CS230

A Deep Learning Approach for Tracheostomy Risk Prediction during Trauma Admission
Healthcare

Jeff Choi, MD MSc
Department of Surgery
Stanford University
jc2226@stanford.edu

Chenyu Li, PhD
Department of Biomedical Data Science
Stanford University
chenyuo@stanford.edu

Eunice Yang
Department of Computer Science
Stanford University
eunicey@stanford.edu

Abstract

Prediction of tracheostomy is an ongoing challenge as clinicians are currently unable to accurately predict whether patients will require prolonged mechanical ventilation. Because early tracheostomy is associated with decreased medical costs and improved patient outcomes, the ability to predict which patients will require the procedure at an earlier timepoint in hospitalization is a crucial area of study. However, there is limited existing research on building these necessary, accurate models for the hospital setting. Here, we develop a binary classification neural network that analyzes a patient's medical history, using data from the National Trauma Data Bank which includes basic demographics as well as much more detailed metrics of injury characteristics and prior medical conditions. Our model performed better than prior models on tracheostomy in the literature, with an AUC of 0.841 and F1 of approximately 0.81. Further work is needed to identify the top factors of importance and to improve upon the deep learning approach we used. Overall, the development of such predictive models and incorporation into the hospital and ICU setting will significantly impact cost improvements and quality of care.

1 Introduction

Traumatic injury can compromise a patient's ability to breathe, necessitating mechanical ventilation for continued life support. Some patients require prolonged mechanical ventilation (>7 to 10 days) while others will recover and be removed from mechanical ventilation. Currently, with poor ability to predict who will require prolonged mechanical ventilation, physicians commonly monitor patients in intensive care units for weeks, which incurs considerable resources and risks other life-threatening complications (e.g. infection) [1, 2].

Tracheostomy (surgery to allow breathing through a neck incision) is a critical procedure for patients who require prolonged respiratory support. Tracheostomy allows patients-who otherwise would have been confined to their beds next to a mechanical ventilator-to walk, breathe more easily, and avoid complications associated with prolonged mechanical ventilation. A number of studies have suggested an association between early tracheostomy and decreased hospital costs, morbidity, and mortality [3, 4]. However, tracheostomy is a surgical procedure with inherent risks and should not be performed on patients who would have self-liberated from mechanical ventilation with more waiting. The ability to predict patients who will require tracheostomy early during hospitalization remains critically needed.

We aimed to build a deep neural network to predict which patients will require tracheostomy after suffering traumatic injury using demographic, injury, and other baseline characteristics readily available to clinicians.

2 Related work

In the last decade, machine learning and deep learning approaches have gained importance in medicine, including emergency, pulmonary, and critical care specialties. ML methods have, for example, been used to develop illness severity scores (multi-

CS230: Deep Learning, Spring 2021, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

classification) in the ICU [5, 6]. There have been very limited attempts to develop earlier predictions regarding ventilation or tracheostomy, and only a handful have used artificial intelligence or machine learning adjacent methods. One study predicted successful extubation in the ICU using a multilayer ANN model with 19 neurons in a hidden layer, achieving F1 of 0.867 and AUC of 0.85 [7]. Another landmark study utilized a gradient-boosted decision tree algorithm to identify patients at risk for prolonged mechanical ventilation [8]. Input was based on the Multiparameter Intelligent Monitoring in Intensive Care III database and used largely as input only six severity-of-illness scores, including the Oxford Acute Severity of Illness Score (OASIS) and the Simplified Acute Physiology Score (SAPS). This classifier achieved an AUC of 0.830 for tracheostomy prediction and an AUC of 0.820 for PMV prediction—this was the only existing study we identified predicting tracheostomy.

A recent study focused more specifically on assessing the optimal timing for tracheostomy insertion to wean COVID-19 patients with pneumonia off of mechanical ventilation [9]. This work was not necessarily meant to create a usable predictive model, but to help determine an optimal time for ventilation weaning in a specific use case. This approach constructed a time-series based decision tree using a C4.5 algorithm. Input variables including vital signs and serum-based biomarkers of disease severity were recorded at different time points during ICU admission (baseline and days 7, 10, 14).

