Deep Learning for Predicting In Hospital Mortality

CHRISTINE LEE, MS, UC IRVINE BIOMEDICAL ENGINEERING
IRA HOFER, MD, UCLA ANESTHESIOLOGY
MAXIME CANNesson, MD PHD, UCLA ANESTHESIOLOGY
PIERRE BALDI, PHD, UC IRVINE COMPUTER SCIENCE
The Need for Identifying Patients At Risk

More than 230 million major surgical procedures are performed annually\(^1\)

Overall mortality rate is less than 2\%\(^2,3\)

10\% of surgical population at high risk, but 80\% of postoperative deaths\(^2,3\)

Less than 15\% of high risk surgical patients are admitted to the ICU\(^3\)
Postoperative Risk Scores

- **ASA Score**
  - 1963
  - 1 = healthy and 5 = not expected to live 24 hours
  - Overall health
  - Preoperative

- **Surgical Apgar Score**
  - 2007
  - Risk of postoperative complication and poor outcome
  - Developed on ~300 patients
  - EBL, lowest MAP, and lowest HR

- **Risk Quantification Index**
  - 2011
  - Risk of 30 day morbidity and mortality
  - Developed on ~635,000 patients
  - CPT code of the performed primary procedure, ASA, AGE
  - Score + Probability of outcome

- **POSPOM**
  - 2016
  - Preoperative score to predict post operative mortality
  - Developed on >2 million patients
  - 17 preoperative variables (such as age, presence of heart disease, surgery type)
Limitations

ASA is subjective

ASA, POSPOM, and RQI are limited to preoperative information

RQI depends on Procedural Severity Score (PSS)

Surgical Apgar score has been shown to have limited accuracy\(^8\)

Adding Surgical Apgar to RQI to leverage both preoperative and intraoperative information does not significantly improve prediction of mortality\(^9\)
Aim of Study

Predict inhospital mortality in surgical patients by using deep neural network models (DNNs) and intraoperative features

- Compare DNNs to ASA, SAS, RQI, POSPOM, and logistic regression
- Assess DNN with a reduced feature set
- Assess DNN with addition of ASA
Deep Feedforward Neural Networks

"Non-deep" feedforward neural network

Deep neural network

Update neurons based on gradient

Loss Function
Deep Feedforward Neural Networks

Hyperparameters
- Number of neurons
- Number of hidden layers
- Batch size
  - Epoch = all training samples have been “seen” and weights updated accordingly
- Activation function
- Learning rate
- Regularization parameters
Data Description

Inclusion Criteria:
◦ All surgical procedures performed since March 1, 2013 with general anesthesia at UCLA

Exclusion Criteria:
◦ No general anesthesia
◦ > 89 years or < 18 years of age
◦ For patients with >1 procedure, only the first procedure was included

N = 59,985 patients
◦ 80% Train and 20% Test

87 features calculated/extracted at the end of surgery
◦ Descriptive intraoperative vital signs
◦ Summary of drugs and fluids interventions
◦ Patient anesthesia descriptions

