing methods. **(ANESTHESIOLOGY 2018; 129:649-62)**

ABOUT 230 million surgeries are performed annually worldwide.[1] While the postoperative mortality is low, less than 2%, about 12% of all patients—the high-risk surgery group—account for 80% of postoperative deaths.[2,3] To assist in guiding clinical decisions and prioritization of care, several perioperative clinical and administrative risk scores have been proposed.

The goal of perioperative clinical risk scores is to help guide care in individual patients by planning clinical management and allocating resources. The goal of perioperative administrative risk scores (based on diagnoses and procedures) is to help compare hospitals. In the perioperative setting, frequently used risk scores include the American Society of Anesthesiologists (ASA) Physical Status Classification (a preoperative score) and the Surgical Apgar score.[4,5] The ASA score was developed in 1963 and remains widely used.[4] Its main limitation is that it is subjective, it presents with high inter- and intrarater variability, it cannot be automated, and it relies on clinicians' experience. The Surgical Apgar score (an intraoperative score) uses three variables: (1) estimated

Editor's Perspective

What We Already Know about This Topic

- Robust predictions are required to compare perioperative mortality among hospitals
- Deep neural network systems, a type of machine learning, can be used to develop highly nonlinear prediction models

What This Article Tells Us That Is New

- The authors' neural network model was comparable in accuracy to, but potentially more efficient at feature selection than logistic regression models
- Deep neural network–based machine learning provides an alternative to conventional multivariate regression

This article is featured in "This Month in Anesthesiology," page 1A. Corresponding articles on pages 619, 663, and 675. Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are available in both the HTML and PDF versions of this article. Links to the digital files are provided in the HTML text of this article on the Journal's Web site (www.anesthesiology.org). This article has an audio podcast. This article has a visual abstract available in the online version. Part of this work was presented at the Society for Technology in Anesthesia Annual Meeting 2017 and received the best of show award for the best abstract presentation of the meeting.

Submitted for publication July 15, 2017. Accepted for publication February 5, 2018. From the Department of Anesthesiology and Perioperative Care (C.K.L., M.C.), Department of Computer Sciences (C.K.L., P.B.), and Department of Bioengineering (M.C.), University of California Irvine, Irvine, California; and Department of Anesthesiology and Perioperative Medicine, University of California Los Angeles, Los Angeles, California (I.H., E.G., M.C.).

Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. All Rights Reserved. Anesthesiology 2018; 129:649-62

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.
<zdoi;10.1097/ALN.0000000000002186>

blood loss, (2) lowest mean arterial pressure, and (3) lowest heart rate during surgery to predict major postoperative complications.[5] Favored for its simplicity, the Surgical Apgar score presents with area under the receiver operating characteristics curve ranging from 0.6 to 0.8 for major complications or death with a correlation varying with subspecialty.[6–9] In addition, the Surgical Apgar score has been shown to not substantially improve mortality risk stratification when combined with preoperative scores.[9] In response to these limitations, there has been work to create more objective and accurate scores. The most popular method used to develop new scoring systems is based on logistic regression, such as the Preoperative Score to Predict Postoperative Mortality.[10] In order to make these scores accessible in clinical practice, the logistic regression coefficients are normalized to easily summed values to be interpreted as a score rather than the direct logistic regression output. Besides the aforementioned clinical risk scores, other recent perioperative administrative risk scores are the Risk Stratification Index (published initially in 2010[11] and validated in 2017 on nearly 40 million patients[12]) and the Risk Quantification Index.[13]

In recent years, and although they are not new,[11] neural networks and deep neural networks, known as "deep learning," have been used to tackle a variety of problems, ranging from computer vision,[12–17] gaming,[18–20] high-energy physics,[21,22] chemistry,[23–25] and biology.[26–28] While there have been studies using other machine-learning methods for clinical applications such as predicting cardiorespiratory instability[29,30] and 30-day readmission,[31,32] the use of deep neural networks in medicine is relatively limited.[33–36]

In this manuscript, we present the development and validation of a deep neural network model based upon intraoperative clinical features, to predict postoperative in-hospital mortality in patients undergoing surgery under general anesthesia. Its performance is presented together with other published clinical risk scores and administrative risk scores, as well as a logistic regression model using the same intraoperative features as the deep neural network. The deep neural networks were also assessed for leveraging preoperative information by the addition of ASA score and Preoperative Score to Predict Postoperative Mortality as features.

Materials and Methods

This manuscript follows the "Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View."[37]

Electronic Medical Record Data Extraction

All data for this study were extracted from the Perioperative Data Warehouse, a custom-built robust data warehouse containing all patients who have undergone surgery at University of California Los Angeles (Los Angeles, California) since the implementation of the electronic medical record (EPIC Systems, USA) on March 17, 2013. The construction of the Perioperative Data Warehouse has been previously described.[38] Briefly, the Perioperative Data Warehouse has a two-stage design. In the first stage, data are extracted from EPIC's Clarity database into 26 tables organized around three distinct concepts: patients, surgical procedures, and health system encounters. These data are then used to populate a series of 800 distinct measures and metrics such as procedure duration, readmissions, admission International Classification of Diseases (ICD) codes, and others. All data used for this study were obtained from this data warehouse, and institutional review board approval (No. 15-000518) has been obtained for this retrospective review.

A list of all surgical cases performed between March 17, 2013, and July 16, 2016, were extracted from the Perioperative Data Warehouse. The University of California Los Angeles Health System includes two inpatient medical centers and three ambulatory surgical centers; however, only cases performed in one of the two inpatient hospitals (including operating room and "off-site" locations) under general anesthesia were included in this analysis. Cases on patients younger than 18 yr of age or older than 89 yr of age were excluded. In the event that more than one procedure was performed during a given health system encounter, only the first case was included.

Model Endpoint Definition

The occurrence of an in-hospital mortality was extracted as a binary event (0, 1) based upon either the presence of a "mortality date" in the electronic medical record between surgery time and discharge or a discharge disposition of expired combined with a note associated with the death (*i.e.*, death summary, death note). The definition of in-hospital mortality was independent of length of stay in the hospital.

Model Input Features

Each surgical record corresponded to a unique hospital admission and contained 87 features calculated or extracted at the end of surgery (table 1). These features were considered to be potentially predictive of in-hospital mortality by clinicians' consensus (I.H., M.C., E.G.) and included descriptive intraoperative vital signs, such as minimum and maximum blood pressure values; summary of drug and fluid interventions, such as total blood infused and total vasopressin administered; and patient anesthesia descriptions, such as presence of an arterial line and type of anesthesia (all features are detailed in table 1).

Data Preprocessing

Before model development, missing values were filled with the mean value for the respective feature. In addition, to account for observations where the value is clinically out of range, values greater than a clinically normal maximum were set to a maximum possible value (table 1). These out-of-range values were due to the data artifact in the raw electronic medical record data. For example, a systolic blood

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

Table 1. Eighty-seven Features Used in Models with Description and Applied Maximum Possible Values as Defined by Domain Experts

