UNIVERSITY OF CALIFORNIA,
IRVINE

Developing Predictive Models for Risk of Postoperative Complications and Hemodynamic Instability in Patients Undergoing Surgery

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Biomedical Engineering

By

Christine Kim Lee

Dissertation Committee:
Distinguished Professor Pierre Baldi, Chair
Professor Maxime Cannesson
Professor Bernard Choi

DEDICATION

To Sam, I couldn't have finished this without you.

We started this journey together and these past long six years you've been not only a loving partner but also my biggest supporter. I am so happy to share this accomplishment with you. You calmed me down when I needed it and rallied me when I felt like giving up. And I am so excited to move on to the next chapter of our lives together. I love you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to first thank all my committee members for their time and support.

My research career started under the guidance of Dr. Maxime Cannesson at UCI Medical Center almost 8 years ago. He has been a mentor to me in all aspects of my life and I owe much of the success in my career to his advice and support. He has set for me the example of how to not only be passionate about your work, but also how to take the time to appreciate what is important in life, like cycling, France, and family. I came to work with him right after college, thinking I wanted to go to medical school, and then thinking maybe I'd do a Master's in biomedical engineering because I didn't know what else to do next. He saw potential in me before I even thought to set bigger goals for myself, without his push I wouldn't have pursued a PhD at all.

I would also like to acknowledge the amazing people of the research team in the Department of Anesthesiology as well as my team at Edwards Lifesciences. In Anesthesiology, there were many collaborators and undergraduate students who deserve recognition for their time and effort in the data collection for this work. Specifically, I would like to point out Dr. Joe Rinehart, Michael Ma, and Paulette Mensah, who did everything from IRB submissions and subject enrollment to making sure my tuition was paid on time. Also, I would like to thank Dr. Ceci Canales who was my first boss ever. She is a role model for how to succeed with humility and good nature, and also how to stand your ground when you need to. At Edwards Lifesciences, I got my first internship and job as an engineer. There are many people to thank for their words of advice and support, especially the Algorithms team, but Feras Hatib gave me that first opportunity.

During my graduate tenure, I have had the opportunity to be advised by two well-recognized leaders in their fields, Professor Pierre Baldi and Professor Bruce Tromberg. Pierre welcomed me into his group despite my lack of experience in coding let alone neural networks. Maxime and I approached him with nothing but a little data we had just started collecting for the eventually larger UCI anesthesiology research database, hoping that under his advisement I could learn how to apply deep neural networks to our data. In his group, I also met Peter Sadowski, who took the time to get me set up in the lab and never turned me down when I had a question or needed help. I'd also like to acknowledge Yuzo Kanomato for all his help in everything from GPU allocation to packages not working to moving large amounts of data; and Muntaha Samad for her help in finishing up my last projects and continuing on the work. Bruce was the previous chair of my committee and is unquestionably one of the most positive and supportive people. He was integral in my transition from Master's to PhD, and championed department support for my decision to work full time while finishing my PhD. Despite never directly advising my research, he never failed to ask if there was anything he could do to help or how Sam was doing.

Most of all, I'd like to thank Sam, Boots, my family, and my friends. I couldn't have gotten here without their support and this achievement

Finally, I'd like to acknowledge the Departments of Anesthesiology at University of California, Los Angeles and Irvine for funding this work.

CURRICULUM VITAE

Christine Kim Lee

EDUCATION

Ph.D. Biomedical Engineering - University of California, Irvine 2019

B.S. Mathematical Biology - Harvey Mudd College 2011

PEER-REVIEWED PUBLICATIONS

Accepted*/In draft**

1. **Lee, C. K.**, Hofer, I., Gabel, E., Baldi, P., & Cannesson, M. Development and Validation of a Deep Neural Network Model to Predict Postoperative Mortality, Acute Kidney Injury, and Reintubation using a single feature set *
2. **Lee, C.,** Ershoff, B., Wray, C., Agopian, V., Urban, G., Baldi, P., Cannesson, M. The Training and Validation of Deep Neural Networks for the Prediction of 90-Day Post-Liver Transplant Mortality Using UNOS Registry Data *
3. **Lee, C.,** Samad, M., Hofer, I., Baldi, P., Cannesson, M. An Interpretable Neural Network for Prediction of Postoperative In-hospital Mortality **
4. **Lee, C.,** Cannesson, M., Baldi, P. Prediction of Postinduction Hypotension with Deep Learning **

Accepted/Published

1. **Lee, C. K.,** Hofer, I., Gabel, E., Baldi, P., & Cannesson, M. (2018). Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, *129*(4), 649-662.
2. Hatib, F., Jian, Z., Buddi, S., **Lee, C.,** Settels, J., Sibert, K., Rinehart, J. & Cannesson, M. (2018). Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, *129*(4), 663-674.
3. Lilot, M., Ehrenfeld, J. M., **Lee, C.**, Harrington, B., Cannesson, M., & Rinehart, J. (2015). Variability in practice and factors predictive of total crystalloid administration during abdominal surgery: retrospective two-centre analysis. *British journal of anaesthesia*, *114*(5), 767-776.
4. Rinehart, J., Lilot, M., **Lee, C.,** Joosten, A., Huynh, T., Canales, C., ... & Cannesson, M. (2015). Closed-loop assisted versus manual goal-directed fluid therapy during high-risk abdominal surgery: a case–control study with propensity matching. *Critical Care*, *19*(1), 94.
5. **Lee, C. K.**, Rinehart, J., Canales, C., & Cannesson, M. (2014). Comparison of automated vs. manual determination of the respiratory variations in the EKG R wave amplitude for the prediction of fluid responsiveness during surgery. *Journal of Computational Surgery*, *1*(1), 5.

6. Rinehart, J., Le Manach, Y., Douiri, H., **Lee, C.**, Lilot, M., Le, K., Canales, C., & Cannesson, M. (2014, March). First closed-loop goal directed fluid therapy during surgery: a pilot study. In *Annales francaises d'anesthesie et de reanimation* (Vol. 33, No. 3, pp. e35-e41). Elsevier Masson.
7. Rinehart, J., **Lee, C.,** Cannesson, M., & Dumont, G. (2013). Closed-loop fluid resuscitation: robustness against weight and cardiac contractility variations. *Anesthesia & Analgesia*, *117*(5), 1110-1118.
8. Rinehart, J., **Lee, C.**, Canales, C., Kong, A., Kain, Z., & Cannesson, M. (2013). Closed-loop fluid administration compared to anesthesiologist management for hemodynamic optimization and resuscitation during surgery: an in vivo study. *Anesthesia & Analgesia*, *117*(5), 1119-1129.
9. Alexander, B. S., Gelb, A. W., Mantulin, W. W., Cerussi, A. E., Tromberg, B. J., Yu, Z., **Lee, C.**, & Meng, L. (2013). Impact of stepwise hyperventilation on cerebral tissue oxygen saturation in anesthetized patients: a mechanistic study. *Acta Anaesthesiologica Scandinavica*, *57*(5), 604-612.

CONFERENCE ABSTRACTS AND PRESENTATIONS

1. **Lee, C.,** Samad, M., Hofer, I., Baldi, P., Cannesson, M. An Interpretable Neural Network for Prediction of Postoperative In-hospital Mortality – Selected for **Best of Abstracts: Clinical Science** session for presentation at American Society of Anesthesiologists, October 2019
2. **Lee, C.,** Cannesson, M., Baldi, P. Prediction of Postinduction Hypotension with Deep Learning – Accepted for presentation at American Society of Anesthesiologists, October 2019
3. **Lee, C.,** Hofer, I., Baldi, P., & Cannesson, M. Deep Learning for Predicting Postoperative Outcomes: AKI, Reintubation, and Mortality. Presented at American Society of Anesthesiologists, October 2017
4. **Lee, C.**, Hofer, I., Cannesson, M., & Baldi, P. Deep Learning for Predicting In Hospital Mortality. In *ANESTHESIA AND ANALGESIA* (Vol. 124, pp. 85-86). Presented at Society for Technology in Anesthesia, January 2017 – **Best of Show Award**
5. **Lee, C.**, Hatib, F., Jian Z., Rinehart, J., Cannesson, M. Feasibility of Real-Time Prescriptive Analytics to Make Predictions and Suggests Decision Options for the Prevention of Hypotension. Presented at American Society of Anesthesiologists, October 2016
6. **Lee, C. K.,** Hatib, F., Jian, Z., Buddi, S., Cannesson, M. Use of Big Data and Machine Learning for Prediction of Hypotensive Events in High Risk ICU Patients From the MIMIC II MIT Database. In *ANESTHESIA AND ANALGESIA* (Vol. 122). Presented at American Society of Anesthesiologists, October 2016
7. **Lee, C.,** Cannesson, M., & Hatib, F. Pilot Study: Feasibility of Predictive Analytics for the Early Detection of Hypotensive Events. In *ANESTHESIA AND ANALGESIA* (Vol. 122). Presented at Society for Technology in Anesthesia, January 2016

BOOK CHAPTERS

Lee, C. K. (2014). Signal Analysis: Acquisition, Storage, and Analysis of Physiological Signals. In *Monitoring Technologies in Acute Care Environments* (pp. 29-33). Springer, New York, NY. Editors: Jesse Ehrenfeld, MD, and Maxime Cannesson, MD; Springer

OTHER

Creator and Lead, Society for Technology in Anesthesia Annual Young Researchers Workshop 2015 to 2019

Speaker and Panel Member, "Artificial Intelligence in the Acute Care Setting: Leveraging UC Wide Physiome Data set to Bring New Discovery and Improve Patient Care"
Lawrence Livermore Data Science Workshop, July 2019.

Speaker and Panel Member, "What is Machine Learning and How it Can Be Leveraged in Anesthesiology?"
Society for Technology in Anesthesia Conference, January 2019

Speaker and Panel Member, "Leveraging Machine Learning to Improve Patient Care: What Is Involved, and What It Really Means"
International Anesthesia Research Society Conference, April 2018

Python Instructor, UC Irvine Data Science Initiative 2015 to 2017

ABSTRACT OF THE DISSERTATION

Developing Predictive Models for Risk of Postoperative Complications and Hemodynamic Instability in Patients Undergoing Surgery

By

Christine Kim Lee

Doctor of Philosophy in Biomedical Engineering

University of California, Irvine, 2019

Professor Pierre Baldi, Chair

Patients undergoing high-risk surgeries are often at higher risk of developing hemodynamic instability during surgery resulting in poor postoperative outcomes. This is usually associated with significantly increased postoperative morbidity and mortality, which therefore makes the early identification of these critical events and those patients at risk of postoperative complications crucial. With these motivations in mind, we first created a large deidentified research dataset of surgical case medical records from University of California, Irvine Medical Center (UCIMC) matched with physiological waveforms as well as intermittent vital sign values, lab values, and ventilator settings. To our knowledge, such a dataset does not currently exist for the intraoperative environment. We hope that creating a such a dataset will allow for advances in machine learning for intraoperative care. Using medical data from UCLA, we have developed deep neural network models to classify the risks of postoperative mortality, acute kidney injury, and reintubation utilizing readily available intraoperative information. Our

risk scores were compared to currently commonly used risk indices ASA and Surgical Apgar as well as logistic regression. While the deep neural network models performed better than the risk scores and logistic regression, clinicians require additional information to assess what led to increased risk of complications. To address this, we also assessed the use of generalized additive neural networks (GANNs) to create a graphical look at how different features contributed to the risk of in hospital mortality. Finally, we were also interested in predicting critical intraoperative events to allow for time for the clinician to avoid such events. We focused on intraoperative hypotension as it is easier to define and has been shown to lead to increased risk of acute kidney injury, stroke, and myocardial injury. For the hypotension prediction models, we looked at the arterial pressure waveform and EMR data as inputs.

Overall, these aims address a gap in current clinical decision guidance and support to reduce adverse events during surgery as well complications after.

INTRODUCTION

Patients undergoing high-risk surgeries are often at higher risk of developing hemodynamic instability during surgery as well as having poor postoperative outcomes. This is usually associated with significantly increased postoperative morbidity and mortality, which therefore makes the early identification of these critical events and those patients likely to have poor outcomes crucial. As instability develops in a patient, there are small but complex changes in multiple vital signs that are not immediately apparent to a clinician. This instability continues to progress until significant large shifts occur in the patient (i.e. elevated heart rate and blood pressure), which are now obvious and detectable by the clinician but at which point may be too late to treat. Known methods are available to monitor invasive and noninvasive hemodynamic parameters to identify instability, however no such mechanism exists to predict the onset of instability in real time.

Although instability can be identified and diagnosed by abnormalities in the values of hemodynamic parameters seen on clinical monitors, the true underlying physiological mechanism of cardiorespiratory instability is neither straightforward nor linear and is much more complex. Currently, most clinician decisions are based solely on the values of individual hemodynamic parameters consistent with severe instability and outside of normal clinical range, such as mean arterial blood pressure less than 55 mmHg or a heart rate greater 120 bpm. However, the definition of “normal” is subjective to the treating clinician. While there are currently available risk scores such as the Surgical Apgar score and real time parameters such as stroke volume variation and heart rate variability that are used to better guide clinical decisions, they are still dependent on absolute thresholds of normal clinical range. It is known

that there is large patient to patient variability, and to focus on such thresholds is a major limitation. While there is the goal to minimize instability during surgery to improve postoperative outcomes, there is also a secondary need to identify patients who are at higher risk for postoperative complications. Being able to identify these patients would allow for more effective care and hopefully avoid complication altogether. In addition, with the payment of healthcare moving towards bundled payments, there is a financial need to efficiently allocate hospital resources and time.

There is a large potential for significant advancements in medicine and machine learning methods exist to help utilize the underlying complexity of patient physiology that current clinical monitors and decision support tools lack. We hypothesize that deep neural networks can leverage the complexity of intraoperative data taken from clinical monitors as well as medical records to better classify risk of specific postoperative complications and better predict postoperative outcomes as well as the onset of instability (hypotension).

With more than 230 million major surgical procedures are performed annually worldwide and an estimated 10% of surgical patients at high risk accounting for 80% of postoperative deaths, many lives can be saved simply by identifying patients at highest risk of specific postoperative complications to avoid onset of those complications.[2–4] In addition, helping to guide the intraoperative anesthesia care can help reduce this risk or avoid the complication altogether. Many of the currently developed models for clinical risk are not robust to patient to patient variability and rely on limited features selected by domain experts. On the other hand, while there is work being done utilizing machine learning, including deep neural networks, to classify patient risk and leverage time series data, there has been no work that we

know of specific to the intraoperative environment or on a shorter time scale. Events in the operating room are on the order of minutes or even seconds to an adverse event, in contrast to the slow decline of patients in the ICU setting. Much of the advancement of model development on surgical patients has also been limited by the availability of data. While high resolution clinical data exists for the critical care setting through the publicly available MIMIC II database created by the Massachusetts Institute of Technology, the same such database does not exist for the surgical setting.[5] Thus, the creation of such a database would significantly help to advance the progress of research on these types of patients.

The Need for Perioperative Risk Assessment

It has been shown that while only about 10% of surgical patients are considered at high risk for complications, this high risk population accounts for 80% of postoperative deaths.[2–4] Postoperative complications, as defined by the National Surgical Quality Improvement Program (NSQIP), comprise of cardiac, neural, renal, pulmonary, and vascular/thrombotic events as well as infections. These complications include cardiac arrest, renal insufficiency or failure, pneumonia, etc. It has been shown that occurrence of these complications within 30 days following a major surgery is a more significant determinant of survival than either preoperative comorbidity or intraoperative adverse events.[6] The top most important predictors of mortality included the following postoperative complications: cardiac arrest, failure to wean, systemic sepsis, cerebrovascular accident, renal failure, myocardial infarction, and renal insufficiency. Therefore, there needs to be a focus on the prevention of postoperative complications.

One such method of avoiding complications would be direct postoperative critical care admission. Despite the high mortality rates of the high risk surgical population, less than 15% of these patients are admitted to the ICU.[2,3] This suggests a systematic failure in the allocation and process of critical care resources. One way to assist with this would be able to identify the patients at most risk of major postoperative complications and death. In addition, there is a need for identifying patients who are at continued risk following discharge. Hospital readmission within 30 days is broadly considered as a healthcare quality measure and cost driver in the United States, and under the Affordable Care Act, Medicare has started penalizing hospitals according to their 30-day readmission rate. Currently, about 1 in 5 Medicare beneficiaries are rehospitalized within 30 days after discharge.[7] Three quarters of these readmissions were considered avoidable. In 2011, there were approximately 3.3 million adult 30-day all-cause hospital readmissions in the United States, resulting in about $41 billion in hospital costs and in 2004 Medicare payments for unplanned readmissions accounted for approximately $17 billion.[7,8] Overall, there is a need for methods to best prioritize care to avoid postoperative complications as well as readmission from both a public health as well as cost stand point. Such methods would allow for hospital systems to more effectively allocate resources available to high risk patients.

Current Parameters for Risk Assessment

Accurate risk prediction is crucial to guiding clinical decision and management. Currently, risk assessment is performed as a one-time risk score at patient admission or presentation, or is calculated as needed, for example at the end of each postoperative day.

Some well-known risk scores include the American Society of Anesthesiologists (ASA) physical status score, Acute Physiology and Chronic Health Evaluation (APACHE) II, and Surgical Apgar.[9–11] The ASA score was developed in 1963 and is still used in the preoperative environment as a subjective assessment of a patient's overall health prior to surgery. An ASA score of 1 means completely health and 5 means not expected live 24 hours.[9] The APACHE score is calculated upon admission into the ICU to estimate mortality, and takes the worst values of vital signs such as mean arterial pressure and heart rate as well as labs such as serum creatinine and hematocrit from the first 24 hours to assign risk points.[11] Out of a possible 71 points, the higher the score, the more likelihood of mortality. The Surgical Apgar score, also a point based system, uses only 3 intraoperative values: estimated blood loss, lowest mean arterial pressure, and lowest heart rate to predict postoperative risk of major complication.[10] There are also risk scores that can be calculated more frequently based on vital signs. These types of scores tend to primarily be used as triggers for immediate action such as calling the rapid response teams to recognize and respond to clinical deterioration. While there are several, two well-known ones include the Modified Early Warning Score (MEWS) and the Shock Index. MEWS is used in the critical care setting and is calculated at intermittent times during admission.[12] MEWS combines values of respiratory rate, heart rate, systolic blood pressure, urine output, temperature and neurological assessment. The Shock Index is mainly used in critical care as well as emergency settings. It is calculated as the ratio of heart rate over systolic blood pressure and can be used to predict cardiac arrest, hypovolemic shock, and sepsis.[13] While measures like MEWS and the Shock Index can be calculated continuously during a patient's admission, they are currently only calculated on a "need base". In summary, current risk scores do exist for patients undergoing

surgery and critical care. However, the scores themselves as well as the clinical variables used in their calculations tend to be specific to the preoperative setting or are calculated once the patient is already in critical care. In response to this there has been work to create newer and potentially more robust scores.

Related Work in Clinical Risk Modeling Efforts

The above scores were developed to create a set of easy to apply rules derived from expert opinion on what leads to a patient's deterioration. With the passing of the HITECH Act in 2009, there has been an explosion in the amount and availability of electronic medical data. This has lead to a growing body of research applying predictive models to medical data. The Preoperative Score to Predict Postoperative Mortality (POSPOM) was developed as a preoperative risk score to predict postoperative in hospital mortality.[14] The POSPOM was developed via a logistic regression model that takes into account preoperatively available patient demographics and conditions such as age, diabetes and chronic heart failure as well as type of surgery. The regression coefficients are then normalized to the regression coefficient for age to create POSPOM points. For example, age has a regression coefficient of 0.303 and diabetes and chronic heart failure have regression coefficients of 0.189 and 1.124, respectively, and so are assigned POSPOM points of 1 and 4, respectively. These POSPOM points are then summed to assign the patient with a final POSPOM. In addition, there is the Rothman Index (RI) which claims "real time" assessment of a patient's current condition, and was developed to predict excess risk of one year mortality, which was defined as the percent increase in one year all cause mortality associated with each clinical variable in the model.[15] A polynomial regression

line was then fit to the data to create an "excess risk function" for each clinical variable. The final RI is essentially the sum of these excess risk functions. The model also allows for infrequently collected lab test values via creating a model for no labs and a model with labs. The RI score is calculated every time a new model input is available, and thus is in "real time" and takes 26 clinical variables, including lab values. RI has been shown to be correlated with mortality as well as 30 day readmission. While RI can be calculated in real time, it is still limited by the time lag of specific variables such as lab values and nursing assessments that are infrequent and irregular and the prediction is not actually predicting any specific in hospital complications or in hospital mortality.

There is also work being done on more specific complications such as reintubation as well as acute kidney injury. Acute kidney injury develops in about 5% of hospitalized patients. Acute kidney failure has been shown to increase cost, length of stay, as well as mortality. This highlights the need for accurate prediction of AKI for early diagnosis and treatment. There have been risk scores developed to predict acute kidney injury following surgery such as one developed by Thakar et al.[16] Similar to the POSPOM, Thakar's AKI risk score used logistic regression model to select the most significant features and those features are assigned points based on the the regression coefficients. Score points were calculated as the regression coefficient multiplied by 2 and rounded to the nearest integer. The final AKI risk score is a sum of these points. This model, like POSPOM, also uses only preoperatively available information such as comorbidities like COPD and diabetes, as well as surgery type and preoperative creatinine. However, it was specifically developed on and for cardiac patients. Another AKI score, but one for non-cardiac surgery patients, was developed also using only preoperative

patient information.[17,18] Following development of a logistic regression model, risk scores were assigned similar to POSPOM and Thakar et al. with the regression coefficient normalized to the smallest coefficient then multiplied by 2 and rounded. This study also assessed intraoperative hypotension as a potential variable for the risk of AKI. Intraoperative hypotension features included amount of time blood pressure was less than an absolute hypotension cutoff such as SBP < 80 mmHg, SBP < 70, as well as intraoperative vasopressor administration and urine output. No specific amount or duration of hypotension was found to be associated with AKI in this study, however in other studies intraoperative hypotension has been shown to highly correlate with post operative complications such as cardiac death, pulmonary edema, mortality, and excess length of stay.[19–24] This highlights the potential for using prevention of hypotension as a continuous, intraoperative way to change decision and management to improve postoperative outcomes.