ML-based prediction in healthcare has increased, but we were ultimately only able to find one study that specifically predicted patient tracheostomies. The researchers did achieve relatively high performance for a first attempt, however we believe the implementation was insufficient in that it did not take full advantage of newer deep learning approaches or the substantial breadth of information generated while patients are hospitalized. Binary classifiers in the clinical setting could benefit in general from utilizing additional more sensitive screening variables. We aimed to address these weaknesses in our project and develop a predictive model with increased performance.

3 Dataset and Features

Our project queries the 2017 National Trauma Data Bank (NTDB), the largest and most recent database of injured patients in the United States [10]. Our study population comprised 443,396 adults (age $\geq$18 years) admitted to the hospital after traumatic injury, among whom 7,828 (1.8%) underwent tracheostomy. We excluded patients who were hospitalized <72 hours, transferred from other hospitals, or underwent emergent tracheostomies (within 24 hours of hospital admission). Of our remaining set of 443,396 patients, we used a 90/10 training/test split, with 438962 in our training set and 4434 in our test set (Fig 1).

Each patient had a total of 1040 potentially relevant predictors in the NTDB for undergoing tracheostomy, including demographic variables (e.g. age, sex), underlying health conditions, injury characteristics (e.g. severity of injury), hospital characteristics (e.g. trauma center level), and 932 diagnoses associated with hospitalization (labeled as binary classifications). For these 932 diagnoses, we only included as input variables those that had one percent or greater prevalence in the population, meaning that extremely rare diagnoses were not included in our model. Finally, any variables with over 10% missingness were also excluded. Ultimately, 113 variables were considered as inputs for our model out of the 1040 officially provided by the NTDB.

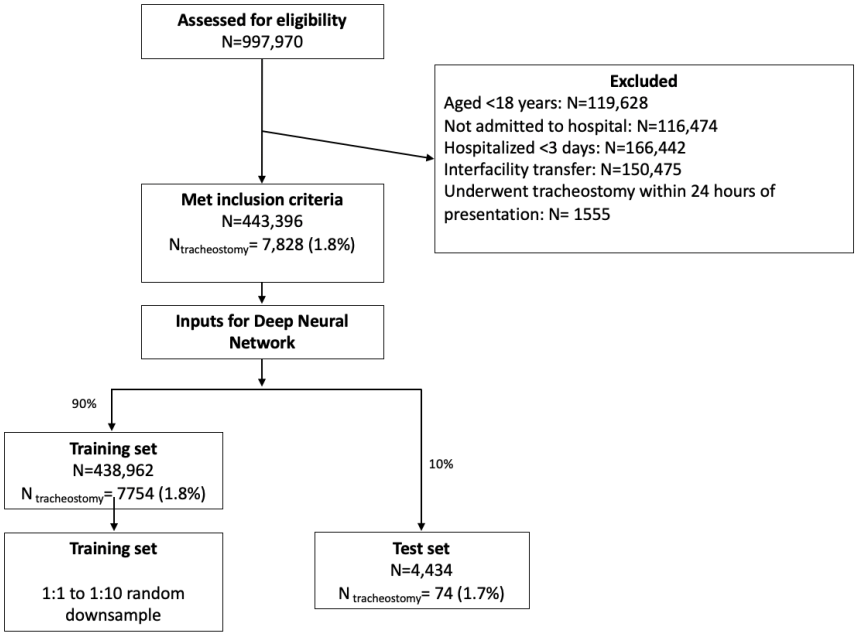

Figure 1: Study Population and Data Processing

4 Methods

We used a neural network for our binary classification problem, implemented via the Keras and TensorFlow packages [11, 12]. After analyzing previous medical studies and ML methods, we developed a final algorithm with 30 nodes per dense layer with Keras. The loss function we choose was binary cross-entropy, where y is the label (1 for tracheostomy and 0 otherwise) and p(y) is the predicted probability of that patient undergoing a tracheostomy.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$$

Our first model included one layer with a Sigmoid activation function - this served as our initial baseline. During the development of our model, we tested with different nodes and determined that 30 nodes per layer yielded most optimal results. Then, we included multiple layers within our model, with two fully connected layers. This additional layer increased the performance of our model. We additionally modified a new model to include dropout layers and chose Adam as our optimization function. Our final model is represented graphically in figure 2, and the summary of our model is shown in figure 3.