Feature Name(s)	Description	No. Features; No. Features in Reduced Feature Set	Maximum Possible Absolute Value (If Applicable)
COLLOID_ML*	Total colloid transfused (ml)	1; 1	—
CRYSTALLOID_ML*	Total crystalloid transfused (ml)	1; 1	—
DBP MAX*, MIN*, AVG, MED, STD	Maximum, minimum, average, median, and SD diastolic blood pressure for the case (mmHg)	5; 2	150
DBP_10min MAX, MIN, AVG, MED, STD	Maximum, minimum, average, median, and SD diastolic blood pressure for the last 10 min of the case (mmHg)	5; 0	150
EBL*	Total estimated blood loss (ml)	1; 1	—
EPHEDRINE BOLUS*	Total bolus dose of ephedrine (mg) during the case	1; 1	—
EPINEPHRINE BOLUS*, END RATE*, MAX RATE*	Total bolus dose (mcg), end of case infusion rate (mcg · kg^{-1} · min^{-1}), and highest infusion rate (mcg · kg^{-1} · min^{-1}) of epinephrine during the case	3; 3	—
ESMOLOL BOLUS*, END RATE*, MAX RATE*	Total bolus dose (mg), end of case infusion rate (mcg · kg^{-1} · min^{-1}), and highest infusion rate (mcg · kg^{-1} · min^{-1}) of esmolol during the case	3; 3	—
HR MAX*, MIN*, AVG, MED, STD	Maximum, minimum, average, median, and SD heart rate (beats/min) for the case	5; 2	180
HR_10min MAX, MIN, AVG, MED, STD	Maximum, minimum, average, median, and SD heart rate (beats/min) for the last 10 min of the case	5; 0	180
INVASIVE_LINE_YN*	Invasive central venous, arterial, or pulmonary arterial line used for the case (Yes/No)	1; 1	—
MAP MAX*, MIN*, AVG, MED, STD	Maximum, minimum, average, median, and SD mean blood pressure (mmHg) for the case	5; 2	300
MAP_10min MAX, MIN, AVG, MED, STD	Maximum, minimum, average, median, and SD mean blood pressure (mmHg) for the last 10 min of the case	5; 0	300
DES MAX*	Maximum and minimum alveolar concentration of desflurane during the case (note: this is not age adjusted)	1; 1	12
GLUCOSE MAX*, MIN*	Maximum and minimum plasma glucose concentration for the case (mg/dl)	2; 2	400
ISO MAX*	Maximum and minimum alveolar concentration of isoflurane during the case (note: this is not age adjusted)	1; 1	12
SEVO MAX*	Maximum and minimum alveolar concentration of sevoflurane during the case (note: this is not age adjusted)	1; 1	10
MILRINONE END RATE*, MAX RATE*	End of case infusion rate and highest infusion rate of milrinone during the case (mcg · kg^{-1} · min^{-1})	2; 2	—
HGB MIN*	Minimum hemoglobin concentration (g/dl) during the case	1; 1	15
MINUTES MAP < 50	Cumulative min with mean arterial pressure < 50 mmHg (min)	1; 0	—
MINUTES MAP < 60	Cumulative min with mean arterial pressure < 60 mmHg (min)	1; 0	—
NICARDIPINE END RATE*, MAX RATE*	End of case infusion rate and highest infusion rate of nicardipine during the case (mg/h)	2; 2	—
NITRIC_OXIDE_YN*	Nitric oxide used for the case (Yes/No)	1; 1	—
NITROGLYCERIN BOLUS*, END RATE*, MAX RATE*	Total bolus dose (mcg), end of case infusion rate (mcg/min), and highest infusion rate (mcg/min) of nitroglycerin during the case	3; 3	—
NITROPRUSSIDE END RATE*, MAX RATE*	End of case infusion rate and highest infusion rate of nitroprusside (mcg · kg^{-1} · min^{-1}) during the case	2; 2	—
PHENYLEPHRINE BOLUS*, END RATE*, MAX RATE*	Total bolus dose (mcg), end of case infusion rate (mcg/min), and highest infusion rate (mcg/min) of phenylephrine during the case	3; 3	—
SBP MAX*, MIN*, AVG, MED, STD	Maximum, minimum, average, median, and SD systolic blood pressure (mmHg) for the case	5; 2	300
SBP_10min MAX, MIN, AVG, MED, STD	Maximum, minimum, average, median, and SD systolic blood pressure (mmHg) for the last 10 min of the case	5; 0	300
Spo_2 MAX*, MIN*, AVG, MED, STD	Maximum, minimum, average, median, and SD Spo_2 (%) for the case	5; 2	100
Spo_2_10min MAX, MIN, AVG, MED, STD	Maximum, minimum, average, median, and SD Spo_2 (%) for the last 10 min of the case	5; 0	100
UOP*	Total urine output (ml)	1; 1	—
VASOPRESSIN BOLUS*, END RATE*, MAX RATE*	Total bolus dose (units), end of case infusion rate (units/h), and highest infusion rate (units/h) of vasopressin during the case	3; 3	—
XFUSION_RBC_ML*	Total red blood cells transfused (ml)	1; 1	—
	Total number of features	87; 45	

*45 features used in the reduced feature set.

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

pressure of 400 mmHg is not clinically possible; however, it may be recognized as the maximum systolic blood pressure for the case during electronic medical record extraction. The data were then randomly divided into training (80%) and test (20%) data sets, with equal percent occurrence of in-hospital mortality. Training data were rescaled to have a mean of 0 and SD of 1 per feature. Test data were rescaled with the training data mean and SD.

Development of the Model

In this work, we were interested in classifying patients at risk of in-hospital mortality using deep neural networks, also referred to as deep learning. During development of deep neural networks, there are many unknown model parameters that need to be optimized by the deep neural network during training. These model parameters are first initialized and then optimized to decrease the error of the model's output to correctly classify in-hospital mortality. This error is referred to as a loss function. The type of deep neural network used in this study is a feedforward network with fully connected layers and a logistic output. "Fully connected" refers to the fact that all neurons between two adjacent layers are fully pairwise connected. A logistic output was chosen so that the output of the model could be interpreted as probability of in-hospital mortality (0 to 1). To develop a deep neural network, it is important to fine-tune the hyperparameters as well as the architecture. We utilized stochastic gradient descent with momentums (0.8, 0.85, 0.9, 0.95, 0.99) and initial learning rates (0.01, 0.1, 0.5), and a batch size of 200. We also assessed deep neural network architectures of one to five hidden layers with 10 to 300 neurons per layer, and rectified linear unit and hyperbolic tangent activation functions. The loss function was cross entropy. We utilized five-fold cross-validation with the training set (80%) to select the best hyperparameters and architecture based on mean cross-validation performance. These best hyperparameters and architecture were then used to train a model on the entire training set (80%) before testing final model performance on the separate test set (20%).

Overfitting. In addition, overfitting was a major concern in the development of our model. While approximately 50,000 patients is large for clinical data, it is small relative to data sets typically found in deep learning tasks such as vision and speech recognition, where millions of samples are available. Thus, regularization was critical. To address this, we utilized three methods: (1) early stopping, (2) L2 weight decay, and (3) dropout. Early stopping is the halting of model training when the loss of a separate early stopping validation set starts to increase compared to the training loss, indicating overfitting. This early stopping validation set was taken as a random 20% of the training set, and a patience of 10 epochs was utilized. L2 weight decay is a method of limiting the size of the weight of every parameter. The standard L2 weight penalty involves adding an extra term to the loss function that penalizes the squared weights, keeping the weights small unless the error derivative is big. We utilized an L2 weight penalty of 0.0001. Dropout is a method where neurons are removed from the network with a specified probability, to prevent coadapting of the neurons.[39–41] Dropout was applied to all layers with a probability of 0.5.

Data Augmentation. The goal of training was to optimize model parameters to decrease classification error of in-hospital mortality. However, the actual percent of occurrence of in-hospital mortality in the data was low and thus the data were skewed. The percent occurrence of mortality in the training data set was less than 1%. To help with this skewed distribution, training data were augmented by taking only the observations positive for in-hospital mortality and adding Gaussian noise. This was performed by adding a random number taken from a Gaussian distribution with a SD of 0.0001 to each feature's value. This essentially duplicated the in-hospital mortality observations with a slight perturbation. The in-hospital mortality observations in the training data set were augmented using this method to approximately 45% occurrence before training. During cross-validation, this meant that only training folds were augmented. The validation fold was not augmented.

Feature Reduction and Preoperative Feature Experiments

Experiments to assess the impact of (1) reducing the number of features from the clinician chosen 87 to 45 features, and (2) adding ASA score and Preoperative Score to Predict Postoperative Mortality as a feature, were also conducted. The reduced 45 feature set was created by excluding all "derived" features, specifically average, median, SD, and last 10 min of the surgical case features (table 1).

After choosing the best performing deep neural network architecture and hyperparameters with the complete 87 features data set, five additional deep neural networks were each trained with the following: (1) the addition of ASA score as a model feature (88 features); (2) the addition of Preoperative Score to Predict Postoperative Mortality as a model feature (88 features); (3) a reduced model feature set (45 features); (4) the addition of ASA score to the reduced feature set (46 features); and (5) the addition of Preoperative Score to Predict Postoperative Mortality to the reduced feature set (46 features).