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Patients</td>
<td>47,988</td>
<td>11,997</td>
</tr>
<tr>
<td># of Patients with In Hospital Mortality (%)</td>
<td>389 (0.81%)</td>
<td>87 (0.73%)</td>
</tr>
<tr>
<td>Age (yrs)</td>
<td>56 ± 17</td>
<td>56 ± 18</td>
</tr>
<tr>
<td>EBL (cc)</td>
<td>95 ± 540</td>
<td>94 ± 410</td>
</tr>
<tr>
<td>Presence of Arterial Line (%)</td>
<td>17.9</td>
<td>17.8</td>
</tr>
<tr>
<td>Presence of PA Line (%)</td>
<td>3.4</td>
<td>0.9</td>
</tr>
<tr>
<td>Presence of Central Line (%)</td>
<td>5.1</td>
<td>1.3</td>
</tr>
<tr>
<td>ASA (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6.3</td>
<td>1.6</td>
</tr>
<tr>
<td>2</td>
<td>37.4</td>
<td>9.3</td>
</tr>
<tr>
<td>3</td>
<td>49.9</td>
<td>12.5</td>
</tr>
<tr>
<td>4</td>
<td>6.1</td>
<td>1.5</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>CPT Code</td>
<td># Patients</td>
<td>CPT_DESCRIPTION</td>
</tr>
<tr>
<td>----------</td>
<td>------------</td>
<td>----------------------------------------------------------</td>
</tr>
<tr>
<td>43239</td>
<td>304</td>
<td>Esophagogastroduodenoscopy</td>
</tr>
<tr>
<td>45380</td>
<td>230</td>
<td>Colonoscopy</td>
</tr>
<tr>
<td>43259</td>
<td>225</td>
<td>Esophagogastroduodenoscopy</td>
</tr>
<tr>
<td>50360</td>
<td>193</td>
<td>Renal allotransplantation, implantation of graft; without recipient nephrectomy</td>
</tr>
<tr>
<td>47562</td>
<td>190</td>
<td>Laparoscopy, surgical; cholecystectomy</td>
</tr>
<tr>
<td>43242</td>
<td>181</td>
<td>Esophagogastroduodenoscopy</td>
</tr>
<tr>
<td>27447</td>
<td>162</td>
<td>Arthroplasty, knee, condyle and plateau</td>
</tr>
<tr>
<td>27130</td>
<td>153</td>
<td>Arthroplasty, acetabular and proximal femoral prosthetic replacement (total hip arthroplasty)</td>
</tr>
<tr>
<td>59841</td>
<td>149</td>
<td>Induced abortion, by dilation and evacuation</td>
</tr>
<tr>
<td>60500</td>
<td>144</td>
<td>Parathyroidectomy or exploration of parathyroid(s)</td>
</tr>
<tr>
<td>44970</td>
<td>127</td>
<td>Laparoscopy, surgical, appendectomy</td>
</tr>
<tr>
<td>55866</td>
<td>126</td>
<td>Laparoscopy, surgical prostatectomy</td>
</tr>
<tr>
<td>61510</td>
<td>106</td>
<td>Craniectomy, trephination, bone flap craniotomy</td>
</tr>
<tr>
<td>38724</td>
<td>103</td>
<td>Cervical lymphadenectomy (modified radical neck dissection)</td>
</tr>
</tbody>
</table>
87 Model Features

SBP
DBP
MAP
HR
Pulse Ox
SBP of the last 10 minutes of the case
DBP of the last 10 minutes of the case
MAP of the last 10 minutes of the case
HR of the last 10 minutes of the case
Pulse Ox of the last 10 minutes of the case
Current Rate of Phenylephrine
Current Rate of Vasopressin
Current Rate of Epinephrine
Current Rate of Milrinone
Current Rate of Nitroglycerin
Current Rate of Esmolol
Current Rate of Nitroprusside
Current Rate of Nicardipine
Maximum Glucose for the Case
Minimum Glucose for the Case

min, max, avg, med, std
min, max, avg, med, std
min, max, avg, med, std
min, max, avg, med, std
min, max, avg, med, std
min, max, avg, med, std
min, max, avg, med, std
min, max, avg, med, std
min, max, avg, med, std
min, max, avg, med, std
min, max, avg, med, std