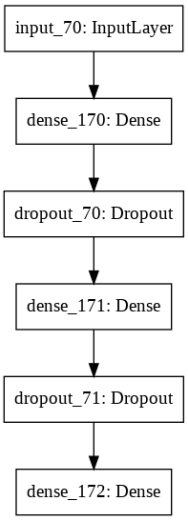

Figure 2: Bi-classification Neural

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 1005)]	0
dense_3 (Dense)	(None, 100)	100600
dropout (Dropout)	(None, 100)	0
dense_4 (Dense)	(None, 30)	3030
dropout_1 (Dropout)	(None, 30)	0
dense_5 (Dense)	(None, 1)	31

Total params: 103,661
Trainable params: 103,661
Non-trainable params: 0

Loss = 0.06581608206033707
Test Accuracy = 0.9821150898933411

Figure 3: Model Summary

We tested a series of evaluation metrics throughout this project, including loss, accuracy, precision, recall, and F1. Though we did generate results from all 5 evaluation metrics, we ultimately relied on our F1 score as the primary metric for evaluating our output. This is because both precision and recall are equally important in the context of our problem, as there is a high cost for false positive (which subjects patients to unnecessary tracheostomy and an invasive surgery) and false negatives (which delays providing a patient with a necessary tracheostomy). We define F1, precision, and recall as follows:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \tag{1}$$

5 Results

Because tracheostomies are considered a "rare event" and comprised only 1.8% of our training set (with 7754 cases), we performed random downsampling to train on a disproportionately lower subset of majority class examples. We created ten training sets from 1:1 to 1:10 and evaluated our model's performance given these different proportions of our majority group (Figure 4). We ultimately proceeded to use the 1:1 set for model training because it generated the best F1 score, our chosen metric. The F1 score (as well as precision and recall), decreased substantially as we included larger majority groups.

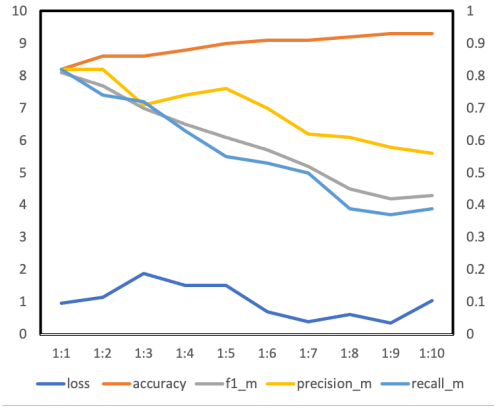

Figure 4: Model Performance based on Downsampling of Majority Group

After training on the 1:1 data set and running the model on our test set, we generated an ROC curve with an AUC of 0.841. This was an exciting development as our deep-learning approach demonstrated a boost over the performance shown in existing literature. Our model's AUC improves upon the highest AUC we identified in published work for tracheostomy prediction classification of approximately 0.830.

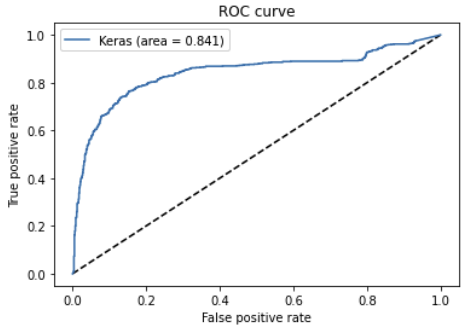

Figure 5: ROC of our Bi-Classification Model

We tried out different iterations on our model and found 40 iterations to be optimal. Evaluating our full suite of metrics, we find that at 40 epoch our model reaches F1 of ~0.8 and accuracy of >0.81 (Fig 6). Precision and recall also reached ~0.8, but precision and especially recall show marked instability across epochs. Though our main F1 metric was generally stable, there are likely aspects of our model that could be further modified. Additionally, we took steps not to overfit our model (ie using dropout layers), and our metrics on the testing set are reasonably close to those on the training set.