Model Performance

All model performances were assessed on 20% of the data held out from training as a test set. Model performance was compared to ASA score, Surgical Apgar, Risk Quantification Index, Risk Stratification Index, Preoperative Score to Predict Postoperative Mortality, and a standard logistic regression model using the same combination of features as in the deep neural network. ASA score was extracted from the University of California Los Angeles preoperative assessment record. Surgical Apgar was calculated using Gawande *et al.*[5] Risk Quantification Index could not be calculated using the downloadable R package from Cleveland Clinic's Web site

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

(http://my.clevelandclinic.org/departments/anesthesiology/depts/outcomes-research; accessed October 16, 2017) due to technical issues with the R version, and so Risk Quantification Index log probability and score were calculated from equations provided in Sigakis *et al.*[42] Uncalibrated Risk Stratification Index was calculated using coefficients provided by the original authors (Supplemental Digital Content, http://links.lww.com/ALN/B681).[43] To calculate Risk Stratification Index, all International Classification of Diseases, Ninth Revision (ICD-9) diagnosis codes for each patient were matched with a Risk Stratification Index coefficient and the coefficients were then summed. Preoperative Score to Predict Postoperative Mortality scores were extracted from the Perioperative Data Warehouse, where they were calculated as described by Le Manach *et al.*[10] Each of the diseases described by Le Manach *et al.*[10] were extracted as a binary endpoint from the admission ICD codes for the relevant hospital admission. In addition to assigning points based on patient comorbidities, the Preoperative Score to Predict Postoperative Mortality also assigns points for the type of surgery performed. These points were assigned based on the primary surgical service for the given procedure.

Area under the Receiver Operating Characteristics Curves. Model performance was assessed using area under the receiver operating characteristics curve and 95% CIs for area under the receiver operating characteristics curve were calculated using bootstrapping with 1,000 samples.

Choosing a Threshold. The F1 score, sensitivity, and specificity were calculated for different thresholds for the deep neural network models, logistic regression model, ASA score, and Preoperative Score to Predict Postoperative Mortality. The F1 score is a measure of precision and recall, ranging from 0 to 1. It is calculated as $F1 = 2 * \frac{precision * recall}{precision + recall}$, where precision is (true positives/predicted true) and recall is equivalent to sensitivity. Two different threshold methods were assessed: (1) a threshold that optimized the observed in-hospital mortality rate, and (2) a threshold based on the highest F1 score. The number of true positives, true negatives, false positives, and false negatives were then assessed for each threshold to assess differences in the number of patients correctly predicted by each model.

Calibration. Calibration was performed to account for the use of data augmentation on the training data set to be used during training of the deep neural network. This data augmentation served to balance classes in the training data set to approximately 45% mortality *versus* the true distribution of mortality (less than 1%). This extreme augmentation of the training data set classes skewed predicted probabilities to be higher than the expected probability based on the true distribution of mortality (less than 1%). Therefore, we performed calibration after finalizing the model. Calibration was performed only on the test data set. Calibration of the deep neural network predicted probability output was performed using the following equation:

$$Calibrated\ Predicted\ Probability = \frac{1}{1 + \left(\frac{1}{Predicted\ Probability} - 1 \right) \frac{P(0)}{P(1)}}$$

where $P(1) = \frac{\#\ Observed\ Mortality\ in\ Test}{\#\ Test\ Patients} = \frac{87}{11997}$ and $P(0) = 1 - P(1)$. This calibration formula was used to maintain the rank of predicted probabilities, and thus not changing any model performance metrics (area under the receiver operating characteristics curve, sensitivity, specificity, or F1 score). Additionally, calibration plots and Brier scores were used to assess calibration of predictions.

Feature Importance. To assess which features are the most predictive in the deep neural network, we performed a feature ablation analysis. This analysis consisted of removing model features grouped by type of clinical feature, and then retraining a deep neural network with the same final architecture, as well as hyperparameters on the remaining features. The change in area under the receiver operating characteristics curve with the removal of each feature was then assessed to evaluate the importance of each group of features. To assess which features are the most predictive in the logistic regression model, we assessed which features corresponded to the largest weights.

All deep neural network models were developed and applied using Keras.[44] Logistic regression models and performance metrics were calculated with scikit-learn.[45]

Results

Patient Characteristics

The data consisted of 59,985 surgical records. Patient demographics and characteristics of the training and test data sets are summarized in table 2. The in-hospital mortality rate of both the training and test set is less than 1%. The presence of invasive lines is also similar for both sets (26.5% in training; 26.7% in test). The most prevalent ASA score is III at 49.9% for both sets.

Development of the Model

The final deep neural network architecture consists of four hidden layers of 300 neurons per layer with rectified linear unit activations and a logistic output (fig. 1). The deep neural network was trained with dropout probability of 0.5 between all layers, L2 weight decay of 0.0001, and a learning rate of 0.01 and momentum of 0.9.

Model Performance

All performance metrics reported below refer to the test data set (n = 11,997).

Area under the Receiver Operating Characteristics Curves. Receiver operating characteristics curves and area under the receiver operating characteristics curve results are shown in figure 2 and table 3. All logistic regression models and all

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

Table 2. Training and Test Data Set Patient Characteristics Reported as Number of Patients (%) or Mean ± SD

Characteristic	Train	Test
No. of patients	47,988	11,997
No. of patients with in-hospital mortality (%)	389 (0.81%)	87 (0.73%)
Age (yr)	56 ± 17	56 ± 18
Estimated blood loss (ml)	95 ± 540	94 ± 410
Presence of arterial line (%)	8,585 (17.9%)	2,135 (18.0%)
Presence of pulmonary artery line (%)	1,641 (3.4%)	430 (3.6%)
Presence of central line (%)	2,444 (5.1%)	635 (5.3%)
ASA score (%)		
I	3,023 (6.3%)	762 (6.4%)
II	17,930 (37.4%)	4,477 (37.3%)
III	23,960 (49.9%)	5,986 (49.9%)
IV	2,911 (6.1%)	735 (6.1%)
V	144 (0.3%)	30 (0.3%)
VI	4 (0.01%)	0 (0%)
HCUP code description (%)		
Upper gastrointestinal, endoscopy, biopsy	3,864 (8.05%)	965 (8%)
Colonoscopy and biopsy	1,693 (3.53%)	388 (3.2%)
Laminectomy, excision intervertebral disc	1,029 (2.14%)	287 (2.4%)
Other therapeutic procedures, hemic, and lymphatic system	1,013 (2.11%)	247 (2.1%)
Other or therapeutic procedures on respiratory system	985 (2.05%)	254 (2.1%)
Incision and excision of central nervous system	942 (1.96%)	255 (2.1%)
Other or procedures on vessels other than head and neck	932 (1.94%)	207 (1.7%)
Other therapeutic endocrine procedures	904 (1.88%)	258 (2.2%)
Hip replacement, total and partial	792 (1.65%)	186 (1.6%)
Arthroplasty knee	768 (1.6%)	193 (1.6%)
Other or therapeutic nervous system procedures	750 (1.56%)	181 (1.5%)
Thyroidectomy, partial or complete	737 (1.54%)	172 (1.4%)
Spinal fusion	735 (1.53%)	150 (1.3%)
Other or therapeutic procedures on bone	722 (1.5%)	195 (1.6%)
Conversion of cardiac rhythm	720 (1.5%)	184 (1.5%)
Heart valve procedures	715 (1.49%)	186 (1.6%)
Cholecystectomy and common duct exploration	700 (1.46%)	216 (1.8%)
Endoscopic retrograde cannulation of pancreas	663 (1.38%)	155 (1.3%)
Kidney transplant	659 (1.37%)	194 (1.6%)
Other or therapeutic procedures on nose, mouth, and pharynx	653 (1.36%)	173 (1.4%)
Other hernia repair	652 (1.36%)	178 (1.5%)
Hysterectomy, abdominal and vaginal	641 (1.34%)	155 (1.3%)
Appendectomy	634 (1.32%)	147 (1.2%)
Other therapeutic procedures on muscles and tendons	629 (1.31%)	154 (1.3%)
Colorectal resection	609 (1.27%)	127 (1.1%)
Insertion, revision, replacement, removal of cardiac pacemaker or cardioverter/defibrillator	601 (1.25%)	128 (1.1%)
Abortion (termination of pregnancy)	587 (1.22%)	162 (1.4%)
Treatment, fracture, or dislocation of hip and femur	570 (1.19%)	155 (1.3%)
Other or gastrointestinal therapeutic procedures	569 (1.19%)	124 (1%)
Open prostatectomy	554 (1.15%)	140 (1.2%)
Diagnostic bronchoscopy and biopsy of bronchus	550 (1.15%)	131 (1.1%)
Nephrectomy, partial or complete	526 (1.1%)	124 (1%)

HCUP code description and distribution is shown only for those representing more than 1% of the train data set.
ASA, American Society of Anesthesiologists; HCUP, Healthcare Cost and Utilization Project.