Apart from AKI, other postoperative complications of concern are respiratory ones such as pneumonia, failure to wean, and post extubation respiratory failure, which have been shown to be the second most frequent type of postoperative complication after wound infection.[6,24,25] Post extubation respiratory failure has been shown to increase poor outcomes and mortality.[6] Thus, being able to predict which patients are at highest risk of post extubation respiratory failure is clinically important. A preoperative risk score to predict risk of postoperative reintubation has been developed combining only 5 features (ASA>3, emergency procedure, high risk service, congestive heart failure, and chronic pulmonary disease).[26] Similar to other models, the Score for Prediction of Postoperative Respiratory Complications (SPORC) also uses a logistic regression model's coefficients to assign points to each feature which are then

summed as the final risk score. In addition to the current work in complications risk, there have also been efforts to predict risk of readmission in specific cohorts of patients such as congestive heart failure, cancer, and emergency.[27,28] The Preadmission Readmission Detection Model (PREADM) was developed as a logistic regression model to predict 30 day readmission using 11 variables that included chronic conditions, prior health services uses, BMI, and geographical location (PREADM). Again, similarly to other models discussed here, the regression coefficients were transformed into scoring points.

There is also current research in utilizing other methods apart from logistic regression. One study compared using support vector machines, logistic regression, decision trees, random forest and generalized boosted modeling for predicting both hospital readmission as well as cost of that readmission, and found that all methods' results were comparable.[29] Variables used in that study were mainly patient demographic and admission information such as age, ethnicity, admission type, number of co morbidities, length of stay, with the only vital sign being blood pressure at discharge. Another study utilized random forests to forecast cardiorespiratory instability, using continuous and high resolution vital signs such as heart rate, respiratory rate, and blood pressure.[30] Deep neural networks, or deep learning, is also becoming a popular approach. One study utilized temporal convolutional neural networks on lab values to predict onset of diseases such as atrial fibrillation and chronic kidney disease.[31] Another applied LSTM recurrent neural networks to predict mortality and number of ventilator free days.[32] They used static features such as demographics and admission diagnosis as well as temporal features such as ventilator settings and blood gas values). Nguyen et al. developed Deepr (Deep net for medical Record), a convolutional neural network that takes the electronic

clinical notes from patient visits to predict future outcomes (specifically diagnoses and procedures).[33] Lipton et al. utilized LSTM recurrent neural networks to recognize patterns in multivariate time series of clinical measurements (vital signs as well as labs) to classify diagnoses.[34]

The ultimate goal of all the above current research is to change modern medicine to becoming more prospective or proactive, rather than reactive. There are 2 ways to think of prospective healthcare: 1) A one-time risk classification to help allocate hospital resources more efficiently to ensure patients receive necessary critical care and 2) Continuous, real time risk classification to avoid onset of complications altogether. One-time risk classification would be to better predict which patients are at risk of which bad outcomes or complications. This would be to help stratify patients prior to care and better allocate hospital time and resources. Continuous risk classification would be similar to a new continuous vital sign for a patient, i.e. an arterial blood pressure signal that outputs blood pressure values every 20 seconds. The point of a continuous risk indicator would be to predict short term onset of adverse events such as atrial fibrillation. We believe that the best way to do both is through deep learning.

Introduction to Deep Neural Networks

Deep neural networks, aka deep learning, is a currently popular approach in machine learning. The aim of deep learning is to learn from raw data and perform desired tasks without any feature engineering. In other words, deep learning is capable of modeling the complex nonlinear and linear features from low level features or raw data that are necessary for an accurate output. Current deep learning is mostly based on multilayered neural networks, where

each layer is connected via neurons.[35] Each neuron applies a nonlinear transform to a linear function of inputs. This is referred to as an activation function and the most commonly used ones are sigmoid, tanh, and the Rectified Linear Unit (ReLU). The sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ has a range of 0 to 1. The tanh nonlinearity has a range of -1 to 1 and can be considered a scaled sigmoid neuron where $\tanh(x) = 2\sigma(2x) - 1$. ReLU simply thresholds a value at zero via the function $f(x) = \max(0, x)$. Deep learning has been developed for decades, but over the past few years it has broken records in visual object recognition, speech recognition, and natural language.[36] There are three main types of deep neural networks: feedforward, recurrent, and convolutional. Feedforward networks pass information from one end to the other, usually input to output, and can be thought of as universal function approximators.[37] Recurrent neural nets (RNNs) model varying length sequential data (sequential over time or space) and can maintain some form of memory and capture long term dependencies.[38] RNNs selectively pass information across sequential steps, while processing sequential data one element at a time. Convolutional neural nets (CNNs) exploit local motifs across time and space, small pieces of data with predefined sizes such as a batch of pixels. CNNs contain convolutional layers that aim to learn feature representations of the inputs and compute different feature maps.[39]

UCI ANESTESHIOLOGY RESEARCH DATABASE

All data collected in this study was obtained with IRB approval from UC Irvine, and was a collaborative effort with Edwards Lifesciences (Irvine, CA), CardioPulmonary Corporation/Bernoulli (Milford, CT) and the Department of Anesthesiology at UC Irvine Medical

Center (UCIMC). All data in this effort were deidentified prior to access by IRB approved researchers.

Creating the Database

Starting in August of 2015, all adult surgical patients presenting at all nineteen UCI operating suites have been being consented for data collection. This database consists of three types of clinical data: high resolution waveforms, intermittent values, and clinical anesthesia record. We are currently collecting three types of high resolution waveforms: arterial blood pressure, EKG, and plethysmography directly from the GE (General Electric Healthcare, Chicago, IL) patient monitors (B850 and Solar 8000) in the operating room (OR). Intermittent values are collected as standard of care and are contained in the anesthesia medical record. Intermittent values consist of the standard vital signs such as heart rate and blood pressure, but also include ventilator values such as end tidal CO2 and PEEP. Intermittent values also consist of any manually input lab values such as hemoglobin. The anesthesia record also consists of all medical record data associated with the surgical case, including patient demographics as well as drug and fluid interventions. Combining all three types of clinical data makes this database unique and novel. Data collection currently remains ongoing with IRB renewal.

Contents of the Database

The data available in the UCI database was collected from two sources: the intraoperative electronic medical record (EMR) and directly from the GE bedside monitor (Figure 1). All EMR data was pulled retrospectively once a week by the UC Irvine Medical

Center (UCIMC) Honest Broker and pushed to a data processing server. All waveforms were collected by Bernoulli systems from bedside monitors and pushed to the data processing server before being aggregated with EMR data, post processed, and deidentified. These data can be separated into specific classes summarized in Table 1.

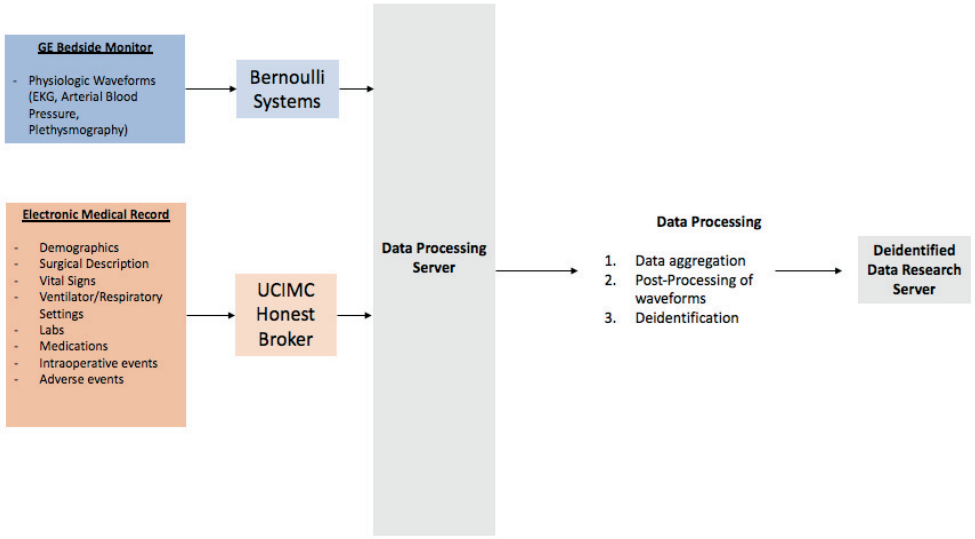


Figure 1. *Overview of data collection points*

Table 1. *Description of data classes and sources found in the UCIMC Anesthesiology Research Database*

Data Class	Source	Description
Descriptive	EMR	Demographic detail, ASA score, admission type, surgical description
Events	EMR	All manually annotated intraoperative events (e.g. induction, intubation) and physician comments
Adverse Events	EMR	All manually annotated intraoperative adverse events or complications
Medications	EMR	All manually annotated administered medications and fluids
Manual Observations	EMR	All manually input observations including intraoperative blood gas values, estimated blood loss, and urine output

Automated Observations	EMR	All automatically collected 1 minute observations such as intermittent vital signs (e.g. heart rate, blood pressure), ventilator settings (e.g. tidal volume, respiratory rate)
Physiologic Waveforms	Bedside monitor	All automatically collected, available waveforms (EKG, Plethysmographic, Arterial Blood Pressure)

Electronic Medical Record (EMR) Data

All EMR data was obtained from the intraoperative EMR system Surgical Information Systems (SIS) and pushed to the data server by a UCIMC once per week. All data was organized to replicate the SIS EMR structure of 6 unique EMR classes, which were saved as individual csv files per patient (Table 2).

Table 2. *Description of the 6 EMR classes of data pulled from the UCIMC Surgical Information Systems (SIS) per patient.*

EMR Type	Description
Patient	• Free text surgery description • Admission type • Height, weight, age, sex • American Society of Anesthesia (ASA) Score • Surgery and anesthesia start and stop times
Category	• Adverse events that occurred intraoperatively
Events	• Comments by the anesthesiologist during surgery • All standard anesthesia events such as intubation, induction, arterial line placement, positional changes etc. with timestamps
Drugs and Fluids	• All start and end (if available) timestamps annotated by the clinician • Medication name • Volume administered or rate of administration
Input/Output	• All Estimated Blood Loss and Urine Output values annotated by the clinician with time stamps • Total sum of volumes of fluids administered
Observations	• All 1 minute sampled vitals and ventilator information with timestamps • All manually input vitals with timestamps • All manually input lab values with timestamps

Physiologic Waveform Data

All physiologic waveform data was collected from the bedside monitors using Bernoulli systems. Every operating room (OR) of UCIMC currently contains one of the following bedside monitors: GE Solar 8000 or GE B850. The Bernoulli system acquires a maximum of 2 (GE Solar 8000) or 3 (GE B850) waveform channels directly from the monitors, and all waveform data per OR was saved into XML files. The newer GE B850 monitors output all 3 waveform channels: EKG sampled at 300 Hz, plethysmograph sampled at 100 Hz, and invasive arterial blood pressure (if available) sampled at 100 Hz. The older GE Solar 8000 monitors have an analog output and output only 2 waveform channels: EKG and invasive arterial blood pressure (if available). It should be noted that due to the analog output of the GE Solar monitors, the sampling rate was approximated in post-processing for corrected time alignment with EMR data. All waveform data from each OR were transmitted through the hospital network to a data processing server, where it was temporarily stored until the EMR data was available on the data server for data processing as described in Figure 1.

EMR and Waveform Data Postprocessing

The surgery start and stop times and OR location from the EMR data were used to parse and match waveform data to the correct patient. In the EMR data, unique surgeries are identified by a unique Case Confirmation Number (CCN). The CCN is unique to the surgical case, while a medical record number is unique to the patient, i.e. a patient can have only 1 MRN but multiple CCNs. We chose to keep each surgical case unique and used the CCN to identify

patients and their associated data. Once the waveforms and EMR data were matched and aggregated, all data was assigned a new unique deidentified ID.

The waveform data is parsed into 30 minute XML files per patient. To be more user-friendly, the waveform data was processed in a secondary step. First, all the waveform data was translated from 16 bit data. Due to hardware and monitor limitations, there also exists data gaps. In the GE B850, these are annotated as flags by the monitor itself. However, in the GE Solar 8000, the data gaps are identified via an algorithm and assumed to be true. After the waveforms are corrected for data gaps they are also corrected for gain and sampling frequencies to be time synced with the EMR data. Waveform data are saved in a .bin file format that includes patient sex, age, height, weight, Body Surface Area (BSA), sampling frequency, and start timestamp (in serial format).

Deidentification and HIPPA Compliance

All patient identifiers such as name, medical record number (MRN) and social security number (SSN) were completely removed and all birthdates were replaced with age at the date of surgery prior to being made available to IRB-approved researchers. Before making a database to be accessible by a larger group of researchers, we intend to do further deidentification of the data. This includes the removal and replacement of all timestamps with timestamps that are shifted by a random offset, and the removal of hospital resources, such as OR location.

Patient Characteristics

The first version of the UCI database includes data from all distinct surgeries from UC Irvine Medical Center (UCIMC) performed between 2015 and 2017. In 2017, UCIMC transitioned from Surgical Information Systems to Epic (Verona, WI) for their intraoperative EMR. Additional data collection is ongoing, however not merged or processed. Table 3 provides a summary of the patient population.

Table 3. *Summary of patient population in UCIMC Research Database 2015 – 2017. Data is represented in mean ± standard deviation, unless otherwise noted.*

# of OR Days	630	
# of Patients	19,636	
ASA	**# Patients**	**% of Patients**
1	561	2.86
1E	78	0.40
2	6,175	31.45
2E	288	1.47
3	9,455	48.15
3E	446	2.27
4	2,039	10.38
4E	386	1.97
5	20	0.10
5E	155	0.79
6	27	0.14
6E	6	0.03
Age (years)	52 ± 19	
Gender	**# Patients**	**% of Patients**
Female	9,985	50.85
Male	9,645	49.12
Other	2	0.01
Unknown	4	0.02
Admission Type	**# Patients**	**% of Patients**
Inpatient	7,630	38.86
23 Hour Observation	2,165	11.03
AM Admission	4,491	22.87
Outpatient	5,189	26.43
Midnight Admission	58	0.30
Day Prior Admission	93	0.47
Unknown	10	0.05
Anesthesia Type	**# Patients**	**% of Patients**
General	17,629	89.78
MAC	1,693	8.62

Regional Block	104	0.53
Spinal	139	0.71
Combined Spinal/Epidural	17	0.09
Combined General/Epidural	10	0.05
Epidural	11	0.06
Local	4	0.02
Bier block	9	0.05
None	18	0.09
Unknown	2	0.01
Total Anesthesia Time	3 hours 20 minutes ± 2 hours 13 minutes	

Data Mining Tools

The goal of this data effort is to promote research. To make the UCI database more user friendly, we have created definition tables for the EMR classes Events, Observations, and Drugs (Table 4, 5, 6, respectively), as well as various data mining resources. The definition tables contain all the unique possible item names found in each EMR class, with metadata in the form of number of times the item is present as well as number of unique patients with the item name. For the Observations class, we included a long label provided by UCIMC SIS as well as manually assigned description of the item name (Table 5). For the Drugs class, we included the manually assigned common drug types (i.e. analgesic, vasopressor, vasodilator, etc.) as well as the unique units found for the specific item name (Table 6). We have also created data mining resources for querying the database based on these definition tables as well as visualization tools. An example of this is shown in Figure 2 below.

Table 4. *Sample subset of definition table for EMR Events.*

Event Name	Event Counts	Patient Counts
MD Maintenance	16130	7038
Anesthesia Positioning Note	10942	7754
Rhythm	10286	7829
Anesthesia Note	9624	4115
SBar Time	8190	7851
Report to RN	8050	7891
Temperature Management	8046	7752
Patient Positioning	7938	7887
Anesthesia Time	7891	7891
Surgery Time	7891	7891
Time out completed	7891	7891
Patient Transport Note	7890	7890
Pre-Induction Patient Safety	7886	7886
OR Time	7861	7861

Table 5. *Sample subset of definition table for EMR Observations.*

Observation Name	Long Label	Description	Observation Counts	Patient Counts
SpO2	Saturation Pulse Oximetry	Measurement of oxygen saturation at periphery	1398284	7854
HR (EKG)	Anesthesia	Heart rate	1378752	7852
NIBP SYS	Non-invasive Blood Pressure Systolic	Non-invasive systolic arterial pressure	295526	7848
NIBP DIA	Non-invasive Blood Pressure Diastolic	Non-invasive diastolic arterial pressure	295431	7847
HR (SpO2)	Anesthesia	Heart rate	1387764	7846
RR	Set Rate Ventilator DC1320	Respiratory rate	1120342	7837
T1		Temperature	887261	7835
ETCO2	End Tidal Carbon Dioxide Amount - Capnometer Wave	End tidal carbon dioxide concentration	1422340	7790
FiO2	Gas Monitor DC3424	Inspired oxygen concentration	1420074	7788

Table 6. *Sample subset of definition table for EMR Drugs.*

Drug Name	Drug Category	Drug Counts	Unique Units	Patient Counts
FENTANYL	analgesic - narcotic	23565	'MICROgm, ML'	7274
PROPOFOL	anesthetic	13642	'MICROgm, mg'	6775
MIDAZOLAM	benzodiazepine	6451	'mcg, mg'	5862
ONDANSETRON	antiemetic	5582	'mg'	5381
LIDOCAINE	analgesic - local	4855	'mL, mg, ml'	4755
GLYCOPYRROLATE	anticholinergic	4446	'mg'	3771
PLASMALYTE	crystalloid	8655	'mL'	3713
PLASMALYTE	crystalloid	8655	'mL'	3713
DEXAMETHASONE	glucocorticoid	3760	'MG, mg'	3674
CEFAZOLIN	antibiotic	4051	'g, gm, mg'	3535
NEOSTIGMINE	acetylcholinesterase inhibitor; paralytic reversal	3474	'mg'	3402
PHENYLEPHRINE	vasopressor	12251	'MICROgm, Mcg, mg'	3114
EPHEDRINE	vasopressor	7723	'mg'	3042
SUCCINYLCHOLINE	paralytic	2881	'mg'	2832
LACTATED RINGERS	crystalloid	4926	'mL'	2819

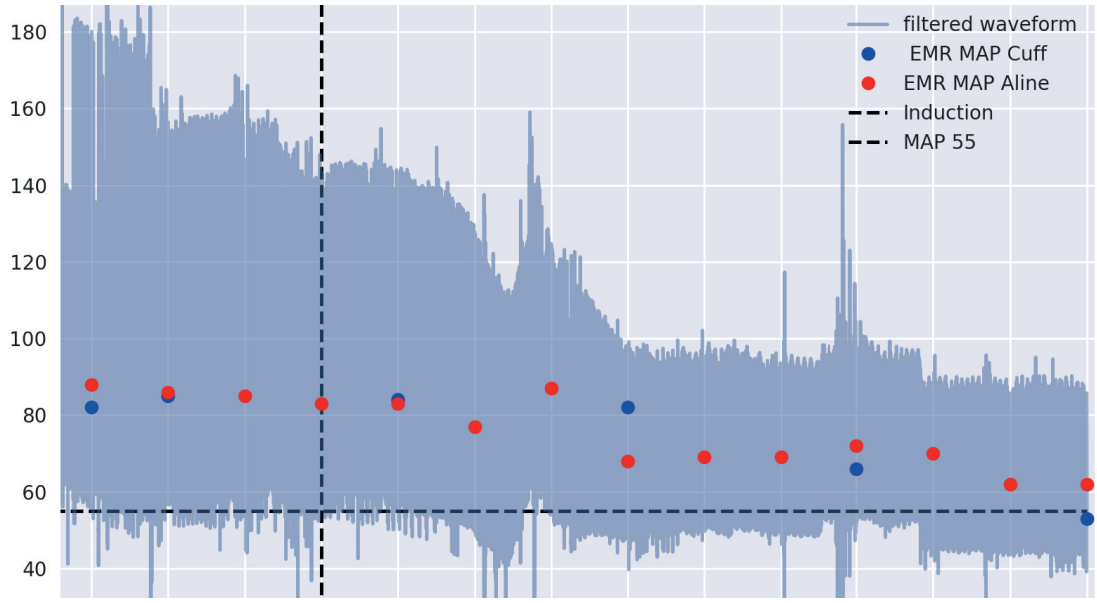


Figure 2. *Example of a UCI surgical patient's arterial blood pressure waveform plotted with noninvasive blood pressure cuff (MAP cuff) and invasive arterial blood pressure (MAP Aline) measurements from the EMR Observations, as well Induction from the EMR Events.*

PREDICTING POSTOPERATIVE IN-HOSPITAL MORTALITY

About 230 million surgeries are performed annually worldwide.[40] While the estimated postoperative mortality is low, less than 2%, studies have shown that about 12% of all patients - the high-risk surgery group - account for 80% of postoperative deaths. To assist in guiding clinical decisions and prioritization of care, several perioperative clinical and administrative risk scores have been proposed. These scores tend to be subjective like the American Society of Anesthesiologists (ASA) physical status score (a preoperative score) or are developed using simple methods like logistic regression, such as with the Preoperative Score to Predict Postoperative Mortality (POSPOM).[9,14]

In collaboration with UCLA Medical Center's Department of Anesthesiology, our first deep neural network (DNN) models were created for predicting in hospital mortality. Performance is presented together with other published clinical risk scores (ASA, Surgical Apgar, POSPOM) and administrative risk scores (Risk Stratification Index and Risk Quantification Index), as well as a logistic regression model using the same intraoperative features as the DNN.[9,10,14,41–43] The DNNs were also assessed for leveraging preoperative information by the addition of ASA and POSPOM as features. This work has been published.[44]

Data Description

Electronic Medical Record (EMR) Data Extraction

All data for this study were extracted from the Perioperative Data Warehouse (PDW), a custom built robust data warehouse containing all patients who have undergone surgery at

UCLA since the implementation of the electronic medical record (EMR) (EPIC Systems, Madison WI) on March 17th, 2013. The construction of the PDW has been previously described.[45]All data used for this study were obtained from this data warehouse and IRB approval was 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 PDW. The UCLA Health System includes two inpatient medical centers as well as 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 years of age or older than 89 years 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 EMR 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 7). These features were considered to be potentially predictive of in-hospital mortality by clinicians' consensus and included descriptive intraoperative vital signs, such as minimum and maximum blood pressure values;

summary of drugs and fluids interventions such as total blood infused and total vasopressin administered; as well as patient anesthesia descriptions such as presence of an arterial line and type of anesthesia (all features are detailed in Table 7).