Nitric Oxide Used for the Case
Presence of invasive central, radial, or pulmonary arterial line
Total Red Blood Cells Transfused
Total Urine Output
Cumulative minutes with MAP<60
Cumulative minutes with MAP<50
Total bolus dose of phenylephrine
Highest infusion rate of phenylephrine during the case
Total bolus dose of ephedrine
Total bolus dose of vasopressin
Highest infusion rate of vasopressin during the case
Total bolus dose of Epinephrine
Highest infusion rate of epinephrine during the case
Highest infusion rate of milrinone during the case
Total bolus dose of nitroglycerin during case
Highest infusion rate of nitroglycerin during the case
Total bolus dose of Esmolol during the case
Highest infusion rate of esmolol during the case
Highest infusion rate of nitroprusside during the case
Highest infusion rate of nicardipine during the case
Minimum Hemoglobin during the case
Maximum MAC of isoflurane during the case (note this is not age adjusted)
Maximum MAC of sevoflurane during the case (note this is not age adjusted)
Maximum MAC of desflurane during the case (note this is not age adjusted)
Methods: Summary

59,985 patients
87 features

- **80% Train**
- **20% Test**

Data Preprocessing

Train DNNs with 5 Fold Cross Validation

- Choose best DNN hyperparameters and architecture

Train DNN on all train data

Test DNN

Data Preprocessing
- Missing values filled with mean values
- Values clinically out of range filled with clinically normal values
- Train data features rescaled to have a mean of 0 and standard deviation of 1
Methods: Model

Feedforward networks with fully connected layers and a **sigmoid output**

Trained using stochastic gradient descent (SGD) with momentum and a batch size of 200.

Trained 4 DNN models using

1. All 87 features
2. Reduced feature set of 46 features
   ▪ This reduced feature set was created by excluding any average, median, standard deviation, and last 10 minutes of the surgical case features.
3. 87 original features + ASA = 88 features
4. 46 reduced features + ASA = 47 features

Model performance was assessed with AUC

For comparison, the AUCs of logistic regression (87 features), ASA, Surgical Apgar, RQI, and POSPOM were also calculated.
   ▪ ASA and POSPOM provided by UCLA
   ▪ Surgical Apgar calculated using Gawande et al.\(^5\)
   ▪ RQI could not be calculated using published R model from Cleveland Clinic’s website [1] due to technical issues with R version
   ▪ RQI log probability and score calculated from Sigakis et al.\(^12\)

Methods: Training Data Augmentation

Prior to training, positive training examples were augmented by adding Gaussian noise with a standard deviation of 0.0001

<table>
<thead>
<tr>
<th>% Occurrence</th>
<th>Data Augmentation</th>
<th>Augmented % Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inhospital Mortality</td>
<td>0.81</td>
<td>100x</td>
</tr>
</tbody>
</table>
Methods: Dealing with Overfitting

Solutions:
- Collect more data
Methods: Dealing with Overfitting

Early stopping\(^{10}\) with a patience of 10 epochs
- Stops training when validation loss starts to increase

L2 weight decay
- Penalize squared weights
- Keeps weights small unless error derivative is big

Dropout\(^{11}\) applied at all layers
- Neurons are removed from the network with a specified probability during training.
- This prevents neurons from co-adapting too much.

#### Log Loss Function with L2 Regularization

\[
L = -\frac{1}{n} \sum_x y \log a + (1 - y) \log(1 - a)
\]

\[
C = L + \frac{\lambda}{2} \sum_i w_i^2
\]

---

**Figure 1:** Dropout Neural Net Model. Left: A standard neural net with 2 hidden layers. Right: An example of a thinned net produced by applying dropout to the network on the left. Crossed units have been dropped.
## Results: Best Neural Network Architecture and Hyperparameters

<table>
<thead>
<tr>
<th>Activation</th>
<th>Output Activation</th>
<th>Initialization</th>
<th># Hidden Layers</th>
<th># Neurons</th>
<th>L2 Weight Decay</th>
<th>Dropout Probability</th>
<th>Learning Rate</th>
<th>Momentum</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReLu</td>
<td>Sigmoid</td>
<td>he_normal</td>
<td>4</td>
<td>[300, 300, 300, 300]</td>
<td>0.0001</td>
<td>0.5</td>
<td>0.01</td>
<td>0.9</td>
</tr>
</tbody>
</table>
**Results: AUC**