deep neural networks had higher area under the receiver operating characteristics curves than Preoperative Score to Predict Postoperative Mortality (0.74 [95% CI, 0.68 to 0.79]) and Surgical Apgar (0.58 [95% CI, 0.52 to 0.64]) for predicting in-hospital mortality (fig. 2, table 3). All deep neural networks had higher area under the receiver operating characteristics curves than logistic regressions for each combination of features except for the reduced feature set with Preoperative Score to Predict Postoperative Mortality (logistic regression 0.90 [95% CI, 0.86 to 0.93] *vs.* deep neural network 0.90 [95% CI, 0.87 to 0.93]). In addition, reducing the feature set from 87 to 45 features did not reduce the

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

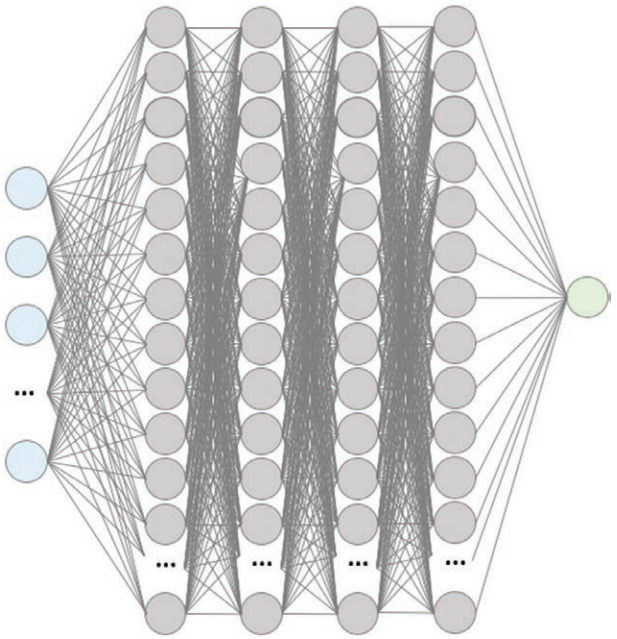

Fig. 1. Summary visualization of the deep neural network. Input layer (*blue*) of features feed into the first hidden layer of 300 neurons with rectified linear unit activations (*grey*). All the activations of neurons in the first hidden layer are fed into each of the neurons in the second, then all the second are fed into the third, and finally, all the third are fed into the fourth. All the activations of the neurons in the fourth hidden layer are then fed into a logistic output layer to produce a probability for in-hospital mortality between 0 and 1.

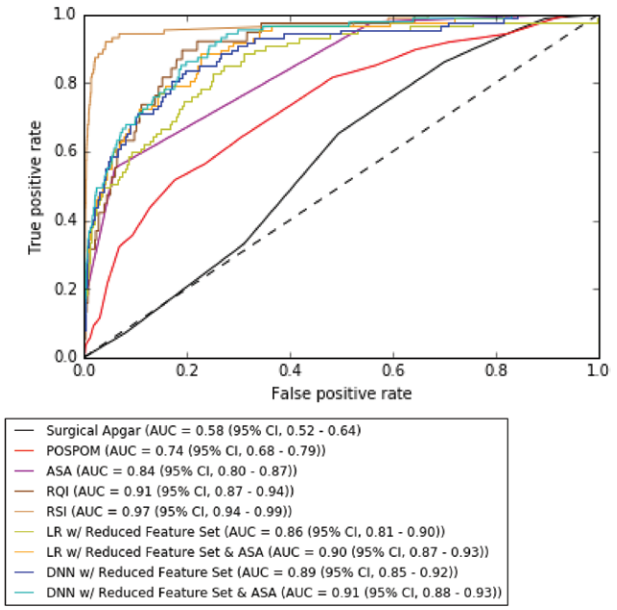

Fig. 2. Receiver operating characteristic curves to predict postoperative in-hospital mortality.

deep neural network model area under the receiver operating characteristics curve performance, and the addition of ASA score and Preoperative Score to Predict Postoperative Mortality as features modestly improved the area under the receiver operating characteristics curves of both the full and reduced feature set deep neural network models. The highest deep neural network area under the receiver operating characteristics curve result was the deep neural network with reduced feature set and ASA score (0.91 [95% CI, 0.88 to 0.93]). The highest risk score area under the receiver operating characteristics curve was Risk Stratification Index (0.97 [95% CI, 0.94 to 0.99]), and the highest logistic regression area under the receiver operating characteristics curves were the logistic regression with reduced feature set and ASA score (0.90 [95% CI, 0.87 to 0.93]), and the logistic regression with reduced feature set and Preoperative Score to Predict Postoperative Mortality (0.90 [95% CI, 0.86 to 0.93]).

Table 3. Area under the Receiver Operator Characteristic Curve Results with 95% CIs for the Test Set (N = 11,997)

	AUC (95% CI)
Clinical risk score	
Surgical Apgar	0.58 (0.52–0.64)
POSPOM	0.74 (0.68–0.79)
ASA score	0.84 (0.80–0.87)
RQI*	0.91 (0.87–0.94)
RSI uncalibrated†	0.97 (0.94–0.99)
Model	
Logistic regression	
With all 87 features	0.86 (0.81–0.90)
With ASA score (88 features)	0.89 (0.86–0.92)
With POSPOM (88 features)	0.89 (0.85–0.93)
With reduced feature set (45 features)	0.86 (0.81–0.90)
With reduced feature set and ASA score (46 features)	0.90 (0.87–0.93)
With reduced feature set and POSPOM (46 features)	0.90 (0.86–0.93)
DNN	
With all 87 features	0.88 (0.85–0.91)
With ASA score (88 features)	0.90 (0.87–0.93)
With POSPOM (88 features)	0.91 (0.87–0.95)
With reduced feature set (45 features)	0.89 (0.85–0.92)
With reduced feature set and ASA score (46 features)	0.91 (0.88–0.93)
With reduced feature set and POSPOM (46 features)	0.90 (0.87–0.93)

*RQI was calculated on 5,591 test patients (38 mortality). †RSI uncalibrated was calculated on 11,939 test patients (86 mortality).

AUC, area under the receiver operating characteristic curve; ASA, American Society of Anesthesiologists; DNN, deep neural network; POSPOM, Preoperative Score to Predict Postoperative Mortality; RQI, Risk Quantification Index; RSI, Risk Stratification Index.

Choosing a Threshold. For comparison of F1 scores, sensitivity and specificity at different thresholds, deep neural network with original 87 features (DNN), deep neural network with a reduced feature set and Preoperative Score to Predict Postoperative Mortality ($DNN_{rfsPOSPOM}$), and deep neural network with a reduced feature set and ASA score (DNN_{rfsASA}) are compared to ASA score, Preoperative Score to Predict Postoperative Mortality, logistic regression with original 87 features, logistic regression with a reduced feature set and Preoperative Score to Predict Postoperative Mortality ($LR_{rfsPOSPOM}$), and logistic regression with a reduced feature set and ASA

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

score (LR_{rfsASA}; table 4). To compare the number of correctly predicted patients by the deep neural networks at different thresholds, a table of the number of correctly and incorrectly classified patients is shown for all models at different thresholds for all test patients (n = 11,997; table 5).

If we choose a threshold that optimizes the observed in-hospital mortality rate, the thresholds (% observed mortality) for Preoperative Score to Predict Postoperative Mortality, ASA score, and logistic regression, $LR_{rfsPOSPOM}$, and LR_{rfsASA} are 10 (93.1%), 3 (97.7%), 0.00015 (98.9%), 0.002 (97.7%), and 0.0034 (96.66%), respectively (table 4). The thresholds for deep neural network, $DNN_{rfsPOSPOM}$, and DNN_{rfsASA} are 0.05 (98.9%), 0.2 (96.6%), and 0.22 (96.6%), respectively. At these thresholds, Preoperative Score to Predict Postoperative Mortality, ASA score, logistic regression, $LR_{rfsPOSPOM}$, LR_{rfsASA}, deep neural network, $DNN_{rfsPOSPOM}$, and DNN_{rfsASA} all have high and comparable sensitivities. The deep neural network with the highest area under the receiver operating characteristics curve, DNN_{rfsASA}, had a sensitivity of 0.97 (95% CI, 0.92 to 1) and specificity of 0.64 (95% CI, 0.64 to 0.65), and the logistic regression with the highest area under the receiver operating characteristics curve, LR_{rfsASA}, had a sensitivity of 0.97 (95% CI, 0.92 to 1) and specificity of 0.64 (95% CI, 0.63 to 0.65). However, all deep neural networks reduced false positives while maintaining the same or similar number of false negatives (table 5). The deep neural network with all 87 original features decreased the number of false positives compared to logistic regression, from 11,873 to 9,169 patients. DNN_{rfsASA} decreased the number of false positives compared to LR_{rfsASA}, from 4,332 patients to 4,241 patients; when compared to Preoperative Score to Predict Postoperative Mortality and ASA score, from 9,169 patients and 6,666 patients, respectively.