***Table 7**. Description of model input features and applied maximum possible values as defined by domain experts.*

Feature Name(s)	Description	Maximum Possible Absolute Value (if applicable)
COLLOID_ML*	Total Colloid Transfused (ml)	-
CRYSTALLOID_ML*	Total Crystalloid Transfused (ml)	-
DBP MAX*, MIN*, AVG, MED, STD	Maximum, Minimum, Average, Median, and Standard Deviation Diastolic Blood Pressure for the case (mmHg)	150
DBP_10min MAX, MIN, AVG, MED, STD	Maximum, Minimum, Average, Median, and Standard Deviation Diastolic Blood Pressure for the last 10 minutes of the case (mmHg)	150
EBL*	Total Estimated Blood Loss (ml)	-
EPHEDRINE BOLUS*	Total bolus dose of Ephedrine (mg) during the case	-
EPINEPHRINE BOLUS*, END RATE*, MAX RATE*	Total bolus dose (mcg), End of case infusion rate (mcg/kg/min), and Highest infusion rate (mcg/kg/min) of Epinephrine during the case	-
ESMOLOL BOLUS*, END RATE*, MAX RATE*	Total bolus dose (mg), End of case infusion rate (mcg/kg/min), and Highest Infusion rate (mcg/kg/min) of Esmolol during the case	-
HR MAX*, MIN*, AVG, MED, STD	Maximum, Minimum, Average, Median, and Standard Deviation Heart Rate (bpm) for the case	180
HR_10min MAX, MIN, AVG, MED, STD	Maximum, Minimum, Average, Median, and Standard Deviation Heart Rate (bpm) for the last 10 minutes of the case	180
INVASIVE_LINE_YN*	Invasive Central venous, arterial, or Pulmonary Arterial Line used for the case (Yes/No)	-
MAP MAX*, MIN*, AVG, MED, STD	Maximum, Minimum, Average, Median, and Standard Deviation Mean Blood Pressure (mmHg) for the case	300
MAP_10min MAX, MIN, AVG, MED, STD	Maximum, Minimum, Average, Median, and Standard Deviation Mean Blood Pressure (mmHg) for the last 10 minutes of the case	300

DES MAX*	Maximum Minimum alveolar concentration of desflurane during the case (note this is not age adjusted)	12
GLUCOSE MAX*, MIN*	Maximum and Minimum plasma Glucose concentration for the Case (mg/dl)	400
ISO MAX*	Maximum Minimum alveolar concentration of isoflurane during the case (note this is not age adjusted)	12
SEVO MAX*	Maximum Minimum alveolar concentration of sevoflurane during the case (note this is not age adjusted)	10
MILRINONE END RATE*, MAX RATE*	End of case Infusion Rate and Highest Infusion rate of Milrinone during the case (mcg/kg/min)	-
HGB MIN*	Minimum Hemoglobin concentration (g/dl) during the case	15
MINUTES MAP < 50	Cumulative minutes with mean arterial pressure <50 mmHg (min)	-
MINUTES MAP < 60	Cumulative minutes with mean arterial pressure < 60 mmHg (min)	-
NICARDIPINE END RATE*, MAX RATE*	End of case infusion Rate and Highest Infusion Rate of Nicardipine during the case (mg/hr)	-
NITRIC_OXIDE_YN*	Nitric Oxide Used for the Case (Yes/No)	-
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	-
NITROPRUSSIDE END RATE*, MAX RATE*	End of case infusion Rate and Highest Infusion Rate of Nitroprusside (mcg/kg/min) during the case	-
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	-
SBP MAX*, MIN*, AVG, MED, STD	Maximum, Minimum, Average, Median, and Standard Deviation Systolic blood pressure (mmHg) for the case	300
SBP_10min MAX, MIN, AVG, MED, STD	Maximum, Minimum, Average, Median, and Standard Deviation Systolic blood pressure (mmHg) for the last 10 minutes of the case	300
SpO2 MAX*, MIN*, AVG, MED, STD	Maximum, Minimum, Average, Median, and Standard Deviation SpO2 (%) for the case	100
SpO2_10min MAX, MIN, AVG, MED, STD	Maximum, Minimum, Average, Median, and Standard Deviation SpO2 (%) for the last 10 minutes of the case	100
UOP*	Total Urine Output (ml)	-
VASOPRESSIN BOLUS*, END RATE*, MAX RATE*	Total bolus dose (units), End of case infusion rate (units/hr), and Highest Infusion rate (units/hr) of Vasopressin during the case	-

XFUSION_RBC_ML*	Total Red Blood Cells Transfused (ml)	-

Data Preprocessing

One of the biggest issues with this data and other clinical data is missing values. For example, minimum hemoglobin during surgery could be missing for a specific patient. This missing value is not due to error, but rather that the anesthesiologist felt the patient was normal and no blood samples were taken because the patient did not need one. It should also be noted that the fact that the variable is missing is important information in itself and a different way to address these gaps would be to include binary variables that indicate whether or not a value was missing. Another way to address these gaps for future work would be to fill them with clinically normal values as defined by domain experts.

Prior to model development, missing values were filled with the mean value for the respective feature, or filled with the most common value or zero (Table 8). 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 7). These out of range values were due to the data artifact in the raw EMR data. For example, a systolic blood pressure of 400 mmHg is not clinically possible, however, it may be recognized as the maximum systolic blood pressure for the case during EMR extraction. The data was then randomly divided into training (80%) and test (20%) data sets, with equal % occurrence of in-hospital mortality. Training data was rescaled to have a mean of 0 and standard deviation of 1 per feature. Test data was rescaled with the training data mean and standard deviation.

Table 8. *Description of missing value preprocessing per feature.*

Feature	Number Patients With Missing Data	Data Fill Type	Mean Value
MAX_ISO	57230	0	1.3
CURRENT_HB	51659	Mean	10.7
MIN_HB	51512	Mean	10.2
MAX_GLUCOSE	51211	Mean	161.0
MIN_GLUCOSE	51211	Mean	121.0
MAX_DES	46961	0	1.0
STARTING_HB	26115	Mean	12.3
MAX_SEVO	20017	0	1.2
BASELINE_GFR	16599	Mean	86.5
MAX_MAP	1510	Mean	114.3
MIN_MAP	1510	Mean	60.2
MIN_MAP_LT_40	535	Mean	0.5
MIN_MAP_LT_45	535	Mean	1.1
MIN_MAP_LT_50	535	Mean	2.6
MIN_MAP_LT_55	535	Mean	6.8
MIN_MAP_LT_60	535	Mean	16.7
MIN_MAP_LT_65	535	Mean	33.6
MAX_PULSE_OX	212	Mean	99.9
MIN_PULSE_OX	212	Mean	91.3
MAX_SBP	207	Mean	164.4
MIN_SBP	207	Mean	79.4
MAX_DBP	207	Mean	94.5
MIN_DBP	207	Mean	43.8
MAX_HR	194	Mean	109.1
MIN_HR	194	Mean	55.7
ASA_SCORE	22	Most common ASA Score 3	2.6

Model Development

In this work, we were interested in classifying patients at risk of in-hospital mortality using deep neural networks (DNNs), also referred to as deep learning. During development of DNNs, there are many unknown model parameters that need to be optimized by the DNN 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 DNN 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 [0-1]. To develop a DNN, it is important to fine-tune the hyperparameters as well as the architecture. We utilized stochastic gradient descent (SGD) 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 DNN architectures of 1 to 5 hidden layers with 10 - 300 neurons per layer, and rectified linear unit (ReLU) and hyperbolic tangent (tanh) 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%) prior to testing final model performance on the separate test set (20%).

Overfitting

While ~50,000 examples is large for clinical data, it is small relative to datasets found in deep learning tasks like vision and speech recognition where millions of examples are available. Thus, overfitting was a major concern and regularization is critical. The first and most obvious solution to this would be to just collect more data. This is currently being addressed by data collection efforts at UC Irvine, but to collect more data at a large enough scale can take years, as it is limited by the number of patients that come through the hospital. Thus, early stopping, L2 weight decay, and dropout were all used to address overfitting. Early stopping is used during the training process. A loss function is calculated after each epoch on a validation set and once the validation loss starts to increase, indicating overfitting, training is stopped. The point at which to stop training depends on a “patience” parameter, corresponding to the number of

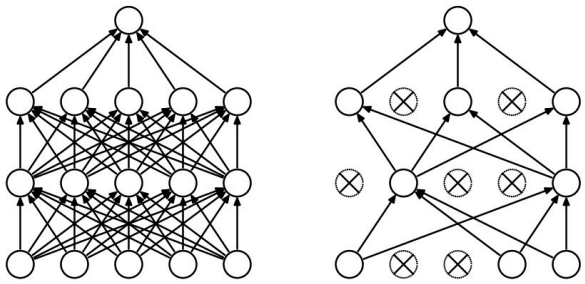
epochs to wait for to see if validation loss continues to increase. For all models, the patience was set to 10. L2 weight decay is a method of limiting the size of the weights. The standard L2 weight penalty involves adding an extra term to the loss function that penalizes the squared weights, keeping the weights small unless the error derivative is big. The loss function used for all models was log loss, also known as cross entropy loss. Log loss is defined by

$$L = -\frac{1}{n}\sum_{x} y \ln a + (1-y)\ln(1-a)$$

where n is the total number of examples in the training data, the sum is over all the training inputs x and y is the corresponding desired output and a is the calculated output. The output used for all models is sigmoid. If L is the loss function, then the new loss function with L2 penalty is the following, where i indicates the i-th example:

$$C = L + \frac{\lambda}{2}\sum_{i} w_i^2$$

We utilized an L2 weight penalty of 0.0001. Dropout is a relatively new way to deal with the limited data as compared to the large number of learning parameters seen with deep neural networks [57]. Neurons are removed from the network with a specified probability during training. This prevents neurons from co-adapting too much. The procedure is repeated for each example at each training epoch. After training is complete, predictions are produced by multiplying the weights by the specified dropout probability. Dropout was only applied before the output layer. The following figure describes drop out and was taken from paper by Srivastava et al.[46] Dropout was applied to all layers with a probability of 0.5.



(a) Standard Neural Net (b) After applying dropout.

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.

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 was skewed. The % occurrence of mortality in the training dataset was < 1%. To help with this skewed distribution, training data was 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 standard deviation 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 prior to 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 and POSPOM as a feature were also

conducted. The reduced 45 feature set was created by excluding all "derived" features, specifically average, median, standard deviation, and last 10 minutes of the surgical case features (Table 7).

After choosing the best performing DNN architecture and hyperparameters with the complete 87 features data set, five additional DNNs were each trained with the following:

1) the addition of ASA as a model feature (88 features)

2) the addition of POSPOM as a model feature (88 features)

3) a reduced model feature set (45 features)

4) the addition of ASA to the reduced feature set (46 features).

5) the addition of POSPOM to the reduced feature set (46 features).

Model Performance Methods

All model performances were assessed on 20% of the data held out from training as a test set. Model performance was compared to ASA, Surgical Apgar, RQI, RSI, POSPOM, and a standard logistic regression model using the same combination of features as in the DNN. ASA was extracted from the UCLA preoperative assessment record. Surgical Apgar was calculated using Gawande et al.[10] RQI could not be calculated using the downloadable R package from Cleveland Clinic's website < http://my.clevelandclinic.org/departments/anesthesiology/depts/outcomes-research> due to technical issues with the R version, and so RQI log probability and score were calculated from equations provided in Sigakis et al.[43] Uncalibrated RSI was calculated using coefficients provided by the original authors. To calculate RSI, all ICD-9 diagnosis codes for each patient

were matched with an RSI coefficient and the coefficients were then summed. POSPOM scores were extracted from the PDW, where they were calculated as described by Le Manach et al.[14] Each of the diseases described by Le Manach et al. were extracted as a binary endpoint from the admission ICD codes for the relevant hospital admission. In addition to assigning points based on patient co-morbidities the POSPOM also assigns points for the type of surgery performed. These points were assigned based on the primary surgical service for the given procedure.

Model performance was assessed using Area Under the ROC Curve (AUC) and 95% confidence intervals for AUC were calculated using bootstrapping with 1,000 samples. The F1 score, sensitivity, and specificity were calculated for different thresholds for the DNN models, logistic regression model, ASA, and POSPOM. 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 DNN. This data augmentation served to balance classes in the training data set to approximately 45% mortality vs the true distribution of mortality (<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 (<1%). Therefore, we performed calibration after finalizing the model. Calibration was performed only on the test data set. Calibration of the DNN 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 (AUC, sensitivity, specificity, or F1 score).

In addition, calibration plots and Brier scores were used to assess calibration of predictions.

Feature Importance

To assess which features are the most predictive in the DNN, we performed a feature ablation analysis. This analysis consisted of removing model features grouped by type of clinical feature, and then re-training a DNN with the same final architecture as well as hyperparameters on the remaining features. The change in AUC 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 DNN models were developed and applied using Keras.[47] Logistic regression models and performance metrics were calculated with scikit-learn.[48]

Results

The data consisted of 59,985 surgical records total. Patient demographics and characteristics of the training and test data sets are summarized in Table 9. 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 is 3 at 49.9% for both sets.

Table 9. *Description of patient demographics*

		Train	Test
# Patients		47985	11996
Age		56 +/- 17	56 +/- 94
EBL		96 +/- 539	18 +/- 410
# with Aline		8583	2135
# with PA		1641	430
# with CVC		2443	635
ASA Score			
	1	3022	762
	2	17930	4477
	3	23960	5985
	4	2910	735
	5	144	30
	6	4	0
	Unknown	15	7
Primary CPT by Specialty			
	Gastroenterology	6615 (13.8%)	1614 (13.5%)
	General Surgery	6552 (13.7%)	1646 (13.7%)
	Urology	4005 (8.3%)	1062 (8.9%)
	Orthopaedics	3916 (8.2%)	979 (8.2%)
	Neurosurgery	3686 (7.7%)	916 (7.6%)
	Otolaryngology	3268 (6.8%)	860 (7.2%)
	Obstetrics and Gynecology	2630 (5.5%)	672 (5.6%)
	Vascular Surgery	1834 (3.8%)	445 (3.7%)
	Cardiac Surgery	1396 (2.9%)	372 (3.1%)
	Thoracic Surgery	1095 (2.3%)	273 (2.3%)
	Other	8497 (17.7%)	2049 (17.1%)
	Unknown	4491 (9.4%)	1108 (9.2%)
AKI			
	Class 1	2501 (5.21 %)	622 (5.19 %)

	Class 2	369 (0.77 %)	99 (0.83 %)
	Class 3	1001 (2.09 %)	246 (2.05 %)
	Null	30616 (63.8 %)	7689 (64.1 %)
Reintubation		548 (1.14 %)	159 (1.33 %)
Mortality		389 (0.81 %)	87 (0.73 %)

The final DNN architecture consists of four hidden layers of 300 neurons per layer with rectified linear unit (ReLu) activations and a logistic output. The DNN 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). ROC curve and AUC results are shown in Figure 2. All logistic regression models (LR) and all DNNs had higher AUCs than POSPOM (0.74 (95% CI, 0.68 – 0.79)) and Surgical Apgar (0.58 (95% CI, 0.52 – 0.64)) for predicting in-hospital mortality (Figure 3). All DNNs had higher AUCs than LRs for each combination of features except for the reduced feature set with POSPOM (LR 0.90 (95% CI, 0.86 – 0.93) vs DNN 0.90 (95% CI, 0.87 – 0.93)). In addition, reducing the feature set from 87 to 45 features did not reduce the DNN model AUC performance and the addition of ASA and POSPOM as features modestly improved the AUCs of both the full and reduced feature set DNN models. The highest DNN AUC result was the DNN with reduced feature set and ASA (0.91 (95% CI, 0.88 – 0.93)). The highest clinical risk score AUC was RSI (0.97 (95% CI, 0.94 – 0.99)) and the highest LR AUCs were the LR with reduced feature set and ASA (0.90 (95% CI, 0.87 - 0.93)) and the LR with reduced feature set and POSPOM (0.90 (95% CI, 0.86 - 0.93)).

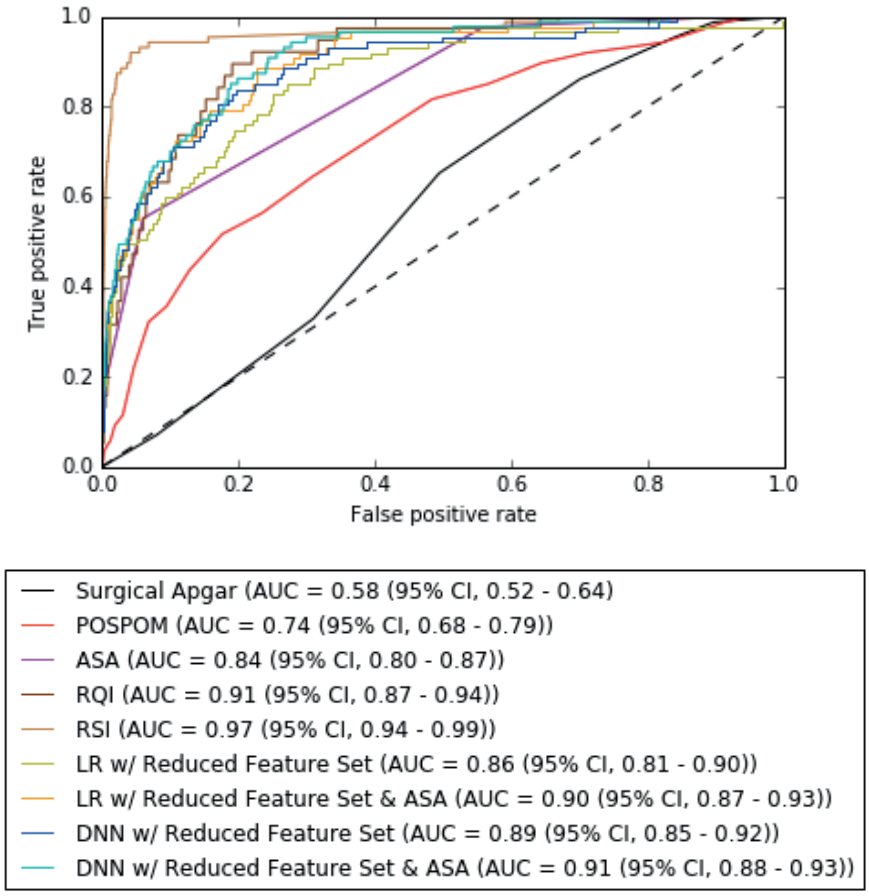


Figure 3. *ROC Curve and AUC (95% CI) results for in-hospital mortality models and scores.*

For comparison of F1 scores, sensitivity and specificity at different thresholds, DNN with original 87 features (DNN), DNN with a reduced feature set and POSPOM ($DNN_{rfsPOSPOM}$), and DNN with a reduced feature set and ASA (DNN_{rfsASA}) are compared to ASA, POSPOM, logistic regression with original 87 features (LR), logistic regression with a reduced feature set and POSPOM ($LR_{rfsPOSPOM}$), and logistic regression with a reduced feature set and ASA (LR_{rfsASA}) (Table 4). If we choose a threshold that optimizes the observed in-hospital mortality rate, the thresholds (% observed mortality) for POSPOM, ASA, and LR, $LR_{rfsPOSPOM}$, LR_{rfsASA} are 10 (93.1%),

3 (97.7%), 0.00015 (98.9%), 0.002 (97.7%), and 0.0034 (96.66%), respectively. The thresholds for DNN, $DNN_{rfsPOSPOM}$ and DNN_{rfsASA} are 0.05 (98.9%), 0.2 (96.6%) and 0.22 (96.6%), respectively. At these thresholds, POSPOM, ASA, LR, $LR_{rfsPOSPOM}$, LR_{rfsASA}, DNN, $DNN_{rfsPOSPOM}$ and DNN_{rfsASA} , all have high and comparable sensitivities. The DNN with the highest AUC DNN_{rfsASA} had a sensitivity of 0.97 (95% CI, 0.92 – 1) and specificity of 0.64 (95% CI, 0.64 – 0.65) and the LR with the highest AUC LR_{rfsASA} had a sensitivity of 0.97 (95% CI, 0.92 – 1) and specificity of 0.64 (95% CI, 0.63 – 0.65). However, all DNNs reduced false positives while maintaining the same or similar number of false negatives. DNN with all 87 original features decreased the number of false positives compared to LR 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; and compared to POSPOM and ASA 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 POSPOM, ASA, LR, $LR_{rfsPOSPOM}$, and LR_{rfsASA} are higher at 20, 5, 0,1, 0.1, and 0.1, respectively (Table 4). All the thresholds for DNN, $DNN_{rfsPOSPOM}$, and DNN_{rfsASA} also increased to 0.3, 0.4, and 0.3, respectively. The highest F1 scores were comparable for ASA, LR_{rfsASA}, and DNN_{rfsASA} at 0.24 (95% CI, 0.14 – 0.35), 0.26 (95% CI, 0.18 – 0.33) and 0.22 (95% CI, 0.12 – 0.30). However, DNN_{rfsASA} had a lower number of false positives at 35 patients compared to LR_{rfsASA} 115 patients.