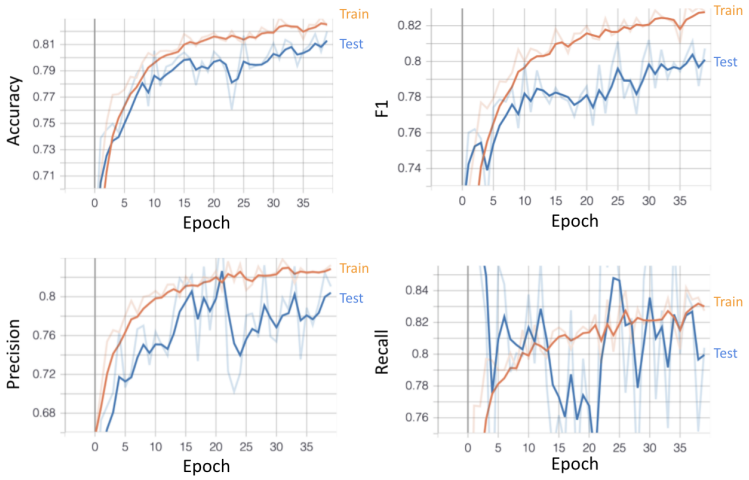

Figure 6: Model Evaluation, Epoch vs Accuracy, Precision, Recall, and F1 Score

Because this classification model is primarily intended for use in the clinical setting, it was important for us to understand what the model was learning and what features it was primarily reliant on (in contrast to the stereotypical ML blackbox that is challenging for medicine). From our 113 input variables, we utilized backwards selection to identify the top 30 variables affecting the F1 score (Table 1, top 12 shown for brevity). When utilizing just these 30 variables (instead of the entire 113 we had previously used), our model demonstrated **precision** of 0.75, **recall** of 0.70, and **F1 score** of 0.72 on our testing set.

On a clinical level, some of these features are intuitive but several are not. For example, a patient presenting with acute opioid poisoning or fractures of the skull and facial bones could very likely require tracheostomy later on during treatment. In contrast, superficial hip or thigh injuries on their own are unlikely to be causal factors for ventilatory failure and subsequent tracheostomy. This suggests that there may be further improvements to be made to our model, and that other approaches may be needed to identify variables of importance. Additionally, the decrease in our F1 score when utilizing just the top 30 variables suggest that there were potentially crucial features eliminated.

Table 1: First 12 of Top 30 Variables for F1

NTDB Code	Variable
S37	Injury of urinary and pelvic organs
S91	Open wound of ankle, foot and toes
Icpparench	Requiring Brain Monitor
Drgscr_oxycodone	Opioid Poisoning
S83	Dislocation and sprain of joints and ligaments of knee
S20	Superficial injury of thorax
S70	Superficial injury of hip and thigh
cc_bleeding	Comorbidity of bleeding
S51	Open wound of elbow and forearm
S02	Fracture of skull and facial bones
cc_anticoagulant	On anticoagulant
cc_chemo	On chemo

6 Conclusion/Future Work

Developing accurate predictions of patient tracheostomy need is challenging area of study that could help lower clinical costs and improve patient health outcomes. In this project, we were able to build a deep learning model that can accurately predict which patients will subsequently require tracheostomy after traumatic injury. After trying out different changes to our architecture (ie adjusting our layers, optimization function, etc), we developed a final binary classification neural network with an AUC of 0.830 and a F1 after 40 epoch of approximately 0.8. This improved on the few existing examples we were able to find in the existing literature, which may be partly due to our use of a deep learning approach and partly due to the amount of patient data and diversity of variables we were able to fully leverage in our project.

Some potential ideas that we believe are immediately applicable and interesting for future work include:

1) Adopting new approaches that can better characterize rare events. Our project was somewhat limited by tracheostomies being rare, with 1.8% in our training set and 1.7% in our testing set. The downsampling method we used to counteract this will limit the generalizability of our model for real world applications. Incorporating encoders, which are effective for rare event classification, could be an important next step.

2) Implementing Lasso/Ridge regression or other approaches that can better consider correlated variables. Our use of backward selection was unable to incorporate correlation, and although we did attempt to run different "combinations" of features, this had too high of a computation cost and was ultimately infeasible.