If we choose a threshold that optimizes precision and recall *via* the F1 score, the thresholds for Preoperative Score to Predict Postoperative Mortality, ASA score, logistic regression, $LR_{rfsPOSPOM}$, and LR_{rfsASA} are higher at 20, 5, 0,1, 0.1, and 0.1, respectively (table 4). All the thresholds for deep neural network, $DNN_{rfsPOSPOM}$, and DNN_{rfsASA} also increased to 0.3, 0.4, and 0.3, respectively. The highest F1 scores were comparable for ASA score, LR_{rfsASA}, and DNN_{rfsASA} at 0.24 (95% CI, 0.14 to 0.35), 0.26 (95% CI, 0.18 to 0.33), and 0.22 (95% CI, 0.12 to 0.30). However, DNN_{rfsASA} had a lower number of false positives at 35 patients, compared to LR_{rfsASA} at 115 patients (table 5).

Calibration. For comparison of calibration, Brier scores and calibration plots were assessed for logistic regression, DNN_{rfsASA}, and calibrated DNN_{rfsASA}. DNN_{rfsASA} had the worst Brier score of 0.0352, and logistic regression had the best score of 0.0065 (fig. 3). However, the calibrated DNN_{rfsASA} had a comparable Brier score of 0.0071. Calibration of DNN_{rfsASA} shifted the best thresholds for observed mortality optimization and F1 optimization from 0.2 and 0.4 to 0.0018 and 0.0048, respectively.

Feature Importance. To assess feature importance in the deep neural network, we assessed the decrease in area under

Table 4. Percentage of Observed Mortality Patients Correctly Identified, F1 Score, Sensitivity, and Specificity Performance of ASA Score; POSPOM; Logistic Regression Model and DNN Model with 87 Features; Logistic Regression Model and DNN Model with Reduced Feature Set and ASA Score; and Logistic Regression Model and DNN Model with Reduced Feature Set and POSPOM at Different Thresholds

	Threshold	No. Observed Mortality (%)	F1 (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
ASA score	3	85 (97.7%)	0.02 (0.02–0.03)	0.98 (0.94–1)	0.44 (0.43–0.45)
	5	14 (16.1%)	0.24 (0.14–0.35)	0.16 (0.09–0.25)	1 (1–1)
POSPOM	10	81 (93.1%)	0.02 (0.01–0.02)	0.93 (0.87–0.98)	0.23 (0.22–0.24)
	20	31 (35.6%)	0.05 (0.03–0.07)	0.36 (0.25–0.47)	0.91 (0.90–0.91)
Logistic regression with 87 features	0.00015	86 (98.9%)	0.01 (0.01–0.02)	0.99 (0.96–1)	0.003 (0.002–0.004)
	0.1	28 (32.2%)	0.24 (0.16–0.30)	0.32 (0.22–0.42)	0.99 (0.99–0.99)
DNN with 87 features	0.05	86 (98.9%)	0.02 (0.01–0.02)	0.99 (0.96–1)	0.20 (0.20–0.21)
	0.3	35 (40.2%)	0.23 (0.17–0.30)	0.40 (0.30–0.51)	0.99 (0.98–0.99)
Logistic regression with reduced feature set and ASA score	0.0034	84 (96.6%)	0.04 (0.03–0.05)	0.97 (0.92–1)	0.64 (0.63–0.65)
	0.1	30 (34.5%)	0.26 (0.18–0.33)	0.34 (0.24–0.44)	0.99 (0.99–0.99)
DNN with reduced feature set and ASA score	0.22	84 (96.6%)	0.04 (0.03–0.05)	0.97 (0.92–1)	0.64 (0.64–0.65)
	0.4	15 (17.2%)	0.22 (0.12–0.30)	0.17 (0.09–0.25)	1 (1–1)
Logistic regression with reduced feature set and POSPOM	85 (97.7%)	0.03 (0.02–0.03)	0.98 (0.94–1)	0.48 (0.48–0.49)	85 (97.7%)
	26 (29.9%)	0.22 (0.15–0.29)	0.30 (0.20–0.39)	0.99 (0.99–0.99)	26 (29.9%)
DNN with reduced feature set and POSPOM	0.2	84 (96.6%)	0.04 (0.03–0.04)	0.97 (0.92–1)	0.63 (0.63–0.64)
	0.3	40 (46%)	0.18 (0.13–0.22)	0.46 (0.36–0.56)	0.97 (0.97–0.98)

Results for best thresholds chosen by (1) highest percent of observed mortality and (2) highest F1 score.

ASA, American Society of Anesthesiologists; DNN, deep neural network; POSPOM, Preoperative Score to Predict Postoperative Mortality.

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

Table 5. Number of Correctly and Incorrectly Classified Patients for ASA Score; POSPOM; Logistic Regression Model and DNN Model with 87 Features; Logistic Regression Model and DNN Model with Reduced Feature Set and ASA; and Logistic Regression Model and DNN Model with Reduced Feature Set and POSPOM at Different Thresholds

	Threshold	No. True Negative	No. False Positive	No. False Negative	No. True Positive
ASA score	3	5,244	6,666	2	85
	5	11,894	16	73	14
POSPOM	10	2,741	9,169	6	81
	20	10,782	1,128	56	31
Logistic regression with 87 features	0.00015	37	11,873	1	86
	0.1	11,788	122	59	28
DNN with 87 features	0.05	2,414	9,496	1	86
	0.3	11,734	176	52	35
Logistic regression with reduced feature set and ASA score	0.0034	7,578	4,332	3	84
	0.1	11,795	115	57	30
DNN with Reduced Feature Set and ASA	0.22	7,669	4,241	3	84
	0.4	11,875	35	72	15
Logistic regression with reduced feature set and POSPOM	0.002	5,772	6,138	2	85
	0.1	11,790	120	61	26
DNN with reduced feature set and POSPOM	0.2	7,550	4,360	3	84
	0.4	11,897	12	82	5

Results for best thresholds chosen by (1) highest percent of observed mortality and (2) highest F1 score.

ASA, American Society of Anesthesiologists; DNN, deep neural network; POSPOM, Preoperative Score to Predict Postoperative Mortality.

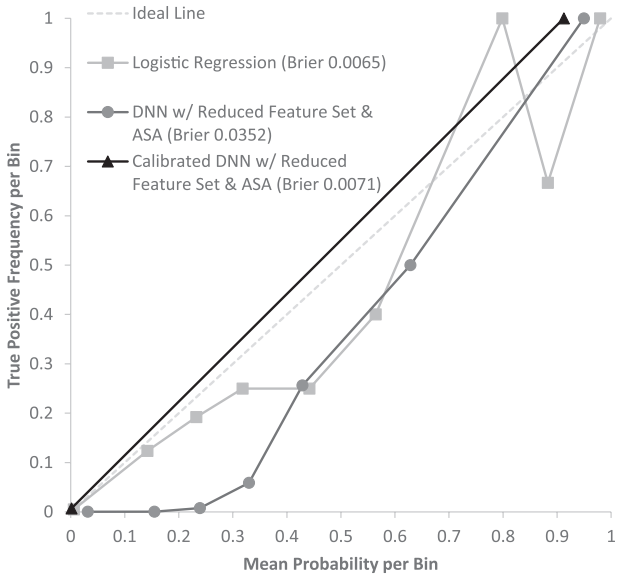

Fig. 3. Calibration plot with mean predicted probability *versus* true positive frequency (number of true positives/number of samples) per probability value bins in the test data set (N = 11,997) for logistic regression, deep neural network (DNN) with reduced feature set and American Society of Anesthesiologists (ASA) score, and calibrated DNN with reduced feature set and ASA score. Bins of predicted probability were at intervals of 0.1: (0 to 0.1), (0.1 to 0.2)…(0.9 to 1.0).

the receiver operating characteristics curve for the removal of groups of features from the best deep neural network (DNN_{rfsasa}; table 6; fig. 4). For the analysis, 13 groups were used (age, anesthesia, ASA score, input, blood pressure, output, vasopressor, vasodilator, labs, heart rate, invasive line, inotrope, and pulse oximetry). To assess feature importance, we assessed the weights for the logistic regression model (LR_{rfsASA}; fig. 5). The top five deep neural network features groups were: labs, ASA score, anesthesia, blood pressure, and vasopressor administration. The top logistic regression feature was ASA score. In addition, similar to the deep neural network, vasopressin administration, hemoglobin, presence of arterial or pulmonary arterial line, and sevoflurane administration are found in the top 10 weights.