<table>
<thead>
<tr>
<th>Risk Score</th>
<th>AUC [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgical Apgar</td>
<td>0.58 [0.52 - 0.64]</td>
</tr>
<tr>
<td>POSPOM SCORE</td>
<td>0.74 [0.69 - 0.78]</td>
</tr>
<tr>
<td>ASA</td>
<td>0.84 [0.81 - 0.87]</td>
</tr>
<tr>
<td>RQI Log Prob**</td>
<td>0.90 [0.87 - 0.93]</td>
</tr>
<tr>
<td><strong>RQI Score</strong></td>
<td><strong>0.91 [0.87 - 0.94]</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.86 [0.81 - 0.89]</td>
</tr>
<tr>
<td>DNN</td>
<td>0.88 [0.85 - 0.91]</td>
</tr>
<tr>
<td>DNN w/ ASA</td>
<td>0.90 [0.87 - 0.93]</td>
</tr>
<tr>
<td>DNN w/ Reduced Feature Set</td>
<td>0.89 [0.85 - 0.92]</td>
</tr>
<tr>
<td><strong>BEST DNN:</strong></td>
<td><strong>DNN w/ Reduced Feature Set &amp; ASA</strong></td>
</tr>
</tbody>
</table>

**It should be noted that RQI could not be calculated for 6,406 of the test patients due to lack of Procedural Severity Scores for their CPT codes.**
Results: Boxplots

SAS < 4: 50% risk of major complications, including a 14% mortality rate

ASA = 3
Severe systemic disease

POSPOM < 20: probability of in-hospital mortality < 0.32%
POSPOM = 25: 1.37%
POSPOM = 30: 5.65%

RQI log prob > 10%: 40-50% of 30 day mortality

RQI score = 40:
~0.01% predicted probability of 30 day mortality
RQI score = 80:
~0.05%
RQI score = 110:
~1%6

RQI score = 40:
~0.01% predicted probability of 30 day mortality
RQI score = 80:
~0.05%
RQI score = 110:
~1%6
Results: Choosing Threshold based on F1 Score

<table>
<thead>
<tr>
<th>Best DNN Threshold</th>
<th>F1 Score</th>
<th>RQI Score</th>
<th>F1 Score</th>
<th>POSPOM</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.02</td>
<td>100</td>
<td>0.03</td>
<td>10</td>
<td>0.02</td>
</tr>
<tr>
<td>0.2</td>
<td>0.03</td>
<td>120</td>
<td>0.08</td>
<td>15</td>
<td>0.03</td>
</tr>
<tr>
<td>0.3</td>
<td>0.14</td>
<td>130</td>
<td>0.12</td>
<td>20</td>
<td>0.05</td>
</tr>
<tr>
<td>0.4</td>
<td>0.22</td>
<td>140</td>
<td>0.16</td>
<td>25</td>
<td>0.04</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>145</td>
<td>0.08</td>
<td>30</td>
<td>0.02</td>
</tr>
<tr>
<td>0.6</td>
<td>0.1</td>
<td>150</td>
<td>0.05</td>
<td>35</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th># True Negative</th>
<th># False Positive</th>
<th># False Negative</th>
<th># True Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best DNN Model</td>
<td>11,875</td>
<td>35</td>
<td>72</td>
<td>15</td>
</tr>
<tr>
<td>(n=11,997 all test patients)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best DNN Model</td>
<td>5,540</td>
<td>13</td>
<td>32</td>
<td>6</td>
</tr>
<tr>
<td>(n=5,591 RQI Score Calculated Patients)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RQI Score</td>
<td>5,502</td>
<td>51</td>
<td>30</td>
<td>8</td>
</tr>
<tr>
<td>(n=5,591 RQI Score Calculated Patients)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POSPOM</td>
<td>10,782</td>
<td>1,128</td>
<td>56</td>
<td>31</td>
</tr>
<tr>
<td>(n=11,997 all test patients)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POSPOM</td>
<td>4,948</td>
<td>605</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>(n=5,591 RQI Score Calculated Patients)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Our DNN has the highest F1 Score.