Calibration

For comparison of calibration, Brier scores and calibration plots were assessed for LR, DNN_{rfsASA}, and calibrated DNN_{rfsASA}. DNN_{rfsASA} had the worst Brier score of 0.0352, and LR had the best score of 0.0065. 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, we assessed the decrease in AUC for the removal of groups of features from the best DNN (DNN_{rfsasa}) (Figure 4). For the analysis, 13 groups were used (Age, Anesthesia, ASA, Input, BP, Output, Vasopressor, Vasodilator, Labs, HR, Invasive Line, Inotrope, and PulseOx). Labs, ASA, anesthesia type, blood pressure, and vasopressor administration were the top features in this analysis. To assess feature importance, we assessed the weights for the logistic regression model (LR_{rfsASA}). ASA had the highest weight. In addition, similar to the DNN, vasopressin, hemoglobin, presence of arterial or pulmonary arterial line, and sevo administration are found in the top 10 weights. (Figure 5)

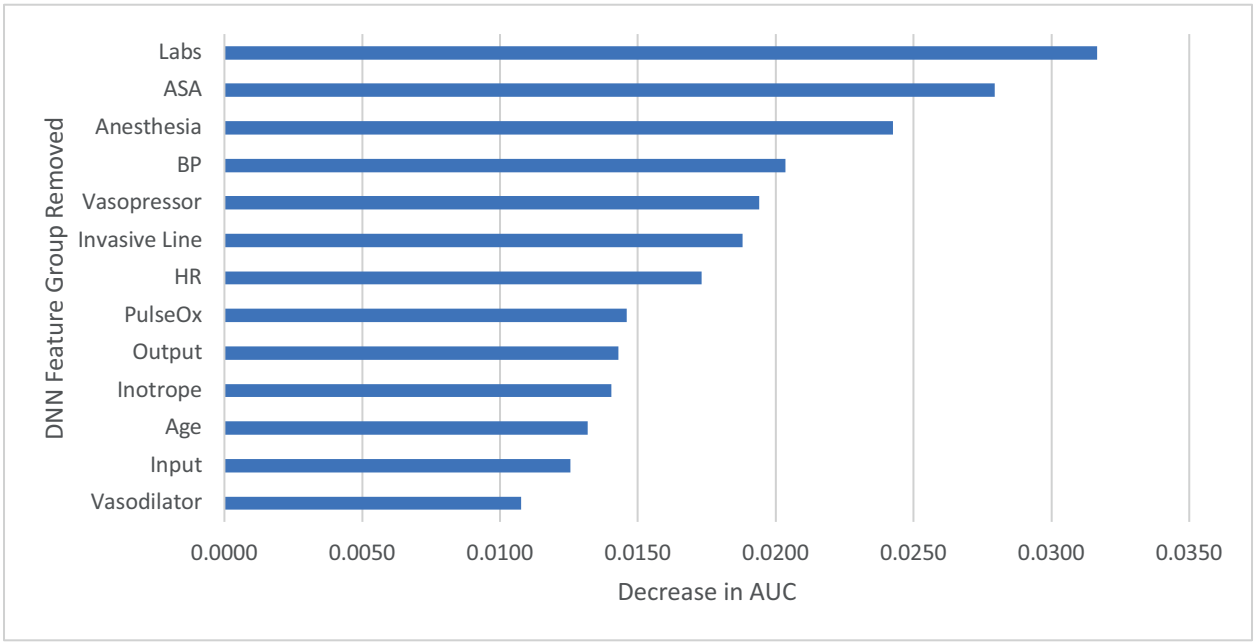


Figure 4. *Feature ablation results for DNN models.*

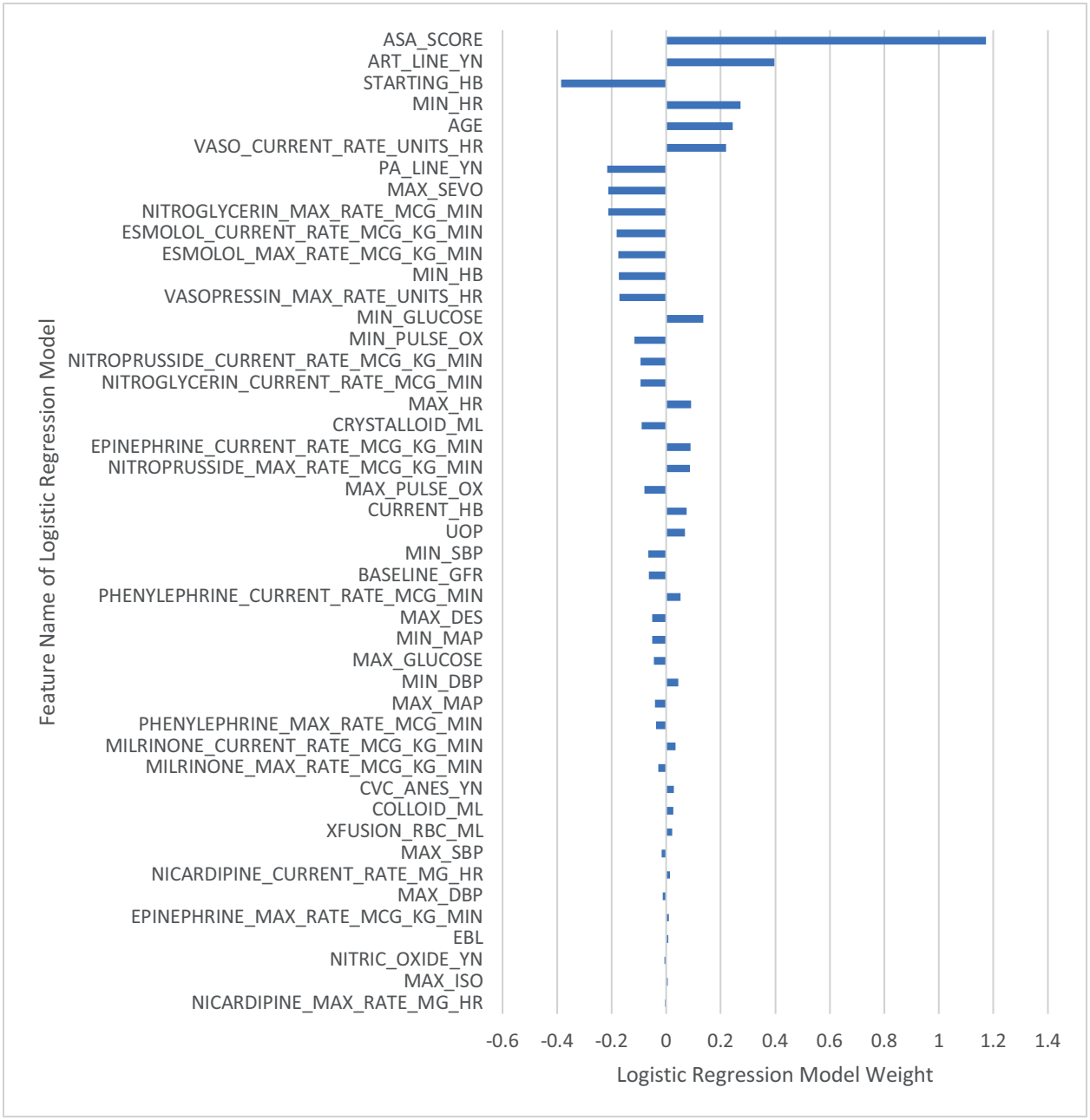


***Figure 5**. Logistic regression models coefficients.*

We have also developed a website 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 <risknet.ics.uci.edu>.

PREDICTING POSTOPERATIVE OUTCOMES: ACUTE KIDNEY INJURY, REINTUBATION, AND MORTALITY

Building off our previous work described above, we were interested in predicting which patients were at risk of poor postoperative outcomes: acute kidney injury (AKI) and reintubation, as well as mortality.

Data Description

Data used in this study was equivalent to the previously described.

Model Endpoint Definition

The occurrence of an in-hospital mortality was extracted as a binary event [0, 1] and described previously. Acute kidney injury (AKI) was determined based upon the change from the patient's baseline serum creatinine (Cr_S) as described in the Acute Kidney Injury Network (AKIN) criteria. Patients were defined as having AKI if they met criteria for any of the AKIN stages based upon changes in their Cr (e.g. had a Cr_S more than 1.5 times their baseline). Patients who lacked either a preoperative or postoperative Cr were excluded only from the AKI and any event models. Postoperative reintubation was determined by documentation of an endotracheal tube or charting of ventilator settings by a respiratory therapist following surgery.

Model Input Features and Data Preprocessing

All data preprocessing and input features were replicated in this study per our previous work. New to this study was the addition of 6 new features: minutes of case time spent with mean arterial pressure (MAP) < 40, 45, 50, 55, 60, and 65 mmHg. These new MAP features were added as potentially relevant features per studies showing the importance of low blood

pressure to the risk of AKI and myocardial infarction.[22,49] For this model, given the addition of 6 new features, we also chose to remove features with a Pearson's correlation > 0.9 with other features and were thus left with a reduced feature set (RFS) of 44 features total. Thus, while the overall architecture of this model is similar to aforementioned model to predict mortality, the various models here have somewhat different input features.

Model Development

We utilized five-fold cross validation with the training set (80%) to select for the best performing deep neural network (DNN) models' hyperparameters and architecture. The hyperparameters assessed were number of hidden layers (1 to 5), number of neurons (10 to 100), learning rate (0.01, 0.1), and momentum (0.5, 0.9). To avoid overfitting, we also utilized L2 regularization (0.001, 0.0001) and dropout probability (0, 0.5, 0.9)[46,50]. These hyperparameters and architecture were then used to train a model on the entire training set (80%) prior to testing final model performance on the separate test set (20%). For patients without a preoperative baseline Cr and/or a postoperative Cr, we could not determine postoperative AKI. Those patients were excluded from training for the individual AKI models and the combined models. In total that amounted to exclusion of 38,305 patients or 63.8% of the total sample.

Individual Models to Predict Each Postoperative Outcome Separately

Similar to our previous work, three separate DNN models were created with each predicting one postoperative outcome of interest: in-hospital mortality, acute kidney injury, reintubation. A logistic output was chosen so that the output of each outcomes model could be

interpreted as probability of each postoperative outcome of interest [0-1]. We also assessed DNN architectures of 3 to 5 hidden layers with [90, 100, 300, 400] neurons per layer, and rectified linear unit (ReLU) and hyperbolic tangent (tanh) activation functions. The loss function was cross entropy. To deal with the highly unbalanced data sets, we also utilized data augmentation during training per our previous work with prediction of in-hospital mortality. Observations positive for reintubation or in-hospital mortality were augmented 100 fold. Observations positive for AKI were augmented 3 fold. Augmentation was done by adding Gaussian noise taken from a Gaussian distribution with a SD of 0.0001.

Combined Model to Predict All Postoperative Outcomes

To assess if a model could leverage the relationship between the three outcomes (i.e. multitask learning), we also created combined models that output probabilities of all three outcomes at once. The same hyperparameters as the individual models were assessed, with the exception of the use of a batch size of 100.

Stacked "Any" Postoperative Outcome Model

We were also interested in predicting the probability of the occurrence of any of the three postoperative outcomes. For the combined DNN model, we took the average of the predicted probability outputs for each outcome (Figure 6). In other words, each predicted probability was given equal weight. The averaged value was considered as the probability of any of the 3 outcomes occurring. For the individual outcome models (DNN and LR) we took the predicted probability of each respective outcome model per equivalent feature set inputs and averaged the three values (Figure 6). For example, the outputs of each of the models for AKI,

reintubation, and mortality with a reduced feature set were averaged to represent the probability of any outcome occurring.

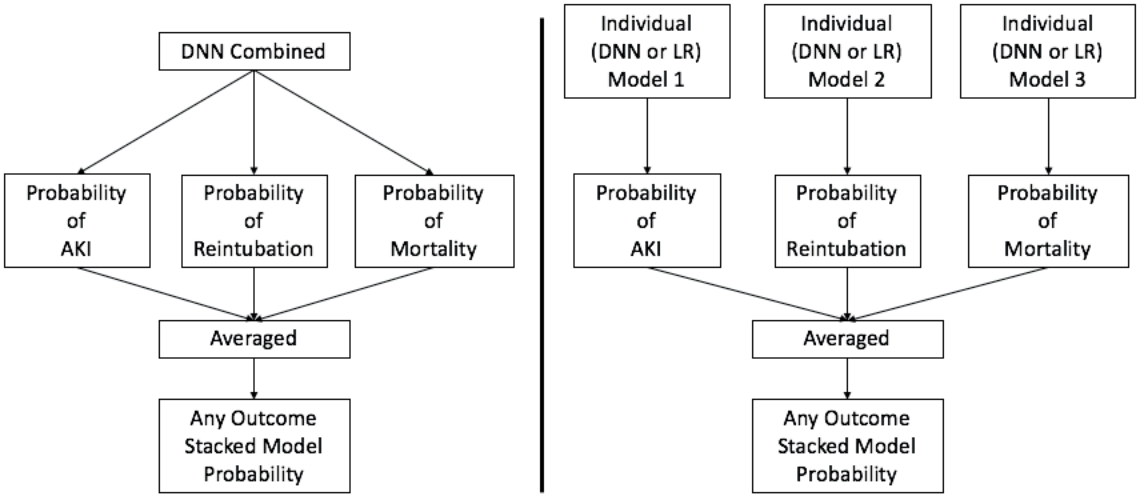


Figure 6. *Summary figure describing the stacked "any" postoperative outcome models for the combined deep neural networks (DNN Combined) trained to output probabilities of all 3 outcomes vs the deep neural networks (DNN Individual) and logistic regression (LR) models.*

Feature Reduction and Clinically Significant Feature Addition Experiments

After choosing the best performing DNN architectures for the reduced features set (RFS), we also assessed the performance of models with two other input feature sets: 1) original 46 features set (OFS) and 2) OFS plus the addition of 6 new MAP features (OFS + MAP). This was done to assess if the reduction of features improved performance compared to a model with more features, and also to assess if the addition of the clinically significant MAP features not used in previous improved performance overall.

Model Performance Methods

All model performances were assessed on 20% of the data held out from training as a test set. Those patients without an AKI label were excluded from evaluation of test set results for AKI, but not for in-hospital mortality, reintubation, or any outcome results. This is due to the input features of each model independence from the determination of AKI, and so all test patients can have an AKI model predicted probability even if AKI class is unknown. Model performance and comparison was performed similar to our previous work, with the addition of average precision (AP) and the McNemar's test.

McNemar's Test to Compare Model Accuracy

To compare the predictions of the DNN and LR models to each other, we utilized McNemar's test.[51] McNemar's test compares the number of correctly predicted samples vs wrongly predicted samples and where they do and do not predict the same label. If the p value of McNemar's test is significant, we can reject the null hypothesis that the 2 models have the same classification performance. McNemar's test was performed using the freely available package MLxtend.[51]

All neural network models were developed using Keras.[47] All performance metrics, except for McNemar's Test, and logistic regression models were developed using scikit-learn.[48]

Results

Patient characteristics have been previously described in Table 9. The final model hyperparameters are described in Table 10 below.

Table 10. *Final postoperative outcomes models hyperparameters.*

	# Layers	# Neurons	L2 Lambda	Learning Rate	Momentum	Dropout Probability
DNN Individual for AKI	4	100	0.0001	0.01	0.9	0.5
DNN Individual for Reintubation	3	100	0.0001	0.01	0.5	0.5
DNN Individual for Mortality	4	90	0.0001	0.01	0.5	0.5
DNN Combined, best AUC for AKI	4	100	0.0001	0.1	0.5	0.5
DNN Combined, best AUC for Reintubation	5	90	0.0001	0.1	0.9	0.5
DNN Combined, best AUC for Mortality	5	90	0.0001	0.1	0.9	0.5

Individual Model Performance

As a baseline, models were created to predict each outcome separately (i.e. AKI, mortality, reintubation or any outcome) using a DNN (DNN OFS). The models all performed well with AUCs of 0.780 (95% CI 0.763-0.796) for AKI, 0.879 (95% CI 0.851-0.905) for reintubation, 0.895 (95% CI .854-0.930) for mortality and 0.866 (95% CI 0.855-0.878) for any outcome. These results as well those for the other models can be found in Table 11.

Table 11. *AUC (95% confidence intervals) for all DNN and LR models as well as risk scores for all outcomes.*

Score	AKI*	Reintubation	Mortality	Any Outcome
ASA	0.652 (0.636 - 0.669)	0.787 (0.757 - 0.818)	0.839 (0.804 - 0.875)	0.76 (0.748 - 0.773)
RQI**	0.652 (0.623 - 0.683)	0.878 (0.842 - 0.909)	0.907 (0.86 - 0.942)	0.8 (0.778 - 0.821)
RSI***	0.594 (0.571 - 0.615)	0.829 (0.783 - 0.873)	0.97 (0.944 - 0.99)	0.597 (0.576 - 0.621)
Model Type	**AKI***	**Reintubation**	**Mortality**	**Any Outcome (Stacked Model)**
LR OFS	0.767 (0.748 - 0.785)	0.856 (0.82 - 0.888)	0.9 (0.865 - 0.93)	0.843 (0.829 - 0.857)
LR OFS + MAP	0.767 (0.749 - 0.785)	0.855 (0.818 - 0.887)	0.898 (0.863 - 0.93)	0.843 (0.829 - 0.857)
LR RFS	0.767 (0.748 - 0.785)	0.862 (0.827 - 0.894)	0.899 (0.864 - 0.93)	0.843 (0.829 - 0.858)
DNN Individual OFS	0.78 (0.763 - 0.796)	0.879 (0.851 - 0.905)	0.895 (0.854 - 0.93)	0.866 (0.855 - 0.878)
DNN Individual OFS + MAP	0.792 (0.775 - 0.808)	0.876 (0.848 - 0.902)	0.903 (0.871 - 0.933)	0.874 (0.864 - 0.886)
DNN Individual RFS	0.783 (0.766 - 0.799)	0.879 (0.851 - 0.905)	0.9 (0.865 - 0.931)	0.866 (0.854 - 0.878)
DNN Combined OFS	0.785 (0.767 - 0.801)	0.858 (0.829 - 0.886)	0.907 (0.872 - 0.938)	0.865 (0.854 - 0.877)
DNN Combined OFS + MAP	0.783 (0.765 - 0.8)	0.84 (0.808 - 0.872)	0.906 (0.87 - 0.937)	0.86 (0.848 - 0.872)
DNN Combined RFS	0.789 (0.772 - 0.806)	0.842 (0.811 - 0.871)	0.906 (0.87 - 0.937)	0.852 (0.84 - 0.864)

Each model was also evaluated for each feature set combination of original feature set (OFS), OFS + the minimum MAP features (OFS + MAP), and reduced feature set (RFS). Note that for the LR and individual models, there is one model per outcome and the predicted outcome probabilities from each model is stacked to predict any outcome. For the combined models, there is one model for all 3 outcomes and those probabilities are stacked to predict any outcome.
**It should be noted that AKI labels were only available for 4307 of the test patients, and so all AUCs reflect results for only those patients with AKI labels.*
*** RQI was calculated on 5,591 test patients (63 Reintubation; 38 Mortality, 491 Any Label); and on 2,319 test patients with AKI labels (445 positive)*
**** RSI was calculated on 11,939 test patients (159 Reintubation; 86 Mortality, 1066 Any Label); and on 4,294 test patients with AKI labels (967 positive)*

Combined Model Performance

In an effort to improve model performance we attempted to train a combined model that would output the risk of each individual outcome. The thought was that in using a model that had information on all of the outcomes the model could "learn" from one outcome in order to predict the others. In fact, these models did not perform better than the original

model: AUC 0.785 (95% CI 0.767-0.801) for AKI, 0.858 (95% CI 0.829-0.886) for reintubation, 0.907 (95% CI 0.872-0.938) for mortality and 0.865 (95% CI 0.854-0.877) for any outcome.

McNemar's Test

In order to assess the ability of the DNN as compared to LR, we used the McNemar Test to look at overall model accuracy. All results were based on the threshold that optimized the F1 score for that model. These results are shown in Table 12. In general, we see that the DNN models and the LR models do perform significantly differently.

Table 12. *McNemar's Test Results*

		AKI*		Reintubation		Mortality		Any Outcome	
DNN Individual	**DNN Combined**	**p**	**p < 0.05**	**p**	**p < 0.05**	**p**	**p < 0.05**	**p**	**p < 0.05**
DNN Individual OFS	DNN Combined OFS	7.78E-03	TRUE	1.00E+00	FALSE	6.16E-01	FALSE	5.58E-01	FALSE
DNN Individual OFS	DNN Combined OFS + MAP Features	2.50E-01	FALSE	6.54E-40	TRUE	1.67E-38	TRUE	7.99E-13	TRUE
DNN Individual OFS	DNN Combined RFS	1.34E-01	FALSE	2.74E-51	TRUE	2.46E-47	TRUE	9.38E-28	TRUE
DNN Individual RFS	DNN Combined OFS	1.42E-07	TRUE	2.76E-05	TRUE	1.05E-07	TRUE	1.50E-02	TRUE
DNN Individual RFS	DNN Combined OFS + MAP Features	1.42E-01	FALSE	1.93E-18	TRUE	2.36E-15	TRUE	4.71E-05	TRUE
DNN Individual RFS	DNN Combined RFS	2.54E-01	FALSE	3.36E-29	TRUE	1.21E-23	TRUE	2.92E-16	TRUE
DNN Individual OFS + MAP Features	DNN Combined OFS	1.80E-10	TRUE	1.97E-27	TRUE	4.81E-31	TRUE	4.93E-07	TRUE
DNN Individual OFS + MAP Features	DNN Combined OFS + MAP Features	2.51E-03	TRUE	1.28E-02	TRUE	4.41E-02	TRUE	1.06E-01	FALSE
DNN Individual OFS + MAP Features	DNN Combined RFS	1.04E-02	TRUE	4.93E-07	TRUE	2.40E-05	TRUE	8.26E-09	TRUE

PREDICTING POST-LIVER TRANSPLANT MORTALITY

Liver transplantation is the definitive treatment for irreversible liver failure, with thousands of lives saved each year in the Unites States through deceased donor organ donation. Unfortunately, with the demand for donor organs far exceeding the supply, thousands of patients die waiting for this life saving procedure.[52]As such, the development of predictive models of post-transplant mortality is crucial to avoid transplanting an individual

with an unacceptably low probability of post-transplant survival. While the prediction of pre-operative mortality among those waiting for an organ has been quite successful with the adoption of the Model for End-Stage Liver Disease (MELD) score to prioritize organ allocation, the accurate prediction of post-transplant mortality has been difficult and less successful.[53–56]

Two of the most commonly cited risk models are the Balance of Risk (BAR) score and the Survival outcomes following liver transplantation (SOFT) score, both of which predict 90-day post-liver transplant mortality using United Network of Organ Sharing (UNOS) registry data.[57,58] The SOFT score incorporated a combination of 18 recipient and donor variables and achieved a c-statistic of 0.7, and the BAR score achieved a C-statistic of 0.7 using a combination of just 6 recipient and donor variables. Despite the popularity of these models in academic circles, their clinical use has been limited due to their modest discriminative performance.

In this study, we attempted to develop a DNN model using pre-operative variables from the UNOS registry to predict 90-day post-liver transplant mortality.