We have developed a Web site application that performs predictions for DNN_{rfsASA} and DNN_{rfs} on a given data set. The application, as well as downloadable model package, are available at http://risknet.ics.uci.edu.

Discussion

The results in this study demonstrate that deep neural networks can be utilized to predict in-hospital mortality based on automatically extractable and objective intraoperative data. In addition, these predictions are further improved *via* the addition of preoperative information, as summarized in a patient's ASA score or Preoperative Score to Predict Postoperative Mortality. The area under the receiver operating characteristics curve of the "best" deep neural network model with a reduced feature set and ASA score (DNN_{rfsASA}) also outperformed Surgical Apgar, Preoperative Score to Predict Postoperative Mortality, and ASA score. Optimizing thresholds to capture the most observed mortality patients, in other words optimizing for sensitivity, DNN_{rfsASA} has higher sensitivity

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

Table 6. Features Removed with Each Group during Each Step of the Feature Ablation Analysis for the DNN

Group Type	Feature Removed in Group
Age	AGE
Anesthesia	MAX_ISO
Anesthesia	MAX_SEVO
Anesthesia	MAX_DES
Anesthesia	NITRIC_OXIDE_YN
ASA	ASA_SCORE
Blood pressure	MAX_DBP
Blood pressure	MIN_DBP
Blood pressure	MAX_MAP
Blood pressure	MIN_MAP
Blood pressure	MAX_SBP
Blood pressure	MIN_SBP
Heart rate	MAX_HR
Heart rate	MIN_HR
Inotrope	MILRINONE_CURRENT_RATE_MCG_KG_MIN
Inotrope	MILRINONE_MAX_RATE_MCG_KG_MIN
Input	XFUSION_RBC_ML
Input	COLLOID_ML
Input	CRYSTALLOID_ML
Invasive line	CVC_ANES_YN
Invasive line	PA_LINE_YN
Invasive line	ART_LINE_YN
Labs	BASELINE_GFR
Labs	MAX_GLUCOSE
Labs	MIN_GLUCOSE
Labs	MIN_HB
Labs	CURRENT_HB
Labs	STARTING_HB
Output	EBL
Output	UOP
Pulse oximetry	MAX_PULSE_OX
Pulse oximetry	MIN_PULSE_OX
Vasodilator	ESMOLOL_CURRENT_RATE_MCG_KG_MIN
Vasodilator	ESMOLOL_MAX_RATE_MCG_KG_MIN
Vasodilator	NICARDIPINE_CURRENT_RATE_MG_HR
Vasodilator	NICARDIPINE_MAX_RATE_MG_HR
Vasodilator	NITROGLYCERIN_CURRENT_RATE_MCG_MIN
Vasodilator	NITROGLYCERIN_MAX_RATE_MCG_MIN
Vasodilator	NITROPRUSSIDE_CURRENT_RATE_MCG_KG_MIN
Vasodilator	NITROPRUSSIDE_MAX_RATE_MCG_KG_MIN
Vasopressor	EPINEPHRINE_CURRENT_RATE_MCG_KG_MIN
Vasopressor	EPINEPHRINE_MAX_RATE_MCG_KG_MIN
Vasopressor	PHENYLEPHRINE_CURRENT_RATE_MCG_MIN
Vasopressor	PHENYLEPHRINE_MAX_RATE_MCG_MIN
Vasopressor	VASO_CURRENT_RATE_UNITS_HR
Vasopressor	VASOPRESSIN_MAX_RATE_UNITS_HR

Feature names are defined in table 1.
ASA, American Society of Anesthesiologists.

than Preoperative Score to Predict Postoperative Mortality, but comparable to ASA score, LR_{rfsASA}, and $LR_{rfsPOSPOM}$. This may make sense as ASA score is a feature in this deep neural network model. Most notably, however, is that DNN_{rfsASA} reduces the number of false positives compared to Preoperative Score to Predict Postoperative Mortality and ASA score by 54% and 36%, respectively. DNN_{rfsASA} also reduced the number of false positives to the most comparably performing logistic regression model LR_{rfsASA} by 2%. In addition, it should be noted that for each feature set combination (all 87 features, 87 features with ASA score, 87 features with Preoperative Score to Predict Postoperative Mortality, reduced features, reduced features with ASA score, and reduced features with Preoperative Score to Predict Postoperative Mortality), the deep neural network slightly outperformed logistic regression, with the exception of the reduced feature set with Preoperative Score to Predict Postoperative Mortality. However, the addition of Preoperative Score to Predict Postoperative Mortality is adding a logistic regression model output as a feature to another logistic regression model, which can be thought of as adding one hidden layer to a neural network with a logistic output. While the area under the receiver operating characteristics curve of logistic regression with the same reduced feature set and ASA score (LR_{rfsASA}) was not significantly lower than DNN_{rfsASA}, the deep neural network with all 87 original features outperformed logistic regression with the same 87 features in area under the receiver operating characteristics curve and significantly decreased the number of false positives by 2,377 patients (20%). This suggests that without careful feature selection to reduce the number of features, as well adding preoperative information, logistic regression did not perform comparably to a deep neural network. Logistic regression can be thought of as a neural network with no hidden layers. When preserving complexity, such as not performing careful feature selection or more rigorous preprocessing, neural networks with many hidden layers are able to perform well and in some cases better than logistic regression.

Due to such a low incidence of true positives (n = 87), the numbers for false negatives are hard to compare in this very small mortality population. This small number of mortality patients also affects the interpretation of the calibration results. Extensive data augmentation was used in training the deep neural network on balanced classes, resulting in predicted probabilities that were shifted up. The deep neural network's predicted probability was calibrated to the expected probability of mortality (less than 1%), and all predicted probabilities were then shifted down significantly less than 0.01 to reflect the % occurrence of in-hospital mortality, while maintaining all performance metrics. After calibration, the calibrated DNN_{rfsASA} resulted in a better Brier score that was also closer to that of logistic regression, and the optimal mortality threshold for DNN_{rfsASA} was shifted down from 0.2 to 0.0018, a more reasonable threshold considering the low percent occurrence of mortality. For direct comparison in the calibration plot, the same probability bins at intervals of 0.1 were chosen for the DNN_{rfsASA} calibrated and uncalibrated as well as logistic regression. A limitation of the calibration plot is that it is highly dependent on the choice of bins. This limitation is reflected in the resulting calibration plot for the calibrated DNN_{rfsASA}, where 86 mortality patients were predicted in the bin (0 to 0.1) and one

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

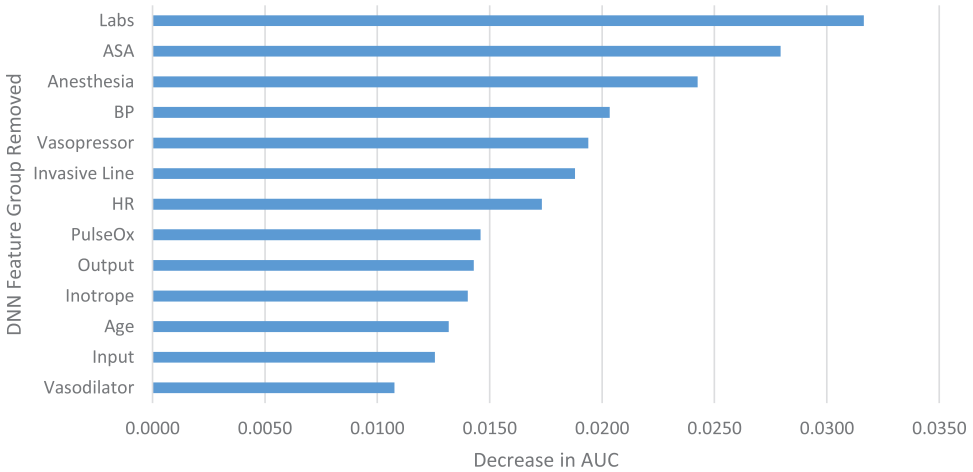

Fig. 4. Decrease in area under the receiver operating characteristic curve (AUC) performance for each feature group removed during feature ablation analysis for deep neural network with reduced feature set and American Society of Anesthesiologists (ASA) score. BP, blood pressure; DNN, deep neural network; HR, heart rate.