Choosing thresholds based on best F1 score optimizes true negatives.

DNN and RQI model were comparable.
## Results: % Mortality by Model

<table>
<thead>
<tr>
<th>Best DNN Model</th>
<th># Mortality</th>
<th>% of Mortality Patients (n=87)</th>
<th>RQI Score</th>
<th># Mortality</th>
<th>% of Mortality Patients (n=38)</th>
<th>POSPOM</th>
<th># Mortality</th>
<th>% of Mortality Patients (n=87)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-0.1</td>
<td>2</td>
<td>2.3</td>
<td>0-100</td>
<td>1</td>
<td>2.63</td>
<td>0-10</td>
<td>6</td>
<td>6.9</td>
</tr>
<tr>
<td>0.1-0.2</td>
<td>1</td>
<td>1.15</td>
<td>100-120</td>
<td>9</td>
<td>23.68</td>
<td>10-20</td>
<td>50</td>
<td>57.5</td>
</tr>
<tr>
<td>0.2-0.3</td>
<td>37</td>
<td>42.53</td>
<td>120-130</td>
<td>11</td>
<td>28.95</td>
<td>20-25</td>
<td>26</td>
<td>29.9</td>
</tr>
<tr>
<td>0.3-0.4</td>
<td>32</td>
<td>36.78</td>
<td>130-140</td>
<td>9</td>
<td>23.68</td>
<td>25-30</td>
<td>4</td>
<td>4.6</td>
</tr>
<tr>
<td>0.4-0.5</td>
<td>10</td>
<td>11.49</td>
<td>140-150</td>
<td>7</td>
<td>18.42</td>
<td>30-40</td>
<td>1</td>
<td>1.1</td>
</tr>
<tr>
<td>&gt;= 0.5</td>
<td>5</td>
<td>5.75</td>
<td>150-160</td>
<td>1</td>
<td>2.63</td>
<td>&gt;40</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Set **0.2** as threshold 97% of Mortality Patients

Set **100** as threshold 97% of Mortality Patients

Set **10** as threshold 93% of Mortality Patients
Results: Choosing Threshold Based on True Positives

At threshold values:
- Best DNN : 0.2
- RQI Score : 100
- POSPOM : 10

The best DNN with a threshold of 0.2 decreases the # of false positives compared to RQI by 352 patients, while comparably labeling true positives.

<table>
<thead>
<tr>
<th>Model</th>
<th># True Negative</th>
<th># False Positive</th>
<th># False Negative</th>
<th># True Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best DNN Model (n=11,997 all test patients)</td>
<td>6,680</td>
<td>5,230</td>
<td>3</td>
<td>84</td>
</tr>
<tr>
<td>Best DNN Model (n=5,591 RQI Score Calculated Patients)</td>
<td>3,385</td>
<td>2,168</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>RQI Score (n=5,591 RQI Score Calculated Patients)</td>
<td>3,033</td>
<td>2,520</td>
<td>1</td>
<td>37</td>
</tr>
<tr>
<td>POSPOM (n=11,997 all test patients)</td>
<td>2,741</td>
<td>9,169</td>
<td>6</td>
<td>81</td>
</tr>
<tr>
<td>POSPOM (n=5,591 RQI Score Calculated Patients)</td>
<td>1,312</td>
<td>4,241</td>
<td>1</td>
<td>37</td>
</tr>
</tbody>
</table>
Conclusions

DNN models predict inhospital mortality better or comparably to currently published risk scores

The addition of ASA and reducing the number of features improves the DNN models

RQI is comparable to our models, but can only be calculated on ~50% of patients

Our models can be calculated on all patients and leverages both preoperative and intraoperative information

Future Work:
- Testing on a different hospital’s patient population
- Leveraging time series data during operation
- Patient specific
References