Data Description

Data Extraction

All data for this study were extracted from the standard transplant analysis and research (STAR) dataset which contains patient-level data for all transplants in the Unites States reported to the Organ Procurement and Transplantation Network (OPTN) since October 1, 1989. The database has been used in numerous important studies of transplantation and contains data on pre-transplant variables pertaining to the recipient, donor variables reported from the organ procurement organization, as well as post-transplantation outcome data. The OPTN mortality

data are linked by UNOS to the Social Security Death Master file to improve ascertainment of recipient. [59]

The study sample included adult deceased donor liver transplants performed from 2005 to 2015. Transplants performed from 2016 onwards were not included in this analysis to ensure adequate time for ascertainment of outcome data, and transplants performed prior to 2005 were excluded because 1) transplants before 2002 were performed prior to implementation of the MELD score allocation system, and 2) data on several predictor variables were either not reported or were inconsistently recorded prior to that time. Exclusion criteria included age less than 18 years, living donor transplantation (n=2,347), multiple-organ transplantation (n=5,267), as well as those lost to follow-up within 90 days post-transplantation (n=70) as these cases were excluded in the development of the SOFT score and BAR score. For patients who underwent more than one liver transplantation (n=3,503), we included each of the transplantations in the analysis as did other comparable prediction models. The study sample included split liver as well as Donation after Cardiac Death (DCD) donors. In sum, we analyzed 57,544 recipients.

Model Endpoint Definition

The occurrence of death within 90 days from transplantation was extracted as a binary event [0, 1]. An event occurred if the value of the variable "pstatus" from the STAR dataset was equal to "1" and the variable "ptime" was less than or equal to 90. The variable "pstatus" indicates whether the recipient had died post-transplant, and the variable "ptime" indicates the time from transplantation to either death or censoring. These variables are based on the combination of mortality data from OPTN database as well as verified external sources of death

(described above), and not based on the variable "PX_STAT" which only accounts for death as documented by the OPTN alone.

Model Input Features

The original STAR dataset contained 395 variables, many of which were not considered for inclusion in the model. Variables that were excluded from model development included those pertaining to post-transplant data, living donor transplants, multi-organ transplants, and identifier code variables. Variables with zero or near zero variances, high levels of missing data (>98%) or those that were highly correlated to other variables (r>0.99) were removed. A few variables with >50% missing data combined with low clinical significance based on domain experts were not analyzed. This resulted in 202 features including 132 recipient variables and 70 donor-related variables (Appendix A). To further reduce the feature set, variables with greater than 50 percent missing data or those containing greater than 95% zero values were removed and the remaining variables comprised a reduced feature set (RFS).

While most of the categorical features had a simple binary encoding (Appendix A), categorical features identified by domain expert that required more complex encoding were encoded based on clinician judgment. For example, the variable "DIAG", which indicates a recipient's primary liver disease diagnosis at transplantation, contains 70 possible unique diagnosis codes. Rather than creating 70 new, binary categorical features, groups of diagnosis codes were used to collapse the 70 unique codes into 11 new categorical features.

BAR Score and SOFT Score

The BAR score and SOFT score are two models used to predict 90-day post-liver transplant survival using UNOS data. To compare the discriminative ability of the DNN to that of

these models, the BAR score and SOFT score were calculated for recipients in this dataset. Data on cold ischemia time was missing for 2.8% of recipients, and therefore, the BAR score could not be calculated for these subjects. The amount of missing data for other variables was <0.1%, and these cases were removed from the calculation of the BAR score's area under the receiver operating characteristics curve (AUC). Missing data for the SOFT score was handled by assigning the missing value to the reference group category as indicated by the scoring methodology. One of the 18 variables that comprises the original SOFT score is the presence of a portal bleed within 48 hours of transplantation. This variable was not available in the STAR dataset and therefore was not included in the calculated SOFT score. In our analysis, we calculated the SOFT score using the remaining 17 components.

Data Preprocessing

Prior to model development, missing values were imputed with the mean value for continuous variables and with 0 for categorical variables. The data were then randomly divided into training (80%) and test (20%) data sets. The training data was rescaled to have a mean of 0 and standard deviation of 1 per feature. The test data was rescaled to the training mean and standard deviation.

"Soft" Binning Features

Besides following the standard approach of normalizing individual input features we also experimented with a novel idea that we will refer to as "soft binning". Similar to standard/"hard" binning, the data representation of any feature is replaced by a fixed number of bins, containing numbers between 0 and 1. Ordinary binning discretizes a feature by representing it as a single "1" in one bin, and zeroes in all other bins, potentially resulting in loss

of information and making the classification task harder. "Soft" binning is the most straight-forward generalization of binning without loss of information, where two bins are assigned values in the range of 0 to 1, which sum to one. These values encode the fraction to which the feature's value falls into the given bins. For example, if in standard binning a value would fall exactly on the boundary between two bins, then it would instead be represented as two neighboring entries of "0.5" in the neighboring bins in "soft" binning. Our motivation for creating "soft" binning was that binning alleviates the burden for the neural network to learn individual features thresholds (i.e. "high", "average", or "low"), and thus improves classification accuracy.

Model Development

The primary aim of the study was to classify recipients with 90-day post-liver transplant mortality using deep neural networks (DNNs), also referred to as deep learning. The type of DNN used in this study was a feedforward network with fully connected layers and a logistic output. A logistic output was chosen so that the output of the model could be interpreted as probability of mortality [0-1]. We utilized stochastic gradient descent with momentum [0.2, 0.5, 0.9] and initial learning rates [0.01, 0.001, 0.1], and a batch size of 500. We also assessed DNN architectures of 1 to 5 hidden layers with [10, 50, 100, 110, 115, 120, 130, 140, 150] neurons per layer, and rectified linear unit (ReLU) activation functions. The loss function was cross entropy. To minimize overfitting, we utilized three methods: 1) early stopping with a patience of 10 epochs, 2) L2 weight decay, and 3) dropout.[46,50] We assessed L2 weight penalties of [0.01, 0.001, 0.0001] and dropout was applied to all layers with a probability of [0, 0.2, 0.5, 0.9]. 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%) prior to testing final model performance on the separate test set (20%).

Model Performance Methods

All model performances were assessed on 20% of the data held out from training as a test set. Model performance was assessed using area under the receiver operating curve (AUC) and were compared to the BAR score and the SOFT score.

Results

Best neural network hyperparameters for each DNN and feature set are described in Table 13.

Table 13. *Final model hyperparameters for liver transplant models.*

	# Hidden Layers	# Neurons per layer	L2 Lambda	Dropout probability	Learning Rate	Momentum
DNN w/ original 202 features (OFS)	5	100	0.001	0.5	0.01	0.5
DNN w/ OFS + softbin	5	110	0.001	0	0.01	0.5
DNN w/ reduced 140 features (RFS)	5	100	0.001	0.5	0.01	0.5
DNN w/ RFS + softbin	5	110	0.001	0	0.01	0.5

Model Performance

All performance metrics reported below refer to the test dataset.

The best DNN model (DNN with OFS + softbin) had a higher AUC (0.703 (95%CI: 0.682 - 0.726)) compared to that for the BAR score and SOFT score models (0.655 (95%CI: 0.633 -

0.678); 0.688 (95%CI: 0.667 - 0.711)), respectively on the 11,207 patients with available BAR scores (Table 14). In addition, softbin preprocessing of input features improved performance of both the OFS and RFS models. While the best DNN had a significantly higher AUC than the BAR score, the DNN did not achieve a significantly higher AUC than the SOFT score. The DNN with the reduced feature set and softbin preprocessing (DNN with RFS + softbin) performed comparably (AUC 0.702 (95%CI:0.68 - 0.725)) to the DNN with OFS + softbin.

Table 14. *AUC (95% confidence intervals) for all DNN models as well as BARscore and SOFTscore,*

	AUC (95% CI)	
	n = 11,509	n = 11,207*
BARscore*	0.655 (0.633 - 0.678)	0.655 (0.633 - 0.678)
SOFTscore	0.691 (0.671 - 0.714)	0.688 (0.667 - 0.711)
DNN w/ Original 202 Features Set (OFS)	0.697 (0.678 - 0.72)	0.695 (0.675 - 0.717)
DNN w/ OFS + softbin	0.708 (0.689 - 0.73)	0.703 (0.682 - 0.726)
DNN w/ Reduced 140 Features Set (RFS)	0.699 (0.681 - 0.722)	0.698 (0.679 - 0.72)
DNN w/ RFS + softbin	0.707 (0.688 - 0.729)	0.702 (0.68 - 0.725)

*BARscore was calculated on 11,207 test patients due to missing data.

By choosing a threshold that optimizes the F1 score, the SOFT score achieved the highest F1 score (0.215 (95%CI:0.191 - 0.238)) at a threshold of 20, with sensitivity and specificity of 0.375 (95%CI:0.336 - 0.416) and 0.881 (95%CI:0.875 - 0.888), respectively for the 11,207 patients with available BAR scores. This score was not significantly different from the highest F1 score among the DNN models, which was achieved by DNN with RFS + softbin (0.21 (95%CI:0.187 - 0.236)) at a threshold of 0.106, with sensitivity and specificity of 0.331 (95%CI:0.296 - 0.369) and 0.898 (95%CI:0.892 - 0.904), respectively. At this threshold, the SOFT score had slightly more true positives compared to the DNN model (223 vs 199) as a result of the higher sensitivity, but with more false positives (1194 vs 1099) as a result of the lower

specificity. The best DNN model based on AUC, namely DNN with OFS + softbin, had a comparable F1 score 0.209 (95%CI:0.184 - 0.234) at a threshold of 0.113.

Overall, the results demonstrated that a DNN can be utilized to predict 90-day post-liver transplant mortality using UNOS registry data. While the AUC for the best performing DNN (DNN with OFS + softbin) was the highest among the tested models, significantly outperforming the BAR score, it did not achieve significantly higher performance compared to the SOFT score.

PREDICTING INTRAOPERATIVE HYPOTENSION USING THE ARTERIAL BLOOD PRESSURE WAVEFORM

In collaboration with Edwards Lifesciences, we developed a continuous predictor of risk of hypotension, also known as Hypotension Probability Index (HPI™). It has been shown that even just one minute of intraoperative hypotension can lead to increased risk of poor postoperative outcomes.[22,49] Thus, there is a need in critical care monitoring to help clinicians identify the onset of a hypotensive event.

Data Description

Data used in the development and testing of HPI came from Edwards Lifesciences internal database from clinical studies as well as from the MIMIC II Waveform Database for a total of 1,280 patients.[5] 954 patients came from the Edwards internal database. These patients included both surgical and ICU patients, and represented a wide range of surgical procedures such as cardiac bypass and liver transplants and various acute conditions such as sepsis. 326 patients came from the MIMIC II database, and these were all ICU patients. 302 patients were

used for training, 628 patients were used for validation and 350 patients were set aside as a test set.

Model Endpoint

The first task was to define a hypotensive event. A hypotensive event was defined as MAP < 65 mmHg for at least 1 minute. Non hypotension was defined as MAP > 75 mmHg and at least 20 minutes away from the start or end of an event. These definitions were settled on after discussion with domain experts as well as review of the literature looking at all currently used definitions for hypotension in the clinical and research setting.[60] It should be noted that the model features and results are proprietary to Edwards Lifesciences. Due to the proprietary nature of this model, description of the methods and all results being shown have been either published or are currently accepted conference abstracts.[61]

Model Input Features: Description of Arterial Blood Pressure Waveform Features

After defining hypotensive events, feature selection was performed to select the best features from the radial arterial pressure waveform as calculated by the Edwards Lifesciences FloTrac™. The FloTrac™ is a pressure transducer that transforms the arterial pressure waveform for display on the anesthesia monitor, but more importantly the FloTrac™ calculates physiologically relevant features of the waveform that clinicians use in the hemodynamic management of the patient such as stroke volume, stroke volume variation, blood pressure, systemic vascular resistance etc. There are also other more complex features that are calculated but not currently displayed for the clinician for research and development purposes.

These include spectral features of the waveform, durations between different peaks and troughs in the waveform, areas, slopes, entropy, variability, etc. In addition to these features, the FloTrac™also calculates combinatorial features, nonlinear and linear combinations of all features. All features are calculated on a beat to beat basis and then averaged over 20 seconds. A beat is defined by the systole onset to end of diastole, reflecting the contraction and relaxation of the heart (Figure 7). The FloTrac™ currently calculates ~2,600,000 features.

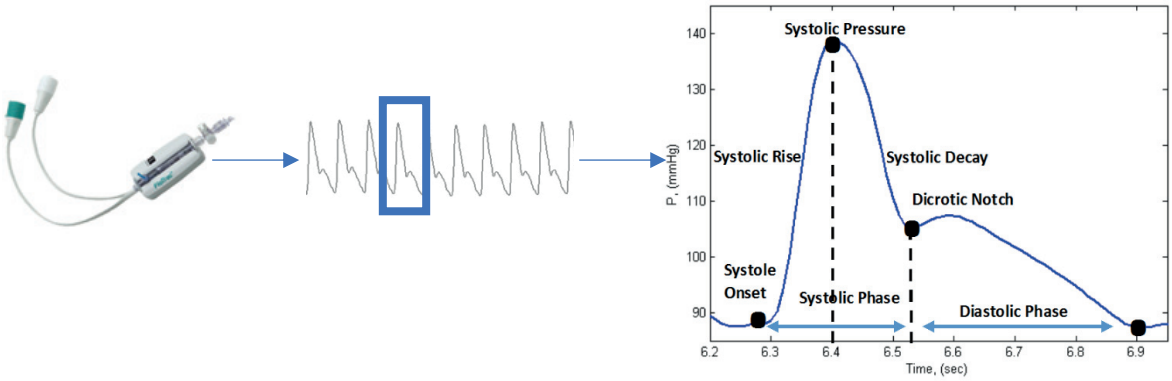


Figure 7*. A typical arterial pressure waveform and a zoomed in view of one cardiac beat.*

Model Development

The 302 patients used for training contained 25,461 positive (for hypotension) observations and 56,143 negative observations. Observations were defined as instantaneous 20 second values calculated by FloTrac™at a positive or negative time point. The 628 patients used for validation contained 25,350 positive observations and 70,864 negative observations. The 350 patients used for test contained 14,969 positive observations and 49,011 negative observations.

Feature Selection

To reduce the number of features from >2 million to a more reasonable number, first the AUC for every feature was calculated and the features with an AUC > 0.85 were selected. This resulted in 62 features. Sequential forward feature selection with logistic regression and ten-fold cross validation was then performed on those features, resulting in 23 features. Sequential forward feature selection is a method that adds features one by one to a logistic regression model, the feature that improves the misclassification rate is kept until some stopping point is set where the addition of another feature only improves the misclassification rate by a negligible amount. These 23 features were used as the final variables for a logistic regression model. The Hypotension Probability Index (HPI™) is the probability outputted by the logistic regression model, displayed as a % value.

Model Performance Methods

AUC of the ROC curve was used to evaluate the performance of the model on the validation set to choose the best model. The best model's performance was then evaluated on the test set using ROC analysis performance as well. For test evaluation, we were specifically interested in if the model could not only detect the start of an event, but also how it performed on the time points leading up to the start of an event. The model was evaluated on the time points at the start of an event as well as time points leading up to an event. It was assumed that these time points should indicate positive for hypotension. In other words, 0 minutes to event were all the time points at the start of a hypotensive event. X minutes to event were all the time points between X minutes prior to the event and start of event. The reason for this is that HPI is intended to be an indicator for not only the detection of an event but also the onset of

the event. In other words, we expected that HPI should increase as blood pressure starts heading towards a hypotensive event. It should be noted that all events are 100% detected by definition. As in, HPI is 100% regardless of the logistic regression model's output if MAP < 65 mmHg.

Results

HPI performed with AUC > 0.9 up to 15 minutes prior the start of a hypotensive event (Table 15).

Table 15. *AUC results at x minutes prior to start of a hypotensive event.*

Time to Start of Hypotensive Event (min)	AUC
0	1
1	1
2	0.998
3	0.994
4	0.99
5	0.987
10	0.973
15	0.964

In addition, we see qualitatively that HPI works as intended, trending towards 100% as MAP approaches a hypotensive event. In Figure 8, we can see that HPI is greater than 50% about 15 minutes prior to the start of an event. After the first event ended, another one started about 10 minutes after and HPI remained high.

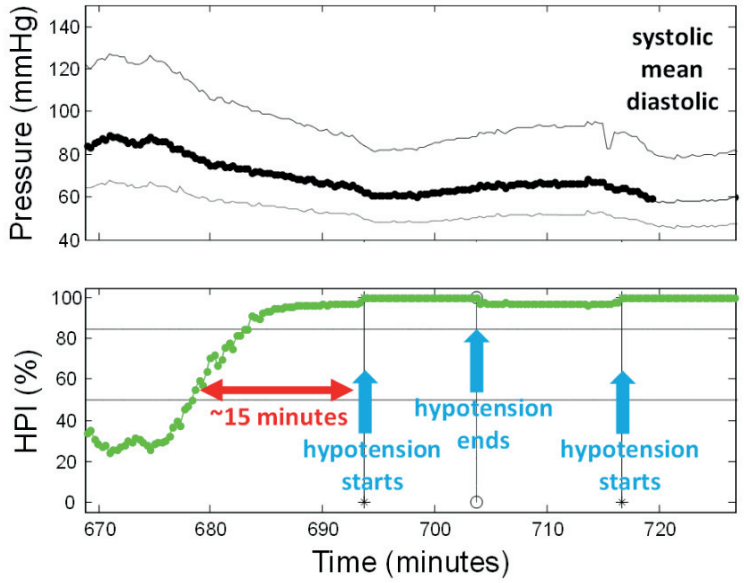


Figure 8. *An example of blood pressure decreasing towards a hypotensive event.*

Overall, HPI shows great performance for the prediction of hypotensive events and we expect that it can be used to help reduce time spent in hypotension and thus improve postoperative outcomes.

PREDICTING POST-INDUCTION HYPOTENSION USING THE ARTERIAL BLOOD PRESSURE WAVEFORM AND EMR

Intraoperative hypotension has been shown to be associated with postoperative morbidity and mortality.[22,49] One time period of specific interest is the time following induction of anesthesia and prior to the start of surgery, aka postinduction. While later on in a patient's case hemodynamic changes are affected by surgical stimulation in combination with general anesthesia, postinduction hypotension can uniquely be attributed to the induction event.

While there has been research into the various predictors of postinduction hypotension, machine learning has not been leveraged until recently by Kendale et al. to tackle this problem.[62–65] Utilizing data from 13,323 surgical procedures, Kendale et al. compared the use of logistic regression, support vector machines, naïve Bayes, k nearest neighbor, linear discriminant analysis, random forest, neural networks, and gradient boosting machines to predict post induction hypotension. The overall best performing model type was gradient boosting machine with an AUC of 0.74. The input features for these models included demographics, ASA, medical comorbidities, preoperative medications, intraoperative medications, intraoperative medications, surgical start, mean peak inspiratory pressure, and first mean arterial pressure. Kendale et al.'s study demonstrated that the prediction of postinduction hypotension was feasible using machine learning and readily available pre-induction information.

In this work, we aimed to expand on Kendale et al.'s work to develop a deep neural network model for predicting postinduction hypotension using data from both electronic medical record (EMR) and arterial blood pressure (ABP) waveform. We hypothesized that the use of more complex features in combination with a well-trained complex model would improve performance of the classification.

Data Description

All data for this study was collected retrospectively from the University of California, Irvine Medical Center (Orange, California), and was described previously in the UCIMC data collection section.

EMR Features

We chose to replicate the features used by Kendale et al. in the prediction of postinduction hypotension[65]. The original Kendale et al. features include patient comorbidities and preoperative medications which we did not have available in our EMR data extraction. Thus, we included only the following subset of Kendale et al.'s original EMR features: age, sex, body mass index (BMI), ASA score, intraoperative medications, mean peak inspiratory pressure, first mean arterial pressure (MAP), and hour of surgery start time (Table 16). Sex was binary encoded as 0 or 1 for male or female, respectively. Intraoperative medications included the first administered amounts for midazolam, propofol, etomidate, fentanyl, rocuronium, and succinylcholine; and maximum sevoflurane concentration and desflurane concentration from case start to 10 minutes from induction start. Mean peak inspiratory pressure was also extracted from the same time window. If no medication or gas was administered, the value was set as 0. All medications were also cleaned for uniform units (i.e. all fentanyl was converted to mg). Incorrect "annotations" for medications and peak inspiratory pressure were identified by values outside the clinically acceptable range and set to either the minimum or maximum possible (Table 16). Prior to extracting the first MAP value, MAP values outside of a physiologic range similar to Kendale et al. (MAP less than 20 mmHg, MAP greater than 200 mmHg, or pulse pressure less than 20 mmHg) were excluded, and then the first MAP was extracted. ASA score was treated as categorical and one hot encoded (each unique ASA score 2 to 6 was binary encoded as its own categorical feature), with the emergency annotation excluded. In addition to the Kendale et al. features, we also included maximum, minimum, and mean heart rate (HR), MAP, and pulse oximetry (SpO2) values, resulting in 28 EMR features total.

Table 16. *Description of model input features.*

Feature	Feature Description	Source	Minimum; Max Values
Age*	Age at date of surgery (years)	EMR	
Sex*	Female = 1; Male = 0	EMR	
BMI*	Body mass index (kg/m2)	EMR	
ASA 2*	ASA Score 2 = 1, else 0	EMR	
ASA 3*	See above	EMR	
ASA 4*	See above	EMR	
ASA 5*	See above	EMR	
ASA 6*	See above	EMR	
First MAP*	First MAP cuff value of case (mmHg)	EMR	
Surgery Start*	Hour of surgery start time	EMR	
Max DES*	Maximum desoflurane concentration administered	EMR	
Max SEVO*	Maximum sevoflurane concentration administered	EMR	
mean PIP*	Mean peak inspiratory pressure	EMR	
First Etomidate *	Amount of first etomidate administered (mg)	EMR	1; 40
Fist Fentanyl*	Amount of first fentanyl administered (mcg)	EMR	0; 500
First Midazolam*	Amount of first midazolam administered (mg)	EMR	0; 10
First Propofol*	Amount of first propofol administered (mg)	EMR	10; 1000
First Succinylcholine*	Amount of first succinylcholine administered (mg)	EMR	0; 400
First Rocuronium*	Amount of first rocuronium administered (mg)	EMR	0; 3000
Max HR	Maximum HR (EKG)	EMR	
Max MAP	Maximum MAP from cuff	EMR	
Max SpO2	Maximum SpO2	EMR	
Mean HR	Mean HR (EKG)	EMR	
Mean MAP cuff	Mean MAP from cuff	EMR	
Mean SpO2	Mean SpO2	EMR	
Min HR	Minimum HR (EKG)	EMR	
Min MAP cuff	Minimum MAP from cuff	EMR	
Min SpO2	Minimum SpO2	EMR	
Mean SBP	Mean systolic BP	Waveform	
Mean DBP	Mean diastolic BP	Waveform	
Mean MAP	Mean MAP	Waveform	
Mean PP	Mean pulse pressure	Waveform	
Mean Period	Mean beat period	Waveform	
Mean Dyneg	Mean of all negative slopes	Waveform	
Mean SYS Area	Mean area under systole	Waveform	
Mean HR	Mean heart rate	Waveform	

*These features were replicated from Kendale et al.