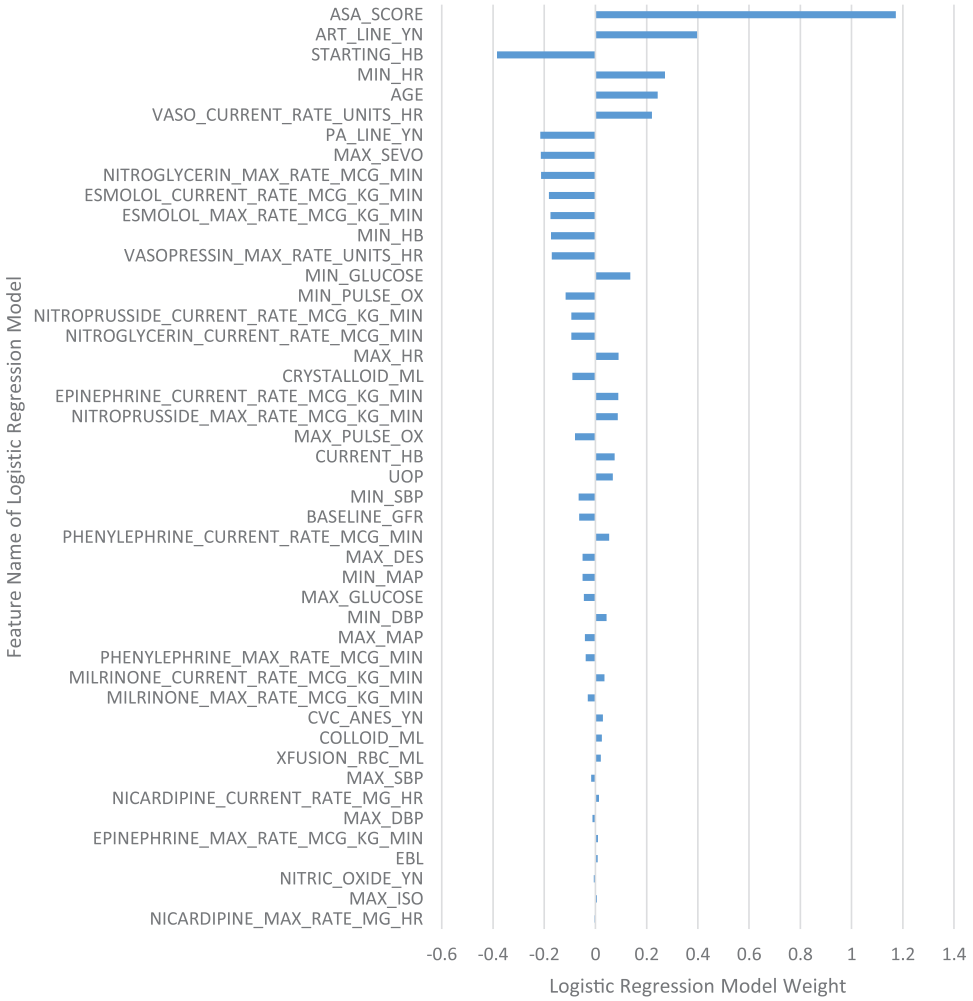

Fig. 5. Logistic regression model weight assigned to each feature in the logistic regression model with reduced feature set and American Society of Anesthesiologists (ASA) score.

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

patient was predicted in the bin (0.9 to 1). Thus, the interpretation of these results is limited to the number of true positives that exist.

While the Risk Quantification Index had a high and comparable area under the receiver operating characteristics curve to the DNN_{rfsASA}, it could only be calculated on 47% of the test patients due to a feature of Risk Quantification Index, specifically the Procedural Severity Score, which was available for only a limited number of Current Procedural Terminology codes. The Risk Stratification Index had the highest area under the receiver operating characteristics curve at 0.97 and, unlike Risk Quantification Index, could be calculated on a vast majority of the patients. Risk Stratification Index requires ICD-9 procedural and diagnosis codes. There are important distinctions to be made between a risk score based on clinical data (ASA score, Surgical Apgar, Preoperative Score to Predict Postoperative Mortality, and the logistic regression and deep neural network models reported here) *versus* administrative data (Risk Stratification Index, Risk Quantification Index). The first is that present-on-admission diagnoses and planned procedures (*i.e.*, ICD-9 and ICD-10 codes) are theoretically available preoperatively. But in practice, the coding is done after discharge, and therefore is not actually available preoperatively to guide clinical care. This makes scores, such as the Risk Stratification Index, appropriate for its intended purpose—comparing hospitals—but not for individual patient care. Finally, point-of-care clinical data contain more information about specific patients than models based only on diagnoses and procedure codes, and therefore should be more specific and useful for guiding the care of individual patients. These distinctions should not be seen as "one is better than another," so much as a matter of selecting the right model for particular purposes.

Perhaps the most attractive feature of this mortality model is that it provides a fully automated and highly accurate way to estimate the mortality risk of the patient at the end of surgery. All data contained in the risk score are easily obtained from the electronic medical record and could be automatically loaded into a model. While the ASA score is subjective, presents with high inter- and intrarater variability, and does require input from the anesthesiologist into the electronic medical record, this input is common practice as a part of preoperative assessment. In addition, we have also trained a deep neural network model using the Preoperative Score to Predict Postoperative Mortality score with comparable performance metrics. Thus, if the clinical need is to be completely objective, the $DNN_{rfsPOSPOM}$ model would be the most automatic and objective, as Preoperative Score to Predict Postoperative Mortality is based on the presence of key patient comorbidities and could be automatically obtained from the electronic medical record.

The input into this mortality model is based heavily on intraoperative data available at the end of surgery. There are 45 intraoperative features in the reduced feature set and one preoperative feature was added accordingly to leverage preoperative information. The ability of the intraoperative-only mortality models (deep neural network and deep neural network with reduced feature set) to maintain high performance with no addition of preoperative features further supports the idea that intraoperative events and management may have a significant effect on postoperative outcomes.

By definition, any screening score will have to trade off between sensitivity (capturing all patients with the condition) and specificity (not capturing those who do not have the condition). As a result, clinically, we generally discuss the number needed to treat—the number of "false positives" that must be treated to capture one true positive. Our deep neural network model not only had the highest area under the receiver operating characteristics curve, but also reduced the number of false positives, thereby reducing the number needed to treat. Given the current transitions toward value-based care, this has some appeal.

Another key advantage of a deep neural network model is its ability to account for the relationships between various clinical factors. For example, in a logistic regression model, excess estimated blood loss might be assigned a certain weight and hypotension a different one, thus assigning a linear relationship between hypotension and blood loss. On the other hand, a deep neural network model could account for the differences and linear or nonlinear associations of hypotension in a minimal blood loss *versus* significant blood loss case. While a feature could be created to reflect this relationship of hypotension and blood loss and used as an input into a logistic regression model, a deep neural network model avoids this need for careful feature extraction and is able to create these features on its own. Eventually, integration of deep neural network models into electronic medical records could result in more accurate risk scores generated automatically per patient, thereby providing real-time assistance in the triaging of patients.

Study Limitations

There are several limitations to this study. Perhaps most significantly, this study is from a single center and of a somewhat limited sample size. As mentioned above, deep learning models in other fields have included millions of samples. In order to address this limitation and avoid overfitting, we chose a limited number of features and implemented regularization training techniques commonly used in deep learning. In addition, there were only 87 mortality patients in the test data set. Thus, it is possible that the results generated here are not fully generalizable to other institutions and will need to be validated on other data sets.

Conclusions

To the best of our knowledge, this study is the first to demonstrate the ability to use deep learning to predict postoperative in-hospital mortality based on intraoperative electronic medical record data. The deep learning model presented in this study is robust, shows improved or comparable

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

discrimination to other risk scores, can be calculated automatically at the end of surgery, and does not rely on any administrative inputs.