3) Developing models that can additionally incorporate timeseries data, such as heartrate or glucose levels from health sensors. LSTM networks for example, would be well-suited to the task of making predictions based on this data, which are abundantly generated in a hospital ER setting. Our dataset did not provide this level of data, but incorporating another datastream could open the avenue to more accurate models.

7 Contributions

Team members contributed equally to this project. Jeff took the lead on data processing and providing medical expertise the analysis of this project. Chenyu took the lead on optimizing our models and experimenting evaluation and backward selection.

Eunice assisted with both of these aspects and compiling the write-up of our report. We regularly met over Zoom to work through our project and were all involved with all aspects. We would also like to thank our project mentor, Shahab Mousavi, for his very useful advice and guidance of our work througout this quarter!

References

[1] Ross BJ, Barker DE, Russell WL, Burns RP. Prediction of long-term ventilatory support in trauma patients. Am Surg. 1996 Jan;62(1):19-25. PMID: 8540640.

[2] Arabi Y, Haddad S, Shirawi N, Al Shimemeri A. Early tracheostomy in intensive care trauma patients improves resource utilization: a cohort study and literature review. Crit Care Lond Engl. 2004;8(5):R347-352. doi:10.1186/cc2924

[3] McCredie VA, Alali AS, Scales DC, et al. Effect of Early Versus Late Tracheostomy or Prolonged Intubation in Critically Ill Patients with Acute Brain Injury: A Systematic Review and Meta-Analysis. Neurocrit Care. 2017;26(1):14-25. doi:10.1007/s12028-016-0297-z

[4] Perry A, Mallah MD, Cunningham KW, et al. PATHway to success: Implementation of a multiprofessional acute trauma health care team decreased length of stay and cost in patients with neurological injury requiring tracheostomy. J Trauma Acute Care Surg. 2020;88(1):176-179. doi:10.1097/TA.0000000000002494

[5] Awad A, Bader-El-Den M, McNicholas J, Briggs J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. Int J Med Inform. 2017 Dec;108:185-195. doi: 10.1016/j.ijmedinf.2017.10.002. Epub 2017 Oct 5. PMID: 29132626.

[6] Cosgriff CV, Celi LA, Ko S, Sundaresan T, Armengol de la Hoz MÁ, Kaufman AR, Stone DJ, Badawi O, Deliberato RO. Developing well-calibrated illness severity scores for decision support in the critically ill. NPJ Digit Med. 2019 Aug 15;2:76. doi: 10.1038/s41746-019-0153-6. PMID: 31428687; PMCID: PMC6695410.

[7] Hsieh MH, Hsieh MJ, Chen CM, Hsieh CC, Chao CM, Lai CC. An Artificial Neural Network Model for Predicting Successful Extubation in Intensive Care Units. J Clin Med. 2018 Aug 25;7(9):240. doi: 10.3390/jcm7090240. PMID: 30149612; PMCID: PMC6162865.

[8] Parreco J, Hidalgo A, Parks JJ, Kozol R, Rattan R. Using artificial intelligence to predict prolonged mechanical ventilation and tracheostomyplacement. J Surg Res. 2018 Aug;228:179-187. doi: 10.1016/j.jss.2018.03.028. Epub 2018 Apr 11. PMID: 29907209

[9] Takhar A, Surda P, Ahmad I, Amin N, Arora A, Camporota L, Denniston P, El-Boghdadly K, Kvassay M, Macekova D, Munk M,Ranford D, Rabcan J, Tornari C, Wyncoll D, Zaitseva E, Hart N, Tricklebank S. Timing of Tracheostomy for Prolonged Respiratory Weanin Critically Ill Coronavirus Disease 2019 Patients: A Machine Learning Approach. Crit Care Explor. 2020 Nov 17;2(11):e0279. doi:10.1097/CCE.0000000000000279. PMID: 33225305; PMCID: PMC7673767.

[10] Comittee on Trauma AC of S. NTDB Version 2016. Chicago, Il,. 2017;

[11] Chollet F. Keras: Theano-based Deep Learning library. Code: https://github. com/fchollet. Documentation: http://keras.io 2015.

[12] Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv Preprint arXiv:1603.04467, 2016.