Arterial Blood Pressure (ABP) Waveform Processing and Feature Extraction

All available ABP waveforms from 5 minutes prior to induction up to induction were extracted and resampled to 100 Hz for uniformity. The waveform was then split into 20 second

windows, and processed for beat detection, beat-to-beat features and beat signal quality index (SQI) using algorithms provided by the blood pressure waveform analysis tools in the PhysioNet Cardiovascular Signal Toolbox (Figure 9).[66] 8 beat-to-beat waveform features were calculated: systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse pressure (PP), mean arterial pressure (MAP), mean of all negative slopes (noise; dyneg), beat period, heart rate, and area under systole. SQI is binary, and any beats considered "bad" were excluded from analysis. The mean of all features' "good" values were then taken as input features into the model.

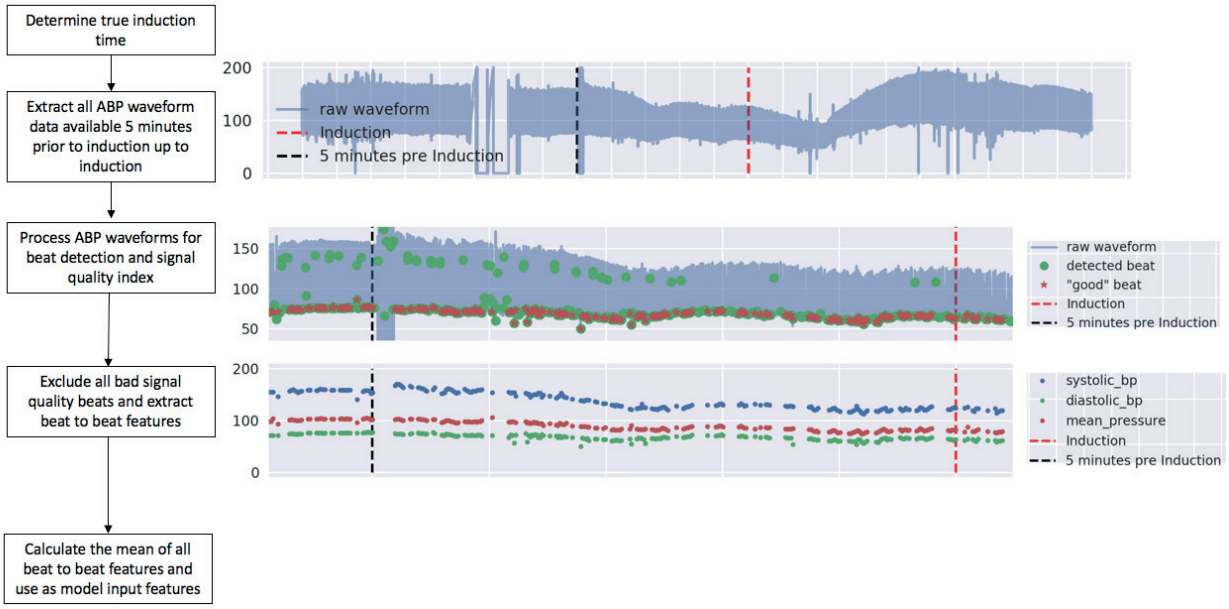


Figure 9. *Description of processing the raw arterial blood pressure (ABP) waveform for model inputs*

Model Endpoint

Induction was defined as the first time between the EMR recorded induction event, propofol administration, and etomidate administration. We chose this definition rather than just the induction event time recorded in the EMR to avoid inaccurate labeling caused by

potential delays in clinician annotations. We extracted "clean" beat-to-beat MAP values from the ABP waveform as described above in waveform processing, and then calculated the median MAP preinduction and the median MAP 0 to 5 minutes and 5 to 10 minutes post induction.[64] We chose to define hypotension from the MAP extracted from the ABP waveform rather than the EMR data to control for signal quality of the MAP values for labeling as well as to ensure time synchronization. In addition, rather than the original Kendale et al. definition of any MAP less than 55 mmHg within 10 minutes of induction, we chose the Reich et al. definition as it includes two time periods for the label prediction and is more robust to single value outliers. Postinduction hypotension was defined as during 0 to 5 or 5 to 10 minutes post induction: 1) postinduction MAP decrease of > 40% from preinduction and postinduction MAP < 70 mmHg or 2) postinduction MAP < 60 mmHg

Model Development

In this work, we were interested in predicting 2 labels: hypotension within 0 to 5 minutes and within 5 to 10 minutes postinduction using deep neural networks (DNN), aka deep learning. We utilized feed forward networks with fully connected layers and a logistic output to output a probability of hypotension (0 to 1). We trained all models with the Adam optimizer[67] with default parameters and initial learning rates (0.01, 0.1, 0.5), and a batch size of (32, 64, 128). The learning rate was reduced by a factor of 10 when validation loss stopped improving.

We also assessed DNN architectures of 1 to 5 hidden layers with 10 to 100 neurons per layer, with hyperbolic tangent (tanh) activation functions. To reduce overfitting, we utilized early stopping with a patience of 10 epochs, L2 weight regularization (0.001, 0.0001) and

dropout [46,50](0, 0.25, 0.5). The loss function was cross entropy. We trained two separate models, one for each label. Due to the limited number of patients available for training, we utilized leave one out (LOO) cross validation. For each iteration of LOO, 1 patient was held out from training for validation and the other 223 were split into 80% for training and 20% for "training-validation" to track loss to avoid overfitting. The model results reported here are for the models with the best LOO validation performance. We also trained models with different input feature sets: 1) ABP waveform features (n=8), 2) EMR features (n=28) and 3) both waveform and EMR features (n=36).

Model Performance Methods

For comparison, we looked at logistic regression (LR) with the same feature sets. For all performance results, we took each leave one out validation result and pooled them together to calculate area under the receiver operating characteristic curve (AUC) and average precision (AP). Sensitivity, specificity, precision, and F1 scores were calculated for thresholds chosen by highest F1 score for each model. 95% confidence intervals were calculated with bootstrapping.

We assessed feature importance in the DNN models with feature ablation analysis. After finalizing the model architectures, we removed each feature, performed leave one out training and validation, and assessed the decrease in AUC for the validation data. To assess which features are the most predictive in the logistic regression model. We calculated the mean weights for each feature following leave one out training.

All deep neural network models were developed using Keras[68]. Logistic regression models and performance metrics were developed with scikit-learn.[48]

Results

There was a total of 19,545 surgical patients from November 2015 to August 2017, of which 1,120 patients had arterial blood pressure (ABP) waveform data prior to induction. After waveform preprocessing for signal quality, there were 224 patients included in the model development. The significant decrease in patient numbers from those with ABP waveforms to those with "good" signal quality is due to the presence of a signal even if there is no arterial line connected to the monitor. As long as an arterial line transducer is connected to the patient monitor, an ABP signal is collected. The transducer is usually set up prior to arterial line placement, and thus noise is collected until the arterial line is placed and zeroed. Preprocessing the waveform for signal quality was essential to exclude noisy data. Of the 224 patients, 22 patients (9.8%) had postinduction hypotension within 0 to 5 minutes and 20 patients (8.9%) had postinduction hypotension within 5 to 10 minutes. Patient demographics are described in Table 17.

The final deep neural network parameters for each DNN model and feature set combination are described in Table 18. All performance metrics reported below refer to the LOO validation data (n=224).

Table 17. *Description of patient demographics*

# Patients	224	
Age	57 +/- 21	
BMI	26.7 +/- 5.7	
Anesthesia Time (hours)	5.2 +/- 2.8	**% of 224**
Female	150	67.0
ASA Score		
1	1	0.4
2	5	2.2
3	75	33.5
4	119	53.1
5	18	8.0
6	6	2.7
Admission Type		
Unknown	1	0.4
Day Prior Admission	4	1.8
23 Hour Observation	1	0.4
Outpatient	5	2.2
AM Admission	26	11.6
Midnight Admission	13	5.8
Inpatient	174	77.7
Anesthesia		
None	1	0.4
General	218	97.3
MAC	4	1.8
Combined Spinal/Epidural	1	0.4
Postinduction Hypotension		
0 to 5 minutes	20	8.9
5 to 10 minutes	22	9.8

Table 18. *Final model hyperparameters for each DNN model and feature combination for predicting hypotension 0 to 5 minutes or 5 to 10 minutes postinduction*

	0 to 5 Minutes Post Induction Hypotension						
	# Features	# Layers	# Neurons	Dropout Probability	L2 Lambda	Batch Size	Learning Rate
Waveform Only	8	2	60	0.25	0.0001	128	0.001
EMR Only	28	2	50	0.25	0.001	128	0.001
Waveform + EMR	36	2	50	0.25	0.001	128	0.001
	5 to 10 Minutes Post Induction Hypotension						
	# Features	# Layers	# Neurons	Dropout Probability	L2 Lambda	Batch Size	Learning Rate
Waveform Only	8	2	70	0.25	0.0001	128	0.001
EMR Only	28	2	100	0.25	0.001	128	0.001
Waveform + EMR	36	2	80	0.25	0.0001	128	0.001

Model Performance

Area under the receiver operating characteristic curve (AUC ROC) and average precision (AP) are summarized in Table 19. All DNN models had higher AUCs than logistic regression (LR) for each feature set, except for the EMR only features model to predict 5 to 10 minutes post induction hypotension (DNN AUC 0.63 (0.497 – 0.76); AP 0.143 (0.084 – 0.266) vs LR AUC 0.667 (0.555 – 0.78); AP 0.151 (0.089 – 0.284)). The best performing model for predicting 0 to 5 minutes post induction hypotension was DNN with waveform only features (AUC 0.88 (0.812-0.934); AP 0.391 (0.241 – 0.631), followed by LR with waveform only features (AUC 0.875 (0.81-0.929); AP 0.372 (0.224 – 0.598)). The best performing model for predicting 5 to 10 minutes postinduction was DNN with waveform only features (AUC 0.703 (0.557-0.823); AP 0.228 (0.115 – 0.433)), followed by LR with EMR only features (AUC 0.667 (0.555-0.78); AP 0.176 (0.089 – 0.345)). When assessing the different feature sets, the use of waveform only features performed best overall, and EMR only features performed the worst, except in predicting 5 to 10 minutes postinduction hypotension with LR.

Table 19. *AUC and AP with 95% CIs for the DNN and LR models for prediction of postinduction hypotension.*

AUC		
0 to 5 Minutes Post Induction		
Feature Set	**DNN Model**	**LR Model**
Waveform Only	0.88 (0.812-0.934)	0.875 (0.81-0.929)
EMR Only	0.51 (0.402-0.623)	0.505 (0.363-0.637)
Waveform + EMR	0.804 (0.703-0.888)	0.792 (0.695-0.873)
5 to 10 Minutes Post Induction		
Feature Set	**DNN Model**	**LR Model**
Waveform Only	0.703 (0.557-0.823)	0.613 (0.452-0.752)
EMR Only	0.63 (0.497-0.76)	0.667 (0.555-0.78)
Waveform + EMR	0.653 (0.512-0.779)	0.603 (0.475-0.725)

AP		
0 to 5 Minutes Post Induction		
Feature Set	**DNN Model**	**LR Model**
Waveform Only	0.391 (0.241-0.631)	0.372 (0.224-0.598)
EMR Only	0.102 (0.064-0.182)	0.106 (0.066-0.193)
Waveform + EMR	0.294 (0.168-0.509)	0.317 (0.164-0.53)
5 to 10 Minutes Post Induction		
Feature Set	**DNN Model**	**LR Model**
Waveform Only	0.228 (0.115-0.433)	0.176 (0.089-0.345)
EMR Only	0.143 (0.084-0.266)	0.151 (0.089-0.284)
Waveform + EMR	0.15 (0.087-0.266)	0.121 (0.069-0.223)

We used the highest F1 score to choose a threshold for all models, as it balances sensitivity and specificity. When predicting 0 to 5 minutes postinduction hypotension, the DNN model with waveform features had the highest F1 score (0.537 (0.324-0.706), followed by the LR model with waveform features (0.5 (0.279-0.667)). The DNN model with waveform features had a higher specificity than the LR model (0.96 (0.931-0.985) vs 0.946 (0.912-0.975)), and equivalent sensitivity (0.5 (0.286-0.714) vs 0.5 (0.278-0.714)). When predicting 5 to 10 minutes postinduction hypotension, the DNN model with waveform features had the highest F1 score (0.364 (0.158-0.533)), followed by the LR model with waveform features (0.312 (0.08-0.513)). The DNN model had higher sensitivity than the LR model (0.4 (0.176-0.625) vs 0.25 (0.059-0.444)), but lower specificity (0.922 (0.884-0.956) vs 0.966 (0.939-0.99)).

Feature Importance

Although the models trained with the waveform and EMR feature sets were not the best performing model, we wanted to compare not only the difference in feature importance across model types (DNN vs LR) but also how each model looked at the EMR vs waveform features. We assessed the results of feature ablation for the DNN trained with EMR and waveform features and the LR model weights for the same feature set.

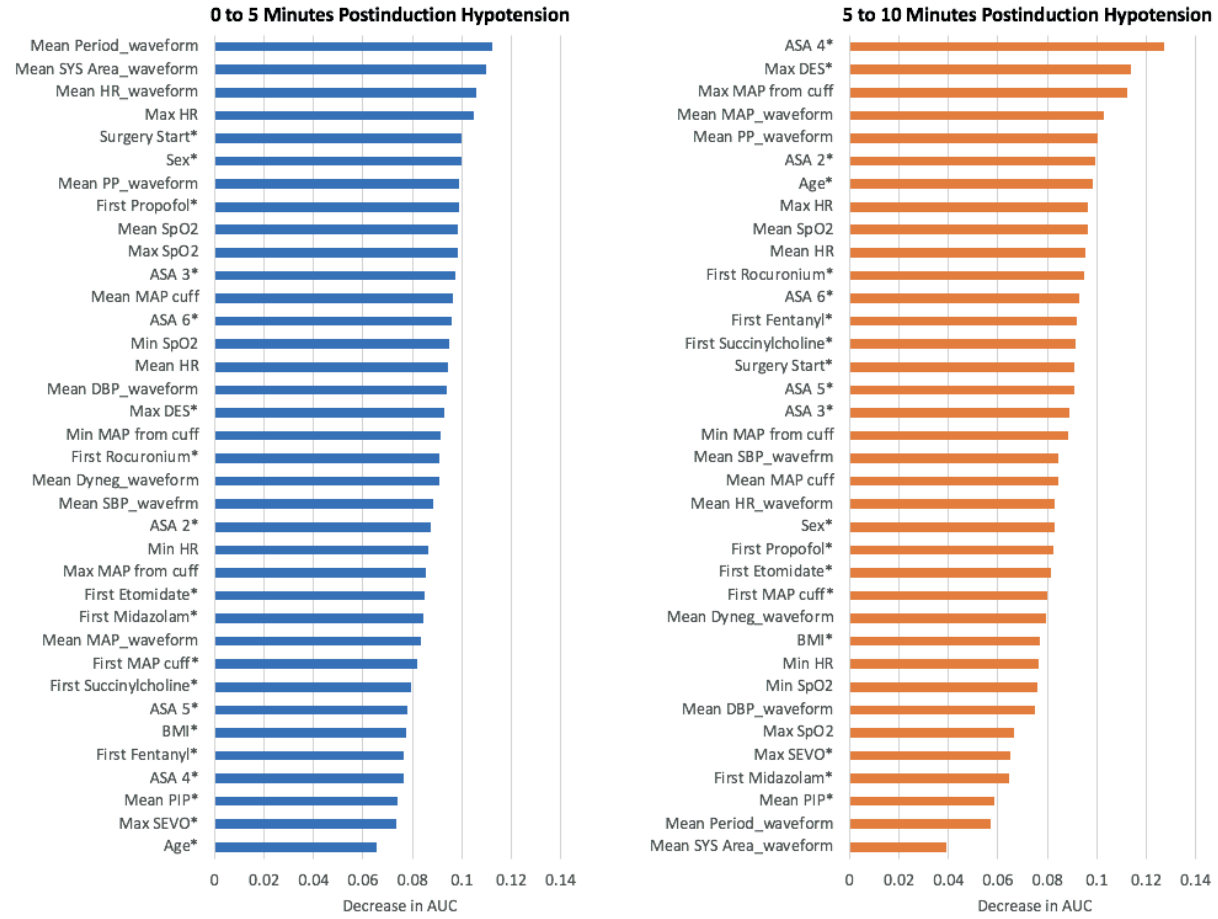


Figure 10. *Feature ablation results for DNN models*

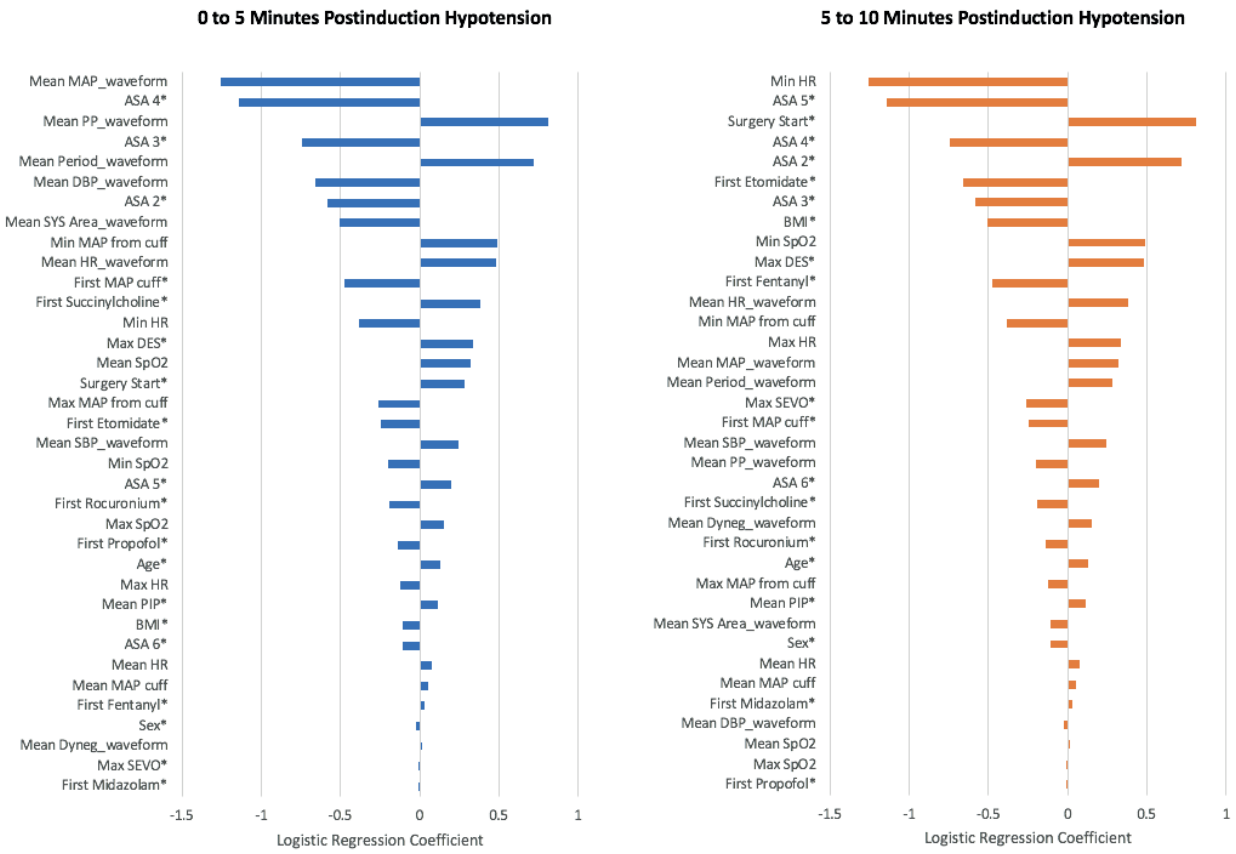


Figure 11. *Logistic regression models coefficients*

The top five DNN features for the 0 to 5 minutes postinduction hypotension model were the mean beat period (waveform), mean systolic area (waveform), mean heart rate (waveform), max heart rate (EMR), and surgery start (EMR; Kendale et al.) (Figure 11). The top five logistic regression features were mean MAP (waveform), ASA 4 (EMR; Kendale et al.), mean pulse pressure (waveform), ASA 3 (EMR; Kendale et al.), and beat period (waveform) (Figure 12). Overall, three of the top five features in both the DNN and logistic regression models were waveform features.

The top five deep neural network features for the 5 to 10 minutes postinduction hypotension model were ASA 4 (EMR; Kendale et al.), maximum desflurane (EMR; Kendale et al.), maximum MAP from cuff (EMR), mean MAP (waveform), and mean pulse pressure

(waveform). The top five logistic regression features were minimum heart rate (EMR), ASA 5 (EMR; Kendale et al.), surgery start (EMR; Kendale et al.), ASA 4 (EMR; Kendale et al.), and ASA 2 (EMR; Kendale et al.). Both the DNN and logistic regression models had ASA 4 as a top feature, however, the logistic regression model had no waveform features in either the top five or the top ten features. In addition, this logistic regression model performed the worst of the 5 to 10 minutes postinduction hypotension models.