Research Support

Support was provided solely from institutional and/or departmental sources.

Competing Interests

Dr. Lee is an Edwards Lifesciences (Irvine, California) employee, but this work was done independently from this position and as part of her Ph.D. Dr. Cannesson has ownership interest in Sironis, a company developing closed-loop systems, and does consulting for Edwards Lifesciences and Masimo Corp. (Irvine, California). Dr. Cannesson has received research support from Edwards Lifesciences through his department and National Institutes of Health (Bethesda, Maryland) grant Nos. R01 GM117622 ("Machine Learning of Physiological Variables to Predict Diagnose and Treat Cardiorespiratory Instability") and R01 NR013912 ("Predicting Patient Instability Noninvasively for Nursing Care-Two [PPINNC-2]"). The other authors declare no competing interests.

Correspondence

Address correspondence to Dr. Cannesson: Department of Anesthesiology and Perioperative Medicine, University of California Los Angeles, 757 Westwood Plaza, Los Angeles, California 90095. mcannesson@mednet.ucla.edu. This article may be accessed for personal use at no charge through the Journal Web site, www.anesthesiology.org.

References

1. Weiser TG, Regenbogen SE, Thompson KD, Haynes AB, Lipsitz SR, Berry WR, Gawande AA: An estimation of the global volume of surgery: A modelling strategy based on available data. Lancet 2008; 372:139–44
2. Pearse RM, Harrison DA, James P, Watson D, Hinds C, Rhodes A, Grounds RM, Bennett ED: Identification and characterisation of the high-risk surgical population in the United Kingdom. Crit Care 2006; 10:R81
3. Pearse RM, Moreno RP, Bauer P, Pelosi P, Metnitz P, Spies C, Vallet B, Vincent J-L, Hoeft A, Rhodes A; European Surgical Outcomes Study (EuSOS) group for the Trials groups of the European Society of Intensive Care Medicine and the European Society of Anaesthesiology: Mortality after surgery in Europe: A 7 day cohort study. Lancet 2012; 380:1059–65
4. American Society of Anesthesiologists: New classification of physical status. Anesthesiology 1963; 24:111
5. Gawande AA, Kwaan MR, Regenbogen SE, Lipsitz SA, Zinner MJ: An Apgar score for surgery. J Am Coll Surg 2007; 204:201–8
6. Reynolds PQ, Sanders NW, Schildcrout JS, Mercaldo ND, St Jacques PJ: Expansion of the surgical Apgar score across all surgical subspecialties as a means to predict postoperative mortality. Anesthesiology 2011; 114:1305–12
7. Haynes AB, Regenbogen SE, Weiser TG, Lipsitz SR, Dziekan G, Berry WR, Gawande AA: Surgical outcome measurement for a global patient population: Validation of the Surgical Apgar Score in 8 countries. Surgery 2011; 149:519–24
8. Regenbogen SE, Ehrenfeld JM, Lipsitz SR, Greenberg CC, Hutter MM, Gawande AA: Utility of the surgical apgar score: Validation in 4119 patients. Arch Surg 2009; 144:30–6; discussion 37
9. Terekhov MA, Ehrenfeld JM, Wanderer JP: Preoperative surgical risk predictions are not meaningfully improved by including the Surgical Apgar score: An analysis of the Risk Quantification Index and present-on-admission risk models. Anesthesiology 2015; 123:1059–66
10. Le Manach Y, Collins G, Rodseth R, Le Bihan-Benjamin C, Biccard B, Riou B, Devereaux PJ, Landais P: Preoperative Score to Predict Postoperative Mortality (POSPOM): Derivation and validation. Anesthesiology 2016; 124:570–9
11. Schmidhuber J: Neural networks. Reviews 2015; 61:85–117
12. Le Cun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD: Handwritten digit recognition with a back-propagation network. Morgan Kaufmann 1990
13. Baldi P, Chauvin Y: Neural networks for fingerprint recognition. Neural Computation 1993; 5
14. Krizhevsky, Sutskever, Hinton E: ImageNet classification with deep convolutional neural networks. Advances in neural information processing systems 2012
15. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A: Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015
16. Srivastava K, Greff, Schmidhuber J: Training very deep networks. Advances in Neural Information Processing Systems 2015
17. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition 2016
18. Wu L, Baldi P: Learning to play Go using recursive neural networks. Neural Netw 2008; 21:1392–400
19. Wu L, Baldi P: A scalable machine learning approach to GO. Advances in Neural Information Processing Systems 2007; 19
20. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D: Mastering the game of Go with deep neural networks and tree search. Nature 2016; 529:484–9
21. Baldi P, Sadowski P, Whiteson D: Searching for exotic particles in high-energy physics with deep learning. Nat Commun 2014; 5:4308
22. Sadowski PJ, Collado J, Whiteson D, Baldi P: Deep learning, dark knowledge, and dark matter. Journal of Machine Learning Research, Workshop and Conference Proceedings 2015; 42
23. Kayala MA, Azencott CA, Chen JH, Baldi P: Learning to predict chemical reactions. J Chem Inf Model 2011; 51:2209–22
24. Kayala MA, Baldi P: ReactionPredictor: Prediction of complex chemical reactions at the mechanistic level using machine learning. J Chem Inf Model 2012; 52:2526–40
25. Lusci A, Pollastri G, Baldi P: Deep architectures and deep learning in chemoinformatics: The prediction of aqueous solubility for drug-like molecules. J Chem Inf Model 2013; 53:1563–75
26. Di Lena P, Nagata K, Baldi P: Deep architectures for protein contact map prediction. Bioinformatics 2012; 28:2449–57
27. Baldi P, Pollastri G: The principled design of large-scale recursive neural network architectures—dag-rnns and the protein structure prediction problem. Journal of Machine Learning Research 2003; 4
28. Zhou J, Troyanskaya OG: Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 2015; 12:931–4
29. Guillame-Bert M, Dubrawski A, Wang D, Hravnak M, Clermont G, Pinsky MR: Learning temporal rules to forecast instability in continuously monitored patients. J Am Med Inform Assoc 2017; 24:47–53
30. Chen L, Dubrawski A, Clermont G, Hravnak M, Pinsky M: Modelling risk of cardio-respiratory instability as a heterogeneous process. AMIA Annual Symposium Proceedings 2015

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.

31. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, Bhatt DL, Fonarow GC, Laskey WK: Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: Comparison of machine learning and other statistical approaches. JAMA Cardiol 2017; 2:204–9
32. Shadmi E, Flaks-Manov N, Hoshen M, Goldman O, Bitterman H, Balicer RD: Predicting 30-day readmissions with preadmission electronic health record data. Med Care 2015; 53:283–9
33. Nguyen, Tran, Wickramasinghe: Deepr: A convolutional net for medical records. arXiv 2016; 1607.07519v1
34. Lipton Z, Kale D, Elkan C, Wetzel R: Learning to diagnose with LSTM recurrent neural networks. International Conference on Learning Representations 2016
35. Razavian N, Sontag D: Temporal convolutional neural networks for diagnosis from lab tests. arXiv 2016; 1511.07938v4
36. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016; 316:2402–10
37. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M: Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. J Med Internet Res 2016; 18:e323
38. Hofer IS, Gabel E, Pfeffer M, Mahbouba M, Mahajan A: A systematic approach to creation of a perioperative data warehouse. Anesth Analg 2016; 122:1880–4
39. Baldi P, Sadowski P: The dropout learning algorithm. Artificial Intelligence 2014; 210:78–122
40. Srivastava N: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 2014; 15
41. Hinton GE, Srivastava N, Krizhevsky A: Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580 2012
42. Sigakis MJ, Bittner EA, Wanderer JP: Validation of a risk stratification index and risk quantification index for predicting patient outcomes: In-hospital mortality, 30-day mortality, 1-year mortality, and length-of-stay. Anesthesiology 2013; 119:525–40
43. Sessler DI, Sigl JC, Manberg PJ, Kelley SD, Schubert A, Chamoun NG: Broadly applicable risk stratification system for predicting duration of hospitalization and mortality. Anesthesiology 2010; 113:1026–37
44. Chollet F: Keras. Available at: https://github.com/fchollet/keras. Accessed October 16, 2017
45. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011; 12

Downloaded from anesthesiology.pubs.asahq.org by Lingzhong Meng on 05/18/2019
Copyright © 2018, the American Society of Anesthesiologists, Inc. Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.