AN INTERPRETABLE NEURAL NETWORK FOR PREDICTING POSTOPERATIVE IN-HOSPITAL MORTALITY

We recently showed that deep neural networks (DNNs) using only readily available intraoperative information extracted from the electronic health record can successfully predict postoperative in-hospital mortality with an AUC of 0.91.[44] While DNNs are great machine learning models and often have higher accuracy than more simple models like logistic regression, they are often thought of as a "black box" and not interpretable. In healthcare, intelligible models not only help clinicians to understand the problem and create more targeted action plans, they also help to gain the clinicians' trust. Thus, logistic regression models remain popular in the healthcare space, as they are robust, easy to implement and usually have good performance, as we have also seen in our previous work comparing DNNs to logistic regression.[44] However, logistic regression is limited by the fact that it can only model a linear relationship between the input features and its target response, which may not only be misleading but also not clinically intuitive. For example, both hypervolemia and hypovolemia have been shown to increase the risk of postoperative complications, reflecting a nonlinear

relationship between a patient's volume status and the risk for complications[69]. While DNNs are capable of learning nonlinear relationships, they lack the interpretability of logistic regression.

One method of overcoming the limitations of a linear model such as logistic regression is through generalized additive models (GAMs). Caruana et al. demonstrated GAMs could be applied to real healthcare problems such as pneumonia risk with high accuracy.[70] Through a graphical representation of each model feature's learned contribution to the predicted risk, the interpretable GAMs help to visualize learned patterns and identify new patterns in the data or confirm what clinicians already know. Inspired by GAMs, the same idea can be applied to neural networks through an architecture referred to as Generalized Additive Neural Networks (GANNs).[71] Bras-Geraldes et al. showed GANNs could be used to predict mortality in the ICU with an AUC of 0.83, using 19 features from vital signs, lab values, demographics, admission information, and comorbidities.[72]

Models like DNNs allow for learning the more complex relationship between the input and class label, however, they are not as easily interpretable as logistic regression. In this work, we applied the same idea of the Generalized Additive Neural Networks architecture to allow for interpretability by visualizing the learned feature patterns related to risk of in-hospital mortality.

Data Description

Data Extraction

All data used in this study came from the UCLA Medical Center described previously. The original 87 features from this data set were reduced to 45 features in our previous work,

and ASA was added as a feature in the final model (46 features) that improved model performance.[44] In this study, we used the same features, but also added previously not included features: total anesthesia case time (1 feature); the time spent with MAP below 40, 45, 50, 55, 60, and 65 mmHg (6 features); and HCUP Code Descriptions of the Primary CPT codes (33 features). There were 183 unique HCUP Code Descriptions in our data set, and we selected 33 HCUP Code Descriptions that were present in at least 1 percent of the total data (Appendix B). These HCUP Code Descriptions were then encoded as 33 binary features.

Data Preprocessing

Before model development, missing values for ASA scores were filled with the most common value (ASA 3); missing values for medications administration features were filled with 0; and all other missing values were filled with the means for that feature. Values that were greater than a clinically normal maximum (determined by M.C. and I.H.) were set to a maximum possible, as described in previous work.[44] Finally, all training data were rescaled to have mean 0 and standard deviation 1 per feature. Test data were rescaled with the training data mean and standard deviation.

Model Development

In this work, we were interested in classifying patients at risk of in-hospital mortality utilizing a proposed generalized additive neural network (GANN) architecture (Figure 13).

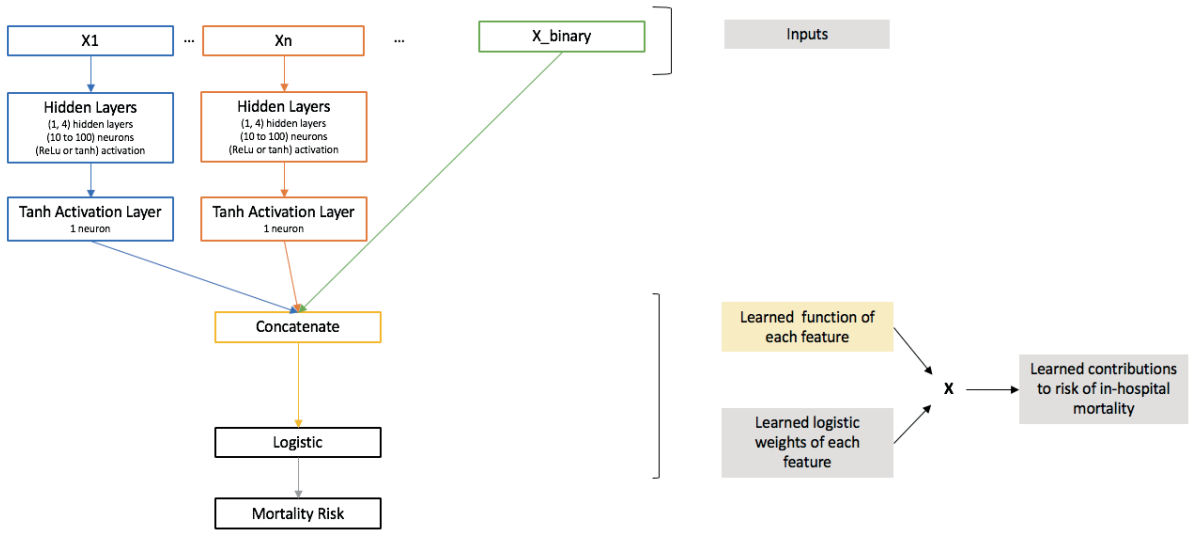


Figure 12. *Proposed generalized additive neural network (GANN) architecture and description of feature contributions calculation, for n individual continuous features vs binary features*

All data was randomly split into 80% for training (n= 47,988) and 20% for test (n=11,997) prior to model development. All GANNs were trained on 80% of the data with 5-fold cross validation to optimize hyperparameters. All models were trained with a batch size of 256 and Adam optimization[67] with default parameters and reduced the learning rate by a factor of 10 when the validation loss stopped improving for a patience of 5 epochs, a batch size of 256, and a maximum of 100 epochs. Dropout (0.25, 0.5, 0.9)[46,50], L2 regularization (0.001, 0.0001) and early stopping with a patience of 5 epochs were used to prevent overfitting. In our GANN architecture, each feature had its own network of hidden layers (1, 4) with (10, 40 to 50, 90, 100) neurons with (rectified linear unit (ReLu), hyperbolic tangent (tanh)) activations (Figure 13). These hidden layers are followed by a last layer with just 1 neuron with a tanh activation. This last tanh layer transforms the previous layer's output into one value and forces the feature's neural network final output to be between -1 and 1. The outputs of all the features'

tanh layers are then concatenated prior to being input into a logistic layer. The feature contributions are calculated as their tanh layer outputs multiplied by their respective logistic weights. Binary features only had a direct connection from input to the logistic layer, and so their feature contributions are calculated as the input value multiplied by their respective logistic weights.

Model Performance Methods

Model performance was assessed using area under the receiver operating characteristic curve (AUC) and average precision (AP). All results reported were calculated on the test set and 95% confidence intervals were calculated using bootstrapping with 1,000 samples. The same training and test sets were used in this work as our previous work on in-hospital mortality for comparison.[44] In addition, as HCUP codes are not immediately available at the end of surgery, we assessed model performance for models developed with and without HCUP features.

All neural network models were developed using Keras.[68] Logistic regression models and performance metrics were calculated with scikit-learn.[48]

Results

The data consisted of 59,985 surgical records, and the % occurrence of in-hospital mortality was 0.81% (n=389) in the training set and 0.72% (n=87) in the test set. The final hyperparameters for each GANN model and feature set combination are described in Table 20.

Table 20. *Final model hyperparameters for each GANN model with and without HCUP category description features*

	# Features	# Hidden Layers	# Neurons	Hidden Layer Activation	Dropout Probability	L2 Lambda
With HCUP Features	88	1	50	tanh	0.5	0.0001
Without HCUP Features	55	1	50	tanh	0.5	0.001

Model Performance

All performance metrics reported below refer to the test set (n=11,997). Area under the receiver operating characteristic curve (AUC ROC) and average precision (AP) are summarized in Table 21. The GANN model with HCUP features had the highest AUC 0.921 (0.895-0.95). Overall, both GANN models had higher AUCs than LR models, however had lower APs. The LR model without HCUP features had the highest AP 0.217 (0.136-0.31).

Table 21. *AUC results of GANN and LR models with and without HCUP features*

Feature Set	Model	AUC	AP
With HCUP Features	GANN	0.921 (0.895-0.95)	0.176 (0.109-0.26)
	LR	0.912 (0.879-0.94)	0.207 (0.127-0.3)
Without HCUP Features	GANN	0.912 (0.883-0.94)	0.197 (0.124-0.29)
	LR	0.906 (0.873-0.94)	0.217 (0.136-0.31)

Interpretability

To assess the interpretability of the GANNs, we visualized the learned contributions of the GANNs vs the learned contributions of the LRs for the models with HCUP features. As described in the methods, the learned contribution of the GANNs for each feature is its last tanh layer's output multiplied by its respective weight from the logistic layer. Since the binary

features have a direct connection from input to the logistic layer, the binary features' learned contributions would be their input values multiplied by their respective weight from the logistic layer. The learned contribution of the LR model is the input value multiplied by its respective weight from the LR model.

In Figure 14, we visualize these contributions and selected a sample of the top 9 contributing features in the GANN model. The top 9 were chosen by selecting the features with the highest mean GANN contribution. We did not include any binary features in this example, such as presence of arterial line, as their visualization would not be as interested, since there would only be 2 values to plot. We see that overall the direction of the learned contributions from both the GANN and LR models were similar, i.e. as MAX_DES increases the contributions for both models decreased. However, while the LR model will always have a linear relationship, the GANN learned non-linear relationships that were unique to each feature. For example, for the feature AVG_MAP_10_MIN we see a non-linear function where GANN contributions increase for MAP < 60 mmHg and for MAP > 60 mmHg. One odd relationship is the one observed between ANES_CASE_HOURS and mortality risk, where with less hours spent under anesthesia there was more contribution to mortality risk. This could be a reflection of the infrequency of extremely high anesthesia case hours (> 10 hours), and that in-hospital mortality patients may not spend significantly longer amounts of time under anesthesia compared to non-mortality patients. In addition, while risk contribution increased with lower MIN_DBP, there was the opposite relationship for AVG_DBP_10_MIN and AVG_DBP, which could indicate that not all summary measures of vital signs are the same, and that these should be taken into consideration when selecting features.

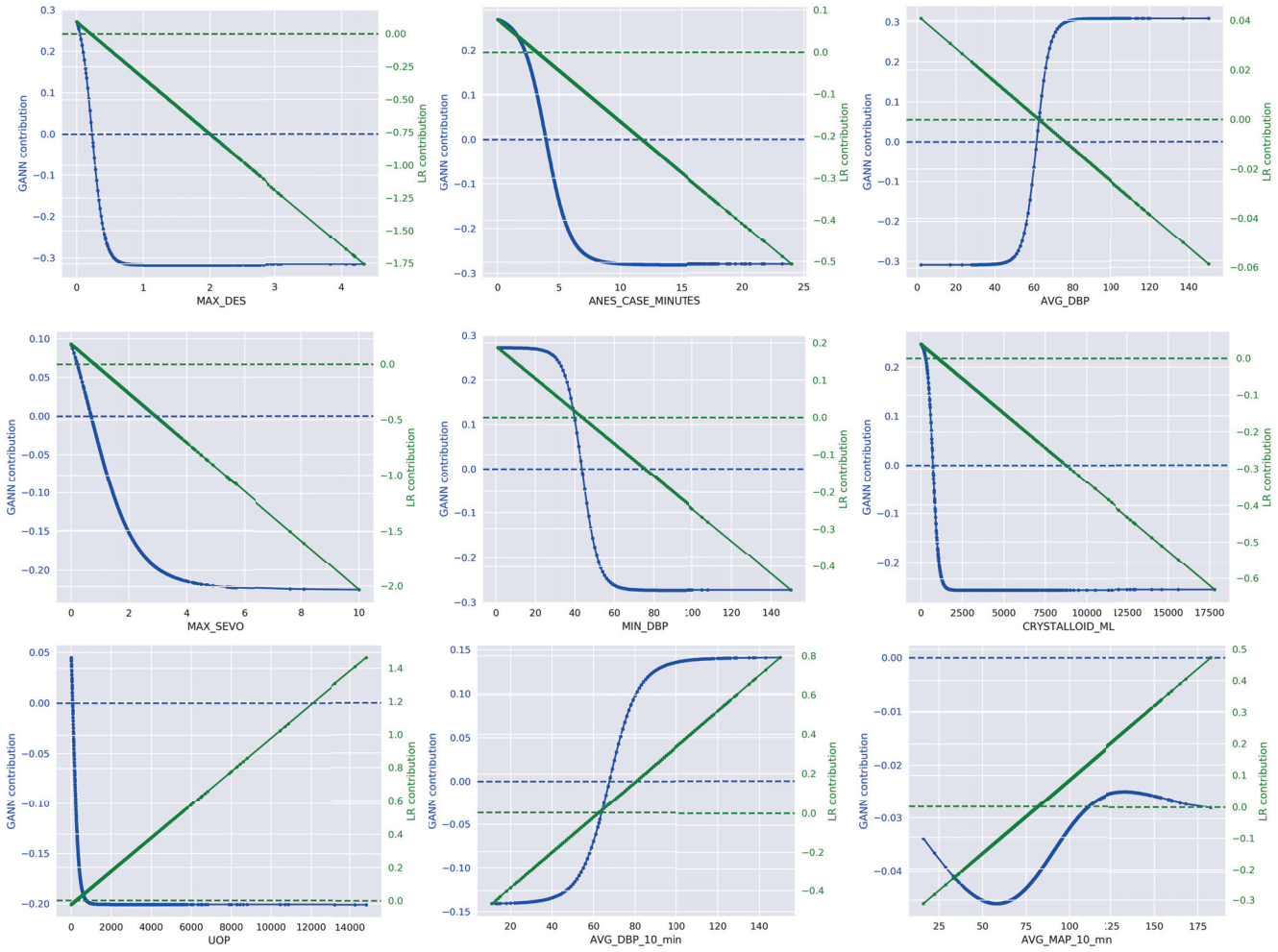


Figure 13. *Sample of 9 continuous features that had the highest mean mortality risk GANN contributions across all patients, in order of highest to lowest (left to right, top to bottom).*

The feature's values for all patients are plotted on the x axis and the respective GANN contribution (blue) on the primary y axis and LR contribution (green) on the secondary y axis. The more negative the risk contribution, the less contribution the respective value has to the risk of mortality.

In our interpretable model, we can also look at the top contributors to a risk of mortality (Table 22). If we look at the top 10 GANN contributions from the best performing GANN with HCUP features for 2 unique in-hospital mortality patients from the test set, we can see that the features that contributed most were different. For example, a high ASA score of 4 contributed highly for Patient Example 1, bud did not show up as a top contributor for Patient Example 2.

Table 22. *Top 10 neural network contributions learned from the best performing GANN model with HCUP features, for 2 in-hospital mortality patient examples from the test set.*

Patient Example 1 (Top 10 Contributions)			Patient Example 2 (Top 10 Contributions)		
Feature	**Value**	**Contribution**	**Feature**	**Value**	**Contribution**
ART_LINE_YN	1	0.993	HCUP_cat_1_YN	1	1.080
ASA_SCORE	4	0.939	ART_LINE_YN	1	0.993
MIN_DBP	22	0.269	MIN_DBP	19	0.271
AGE	81	0.259	MIN_HB	7.6	0.184
AVG_DBP	68	0.234	PHENYLEPHRINE_CURRENT_RATE_MCG_MIN	43	0.177
PHENYLEPHRINE_CURRENT_RATE_MCG_MIN	45	0.191	PHENYLEPHRINE_MAX_RATE_MCG_MIN	43	0.174
PHENYLEPHRINE_MAX_RATE_MCG_MIN	45	0.176	MIN_MAP	17	0.132
MIN_MAP	30	0.122	AGE	69	0.094
AVG_HR	95	0.104	AVG_DBP_10_min	72	0.043
AVG_DBP_10_min	74	0.060	ANES_CASE_HOURS	3.9	0.001

CONCLUSIONS AND RECOMMENDED FUTURE WORK

Modern medicine requires tools to change from “reactive” patient management to a more “prospective” or “proactive” approach. However, these tools need to not only be accurate, but also interpretable and therefore actionable by the clinician. There are 2 ways to approach prospective healthcare: 1) A one-time risk classification to help allocate hospital resources more efficiently to ensure patients receive necessary critical care and 2) A continuous, real time risk classification to avoid onset of complications altogether.

Throughout this thesis, I have shown the deep neural networks can perform with high accuracy when compared to currently used common risk scores for adverse outcomes, such as AKI, reintubation, and mortality. We were also able to develop models for predicting acute intraoperative events, such as hypotension, that can increase the risk of those same adverse outcomes. In addition, while logistic regression performed comparably for many of the

problems, we have also shown its limitations. Logistic regression is often preferred in the medical field due to its easy implementation and interpretability. However, it can only impose linear relationships between features and response labels, such as the risk of in-hospital mortality. One of the drawbacks of deep neural networks has been its "black box" reputation. In response, we proposed a generalized additive neural network (GANN) architecture to learn nonlinear patterns in the data that resembled more of clinical intuition and can be interpreted easily. With these GANNS, we were able to show that we can automatically learn clinically intuitive relationships without domain knowledge or extra featurization, give clinicians interpretability, and maintain model performance.

While one-time risk classifications are useful, clinicians need continuous indicators for managing patients in a real-time, acute setting. We made a first attempt at this by developing a Hypotension Prediction Index (HPI™), which predicts the probability of future onset of hypotension using features from the arterial blood pressure waveform. HPI™ is currently FDA-approved and a commercial product (Edwards Lifesciences, Irvine, CA).

Future work would include leveraging the >19,000 surgical patients worth of data we have collected from the UCI Medical Center and prepared for research-use. Much of the initial years of my research were spent on gathering, cleaning, and understanding this data set. While we were able to successfully unify EMR and arterial blood pressure waveform data to predict postinduction hypotension using deep neural networks with this dataset, we were severely limited by the number of includable patients and thus did not get good results with more complex models such as convolutional neural networks that would require more data. Next steps would be to use more frequently available inputs, such as HR, SpO2, and blood pressure

cuff measurements that are available for all patients, in models like LSTMs which would incorporate time. There is also much to be learned from the waveforms we have collected. Instability and subsequent shock are very complicated physiologically, and thus may require a more complex input that is not captured in the static features like HR and BP which are extracted from the waveform to begin with. Future work could also include utilizing unsupervised learning on the waveforms to determine periods of unique hemodynamic patterns, and see if those correlate to what is occurring or will occur in the patient. Supervised learning could also be used if a definition for shock can be extracted from the EMR.

This thesis work can be summarized as follows: by utilizing readily available patient monitoring data, we can build predictive models to help inform clinicians in the management of their patients as well as help them avoid poor outcomes or acute events. These models need to be easily implementable, interpretable, and accurate. Once trained, most models, even ones large and complex as a deep neural network, are easily implementable. The models created throughout this thesis were developed initially with only the classic approach of just meeting the accuracy need. However, we then shifted our approach to develop for interpretability in addition to accuracy. Next steps would be to leverage the large UCIMC dataset we have created to continue to develop more interpretable models that can also be displayed continuously, in real-time during patient management.

REFERENCES

1. Hou SH, Bushinsky DA, Wish JB, Cohen JJ, Harrington JT: Hospital-acquired renal insufficiency: A prospective study. Am J Med 1983; 74:243–8
2. Pearse RM, Harrison DA, James P, Watson D, Hinds C, Rhodes A, Grounds MR, Bennett DE: Identification and characterisation of the high-risk surgical population in the United Kingdom. Crit Care 2006; 10:1–6
3. Pearse RM, Moreno RP, Bauer P, Pelosi P, Metnitz P, Spies C, Vallet B, Vincent J-L, Hoeft A, Rhodes A, group for the groups of the of, the of Anaesthesiology E: Mortality after surgery in Europe: a 7 day cohort study. Lancet 2012; 380:1059–65
4. Monk TG, Saini V, Weldon CB, Sigl JC: Anesthetic Management and One-Year Mortality After Noncardiac Surgery. Anesth Analg 2005; 100:4
5. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G, Heldt T, Kyaw TH,

Moody B, Mark RG: Multiparameter Intelligent Monitoring in Intensive Care II: A public-access intensive care unit database*. Crit Care Med 2011; 39:952
6. Khuri SF, Henderson WG, DePalma RG, Mosca C, Healey NA, Kumbhani DJ: Determinants of Long-Term Survival After Major Surgery and the Adverse Effect of Postoperative Complications. Trans . Meet Am Surg Assoc 2005; 123:32–48
7. Jencks SF, Williams M V, Coleman EA: Rehospitalizations among Patients in the Medicare Fee-for-Service Program. N Engl J Med 2009; 360:1418–28
8. Hines AL, Barrett ML, Jiang J, Steiner CA: Conditions with the largest number of adult hospital readmissions by payer. Healthc Cost Util Proj Stat Briefs 2014 at <http://www.ncbi.nlm.nih.gov/books/NBK206781/>
9. American Society of Anesthesiologists AS of AAS of A: New classification of physical status. Anesthesiology 1963 at <http://scholar.google.com/scholar?q=New classification of physical status&btnG=&hl=en&num=20&as_sdt=0%2C22>
10. Gawande AA, Kwaan MR, Regenbogen SE, Lipsitz SA, Zinner MJ: An Apgar Score for Surgery. J Am Coll Surg 2007; 204:201–8
11. Knaus WA, Draper EA, Wagner DP, Zimmerman JE: APACHE II: a severity of disease classification system. Crit Care Med 1985; 13:818–29
12. Gardner-Thorpe J, Love N, Wrightson J, Walsh S, Keeling N: The Value of Modified Early Warning Score (MEWS) in Surgical In-Patients: A Prospective Observational Study. Ann R Coll Surg Engl 2006; 88:571–5
13. Berger T, Green J, Horeczko T, Hagar Y, Garg N, Suarez A, Panacek E, Shapiro N: Shock Index and Early Recognition of Sepsis in the Emergency Department: Pilot Study. West J Emerg Med 2013; 14:168–74
14. Manach Y, Collins G, Rodseth R, Bihan-Benjamin C, Biccard B, Riou B, Devereaux PJ, Landais P: Preoperative Score to Predict Postoperative Mortality (POSPOM). Anesthesiology 2016; 124:570–9
15. Rothman MJ, Rothman SI, Beals J: Development and validation of a continuous measure of patient condition using the Electronic Medical Record. J Biomed Inform 2013; 46:837–48
16. Thakar C V, Arrigain S, Worley S, Yared J-P, Paganini EP: A clinical score to predict acute renal failure after cardiac surgery. J Am Soc Nephrol 2004; 16:162–8
17. Kheterpal S, Tremper KK, Englesbe MJ, O'Reilly M, Shanks AM, Fetterman DM, Rosenberg AL, Swartz RD: Predictors of Postoperative Acute Renal Failure after Noncardiac Surgery in Patients with Previously Normal Renal Function. Anesthesiology 2007; 107:892
18. Kheterpal S, Tremper KK, Heung M, Rosenberg AL, Englesbe M, Shanks AM, Campbell DA: Development and Validation of an Acute Kidney Injury Risk Index for Patients Undergoing General Surgery: Results from a National Data Set. Anesthesiology 2009; 110:505
19. Badin J, Boulain T, Ehrmann S, Skarzynski M, Bretagnol A, Buret J, Benzekri-Lefevre D, Mercier E, Runge I, Garot D, Mathonnet A, Dequin P-F, Perrotin D: Relation between mean arterial pressure and renal function in the early phase of shock: a prospective, explorative cohort study. Crit Care 2011; 15:1–12
20. Reich DL, Bodian CA, Krol M, Kuroda M, Osinski T, Thys DM: Intraoperative Hemodynamic Predictors of Mortality, Stroke, and Myocardial Infarction After Coronary Artery Bypass Surgery. Anesth Analg 1999; 89:814

21. Sessler DI, Sigl JC, Kelley SD, Chamoun NG, Manberg PJ, Saager L, Kurz A, Greenwald S: Hospital Stay and Mortality Are Increased in Patients Having a “Triple Low” of Low Blood Pressure, Low Bispectral Index, and Low Minimum Alveolar Concentration of Volatile Anesthesia. Anesthesiology 2012; 116:1195
22. Walsh M, Devereaux PJ, Garg AX, Kurz A, Turan A, Rodseth RN, Cywinski J, Thabane L, Sessler DI: Relationship between intraoperative mean arterial pressure and clinical outcomes after noncardiac surgery: Toward an empirical definition of hypotension. Anesthesiology 2013; 119:507–15
23. Brady K, Hogue CW: Intraoperative Hypotension and Patient Outcome: Does “One Size Fit All?” Anesthesiology 2013; 119:495
24. Sun LY, Wijeysundera DN, Tait GA, Beattie SW: Association of Intraoperative Hypotension with Acute Kidney Injury after Elective Noncardiac Surgery. Anesthesiology 2015; 123:515–23
25. Dimick JB, Chen SL, Taheri PA, Henderson WG, Khuri SF, Campbell DA: Hospital costs associated with surgical complications: A report from the private-sector National Surgical Quality Improvement Program. J Am Coll Surg 2004; 199:531–7
26. Brueckmann B, Villa-Uribe JL, Bateman BT, Grosse-Sundrup M, Hess DR, Schlett CL, Eikermann M: Development and Validation of a Score for Prediction of Postoperative Respiratory Complications. Anesthesiology 2013; 118:1276
27. Francis NK, Mason J, Salib E, Allanby L, Messenger D, Allison AS, Smart NJ, Ockrim JB: Factors predicting 30-day readmission after laparoscopic colorectal cancer surgery within an enhanced recovery programme. Color Dis 2015; 17
28. Gozal D, Daniel JM, Dohanich GP: Behavioral and anatomical correlates of chronic episodic hypoxia during sleep in the rat. J Neurosci 2001; 21:2442–50
29. Sushmita S, Hasan: Predicting 30-Day Risk and Cost of" All-Cause" Hospital Readmissions. Work Thirtieth AAAI Conf Artif Intell Expand Boundaries Heal Informatics Using AI 2016; WS-16-08
30. Chen L, Dubrawski A, Clermont G, Hravnak M, Pinsky M: Modelling Risk of Cardio-Respiratory Instability as a Heterogeneous Process. AMIA Annu Symp Proc 2015 at <http://scholar.google.com/scholar?q=Modelling Risk of Cardio-Respiratory Instability as a Heterogeneous Process&btnG=&hl=en&num=20&as_sdt=0%2C22>
31. Razavian N, Sontag D: TEMPORAL CONVOLUTIONAL NEURAL NETWORKS FOR DIAGNOSIS FROM LAB TESTS. arXiv 2016; 1511.07938
32. Che Z, Purushotham S, Khemani R, Liu Y: Distilling Knowledge from Deep Networks with Applications to Healthcare Domain. arXiv 2015; 1512.03542
33. Nguyen, Tran, Wickramasinghe: Deepr: A Convolutional Net for Medical Records. arXiv 2016; 1607.07519
34. Lipton Z, Kale D, Elkan C, Wetzel R: Learning to Diagnose with LSTM Recurrent Neural Networks. Int Conf Learn Represent 2016 at <http://scholar.google.com/scholar?q=Learning to Diagnose with LSTM Recurrent Neural Networks&btnG=&hl=en&num=20&as_sdt=0%2C22>
35. Schmidhuber J: Neural Networks. Reviews 2015; 61:85–117
36. LeCun Y, Bengio Y, Hinton G: Deep learning. Nature 2015; 521:436–44
37. Hornik K, Stinchcombe M, White H: Multilayer feedforward networks are universal

approximators. Neural Networks 1989; 2:359–66
38. Lipton Z, Berkowitz J, Elkan C: A Critical Review of Recurrent Neural Networks for Sequence Learning. arXiv 2015; 1506.00019
39. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang G: Recent Advances in Convolutional Neural Networks. arXiv 2016; 1512.07108
40. 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
41. 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
42. Dalton JE, Kurz A, Turan A, Mascha EJ, Sessler DI, Saager L: Development and Validation of a Risk Quantification Index for 30-Day Postoperative Mortality and Morbidity in Noncardiac Surgical Patients. Anesthesiology 2011; 114:1336
43. Sigakis MJG, 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
44. Lee CK, Hofer I, Gabel E, Baldi P, Cannesson M: Development and Validation of a Deep Neural Network Model for Prediction of Postoperative In-hospital Mortality. Anesthesiology 2018; 129:649–62
45. 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
46. Srivastava N, Hinton G, Krizhevsky A, Salakhutdinov R: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. 2014
47. Chollet F: Keras <https://github.com/fchollet/keras>. GitHub 2015 at <https://github.com/fchollet/keras>
48. Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, Vanderplas J, Cournapeau D, Varoquaux G, Gramfort A, Thirion B, Grisel O, Dubourg V, Passos A, Brucher M, Perrot M, Duchesnay É: Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011; 12:2825–30
49. Sessler DI, Meyhoff CS, Zimmerman NM, Mao G, Leslie K, Vásquez SM, Balaji P, Alvarez-Garcia J, Cavalcanti AB, Parlow JL, Rahate P V., Seeberger MD, Gossetti B, Walker SA, Premchand RK, Dahl RM, Duceppe E, Rodseth R, Botto F, Devereaux PJ: Period-dependent Associations between Hypotension during and for Four Days after Noncardiac Surgery and a Composite of Myocardial Infarction and Death. Anesthesiology 2018; 128:317–27
50. Baldi P, Sadowski P: The dropout learning algorithm. Artif Intell 2014; 210:78–122
51. Raschka S: MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. J Open Source Softw 2018; 3:638
52. Wertheim JA, Petrowsky H, Saab S, Kupiec-Weglinski JW, Busuttil RW: Major challenges limiting liver transplantation in the United States 2011; 11:pp 1773–84
53. Dutkowski P, Linecker M, Deoliveira ML, Müllhaupt B, Clavien PA: Challenges to liver transplantation and strategies to improve outcomes 2015; 148:pp 307–23
54. Wiesner R, Edwards E, Freeman R, Harper A, Kim R, Kamath P, Kremers W, Lake J,

Howard T, Merion RM, Wolfe RA, Krom R, Colombani PM, Cottingham PC, Dunn SP, Fung JJ, Hanto DW, McDiarmid S V., Rabkin JM, Teperman LW, Turcotte JG, Wegman LR: Model for end-stage liver disease (MELD) and allocation of donor livers. Gastroenterology 2003; 124:91–6
55. Kamath PS, Kim WR: The Model for End-stage Liver Disease (MELD) 2007; 45:pp 797–805
56. Desai NM, Mange KC, Crawford MD, Abt PL, Frank AM, Markmann JW, Velidedeoglu E, Chapman WC, Markmann JF: Predicting outcome after liver transplantation: Utility of the model for end-stage liver disease and a newly derived discrimination function. Transplantation 2004; 77:99–106
57. Dutkowski P, Oberkofler CE, Slankamenac K, Puhan MA, Schadde E, Müllhaupt B, Geier A, Clavien PA: Are there better guidelines for allocation in liver transplantation?: A novel score targeting justice and utility in the model for end-stage liver disease era, Annals of Surgery. 2011, pp 745–53
58. Rana A, Hardy MA, Halazun KJ, Woodland DC, Ratner LE, Samstein B, Guarrera J V., Brown RS, Emond JC: Survival Outcomes Following Liver Transplantation (SOFT) score: A novel method to predict patient survival following liver transplantation. Am J Transplant 2008; 8:2537–46
59. Massie AB, Kuricka LM, Segev DL: Big data in organ transplantation: Registries and administrative claims 2014; 14:pp 1723–30
60. Bijker JB, Klei WA van, Kappen TH, Wolfswinkel L van, Moons KGM, Kalkman CJ: Incidence of Intraoperative Hypotension as a Function of the Chosen Definition: Literature Definitions Applied to a Retrospective Cohort Using Automated Data Collection. Anesthesiology 2007; 107:213
61. Hatib F, Jian Z, Buddi S, Lee C, Settels J, Sibert K, Rinehart J, Cannesson M: Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis. Anesthesiology 2018; 129:663–74
62. Südfeld S, Brechnitz S, Wagner JY, Reese PC, Pinnschmidt HO, Reuter DA, Saugel B: Post-induction hypotension and early intraoperative hypotension associated with general anaesthesia. Br J Anaesth 2017; 119:57–64
63. Hanss R, Renner J, Ilies C, Moikow L, Buell O, Steinfath M, Scholz J, Bein B: Does heart rate variability predict hypotension and bradycardia after induction of general anaesthesia in high risk cardiovascular patients?, Anaesthesia. 2008, pp 129–35
64. Reich DL, Hossain S, Krol M, Baez B, Patel P, Bernstein A, Bodian CA: Predictors of hypotension after induction of general anesthesia. Anesth Analg 2005; 101:622–8, table of contents
65. Kendale S, Kulkarni P, Rosenberg AD, Wang J: Supervised Machine-learning Predictive Analytics for Prediction of Postinduction Hypotension. Anesthesiology 2018; 129:675–88
66. Vest AN, Poian G Da, Li Q, Liu C, Nemati S, Shah AJ, Clifford GD: An open source benchmarked toolbox for cardiovascular waveform and interval analysis. Physiol Meas 2018; 39
67. Kingma DP, Lei Ba J: ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION.
68. Cholett F: Keras 2015 at <https://keras.io>
69. Vincent JL, Pelosi P, Pearse R, Payen D, Perel A, Hoeft A, Romagnoli S, Ranieri VM, Ichai C, Forget P, Rocca G Della, Rhodes A: Perioperative cardiovascular monitoring of high-risk

patients: A consensus of 12 2015; 19
70. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N: Intelligible Models for HealthCare, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15. New York, New York, USA, ACM Press, 2015, pp 1721–30 doi:10.1145/2783258.2788613
71. Potts WJE: Generalized Additive Neural Networks. 1999
72. Brás-Geraldes C, Papoila A, Xufre P: Generalized additive neural network with flexible parametric link function: model estimation using simulated and real clinical data. Neural Comput Appl 2017:1–18 doi:10.1007/s00521-017-3105-6

Appendix A. Description of liver transplant features

feature	feature extraction description	categorical_YN	unq_categories	keep feature in reduced feature set
abo_A	if type A, A1, or A2	Y	[0.0, 1.0]	remove due to >95 percent zero
abo_AB	if type AB, A1B, or A2B	Y	[0.0, 1.0]	keep
abo_B	if type B	Y	[0.0, 1.0]	keep
abo_don_A	see above	Y	[0.0, 1.0]	keep
abo_don_AB	see above	Y	[0.0, 1.0]	remove due to >95 percent zero
abo_don_B	see above	Y	[0.0, 1.0]	keep

abo_don_O	if type O	Y	[1.0, 0.0]	keep
abo_mat		Y	[1.0, 2.0, 3.0]	keep
abo_O	if type O	Y	[1.0, 0.0]	keep
age		N		keep
age_don		N		keep
albumin_tx		N		keep
antihype_don		Y	[1.0, 0.0, nan]	keep
arginine_don		Y	[1.0, 0.0, nan]	keep
ascites_tx		N		keep
bact_perit_tcr		Y	[0.0, nan, 1.0]	keep
bmi_calc		N		keep
bmi_don_calc		N		keep
bmi_tcr		N		keep
bun_don		N		keep
cardarrest_neuro		Y	[0.0, 1.0, nan]	keep
cdc_risk_hiv_don		Y	[0.0, 1.0, nan]	keep
citizenship	if not equal to 1, set to 1; otherwise, 0	Y	[0.0, 1.0]	remove due to >95 percent zero
citizenship_don		Y	[0.0, 1.0]	remove due to >95 percent zero
clin_infect_don		Y	[1.0, 0.0, nan]	keep
cmv_don		Y	[0.0, 1.0, nan]	keep
cmv_igg		Y	[1.0, 0.0, nan]	keep
cmv_igm		Y	[nan, 0.0, 1.0]	remove due to >50 percent null
cmv_status		Y	[1.0, 0.0, nan]	keep
cod_cad_don_1	if cod_cad_don == 1, set to 1	Y	[1.0, 0.0]	keep
cod_cad_don_2	if cod_cad_don == 2, set to 1	Y	[0.0, 1.0]	keep
cod_cad_don_3	if cod_cad_don == 3, set to 1	Y	[0.0, 1.0]	keep
cod_cad_don_4	if cod_cad_don == 4, set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
cold_isch		N		keep
coronary1	if "coronary_angio_norm_don" equals 0 and "coronoary_angio_don" equals 1, set to 1; if "coronary_angio_norm_don" equals 1 and "coronoary_angio_don" equals 1, set to 2;	Y	[0.0, 2.0, 1.0]	keep
creat_don		N		keep
creat_tx		N		keep

dayswait_chron		N		keep
ddavp_don		Y	[0.0, 1.0, nan]	keep
death_circum_don	if not equal to 6, set to 1	Y	[1.0, 0.0]	keep
death_mech_don	if equal to 12, set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
dgn_tcr_AHN	see diag_* description; except looking at "dgn_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero
dgn_tcr_autoimmune	see diag_* description; except looking at "dgn_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero
dgn_tcr_cryptogenic	see diag_* description; except looking at "dgn_tcr"	Y	[0.0, 1.0]	keep
dgn_tcr_etoh	see diag_* description; except looking at "dgn_tcr"	Y	[0.0, 1.0]	keep
dgn_tcr_etoh_hcv	see diag_* description; except looking at "dgn_tcr"	Y	[0.0, 1.0]	keep
dgn_tcr_HBV	see diag_* description; except looking at "dgn_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero
dgn_tcr_HCC	see diag_* description; except looking at "dgn_tcr"	Y	[0.0, 1.0]	keep
dgn_tcr_HCV	see diag_* description; except looking at "dgn_tcr"	Y	[0.0, 1.0]	keep
dgn_tcr_NASH	see diag_* description; except looking at "dgn_tcr"	Y	[0.0, 1.0]	keep
dgn_tcr_PBC	see diag_* description; except looking at "dgn_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero
dgn_tcr_PSC	see diag_* description; except looking at "dgn_tcr"	Y	[1.0, 0.0]	remove due to >95 percent zero
dgn2_tcr_AHN	see diag_* description; except looking at "dgn2_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero
dgn2_tcr_autoimmune	see diag_* description; except looking at "dgn2_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero
dgn2_tcr_cryptogenic	see diag_* description; except looking at "dgn2_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero
dgn2_tcr_etoh	see diag_* description; except looking at "dgn2_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero
dgn2_tcr_etoh_hcv	see diag_* description; except looking at "dgn2_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero

dgn2_tcr_HBV	see diag_* description; except looking at "dgn2_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero
dgn2_tcr_HCC	see diag_* description; except looking at "dgn2_tcr"	Y	[0.0, 1.0]	keep
dgn2_tcr_HCV	see diag_* description; except looking at "dgn2_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero
dgn2_tcr_NASH	see diag_* description; except looking at "dgn2_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero
dgn2_tcr_PBC	see diag_* description; except looking at "dgn2_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero
dgn2_tcr_PSC	see diag_* description; except looking at "dgn2_tcr"	Y	[0.0, 1.0]	remove due to >95 percent zero
diab	if not equal to 1, set to 1; otherwise, 0	Y	[0.0, 1.0]	keep
diabdur_don		N		keep
diabetes_don		Y	[0.0, 1.0, nan]	keep
diag_AHN	if "diag" is in [4100, 4101, 4102, 4103, 4104, 4105, 4106, 4107, 4108, 4110, 4217], set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
diag_autoimmune	if "diag" is in [4212], set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
diag_cryptogenic	if "diag" is in [4213, 4208], set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
diag_etoh	if "diag" is in [4215], set to 1	Y	[0.0, 1.0]	keep
diag_etoh_hcv	if "diag" is in [4216], set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
diag_HBV	if "diag" is in [4202, 4592], set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
diag_HCC	if "diag" is in [4400, 4401, 4402], set to 1	Y	[0.0, 1.0]	keep
diag_HCV	if "diag" is in [4204, 4593], set to 1	Y	[0.0, 1.0]	keep
diag_NASH	if "diag" is in [4214], set to 1	Y	[0.0, 1.0]	keep
diag_PBC	if "diag" is in [4220], set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
diag_PSC	if "diag" is in [4240, 4241, 4242, 4245], set to 1	Y	[1.0, 0.0]	remove due to >95 percent zero
dial_tx		Y	[0.0, 1.0, nan]	keep
distance		N		keep
ebv_igg_cad_don		Y	[1.0, nan, 0.0]	keep

ebv_igm_cad_don		Y	[0.0, nan, 1.0]	remove due to >95 percent zero
ebv_serostatus		Y	[1.0, 0.0, nan]	keep
ecd_donor		Y	[0.0, 1.0]	keep
education		N		keep
enceph_tx		N		keep
end_stat	if equal to 6010 or 6011, set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
ethcat_1	if "ethcat" = 1	Y	[0.0, 1.0]	keep
ethcat_2	if "ethcat" = 2	Y	[1.0, 0.0]	keep
ethcat_4	if "ethcat" = 4	Y	[0.0, 1.0]	keep
ethcat_5	if "ethcat" = 5	Y	[0.0, 1.0]	remove due to >95 percent zero
ethcat_don_1	see above	Y	[1.0, 0.0]	keep
ethcat_don_2	see above	Y	[0.0, 1.0]	keep
ethcat_don_4	see above	Y	[0.0, 1.0]	keep
ethcat_don_5	see above	Y	[0.0, 1.0]	remove due to >95 percent zero
ethcat_don_other	if "ethcat_don" >= 6, set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
ethcat_other	if "ethcat" >= 6, set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
ever_approved		Y	[nan, 1.0, 0.0]	remove due to >50 percent null
exc_case		Y	[0.0, 1.0]	keep
exc_diag_id_cat1	if "exc_diag_id" equals 1, 3, or 10, set to 1	Y	[0.0, 1.0]	keep
exc_diag_id_cat2	if "exc_diag_id" equals 2, set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
exc_diag_id_cat3	if "exc_diag_id" equals 4, set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
exc_diag_id_cat4	if "exc_diag_id" equals 5, set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
exc_diag_id_cat5	if "exc_diag_id" equals 6 or 12, set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
exc_diag_id_cat6	if "exc_diag_id" equals 11, set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
exc_diag_id_cat7	if "exc_diag_id" equals 9, set to 1	Y	[0.0, 1.0]	keep
exc_ever		Y	[0.0, 1.0]	keep
exc_hcc	HCC = 1, non-HCC = 0	Y	[0.0, 1.0]	keep
final_inr		N		keep
final_serum_sodium		N		keep
func_stat_tcr	Replaced [2, 2040, 2050, 2060, 2070] as 1; Replaced [3, 2010, 2020, 2030] as 2; otherwise 0 if not empty	N		keep

func_stat_trr	Replaced [2, 2040, 2050, 2060, 2070] as 1; Replaced [3, 2010, 2020, 2030] as 2; otherwise 0 if not empty	N		keep
gender		Y	[0.0, 1.0]	keep
gender_don		Y	[0.0, 1.0]	keep
hbv_core		Y	[0.0, 1.0, nan]	keep
hbv_core_don		Y	[0.0, 1.0, nan]	keep
hbv_sur_antigen		Y	[0.0, 1.0, nan]	remove due to >95 percent zero
hbv_sur_antigen_don		Y	[0.0, 1.0, nan]	remove due to >95 percent zero
hcc_ever_appr		Y	[nan, 0.0, 1.0]	remove due to >50 percent null
hcv_serostatus		Y	[0.0, 1.0, nan]	keep
hematocrit_don		N		keep
hep_c_anti_don		Y	[0.0, 1.0, nan]	remove due to >95 percent zero
heparin_don		Y	[1.0, 0.0, nan]	keep
hgt_cm_calc		N		keep
hgt_cm_don_calc		N		keep
hgt_cm_tcr		N		keep
hist_cancer_don		Y	[0.0, 1.0, nan]	remove due to >95 percent zero
hist_cig_don		Y	[1.0, 0.0, nan]	keep
hist_cocaine_don		Y	[1.0, 0.0, nan]	keep
hist_insulin_dep_don		Y	[nan, 1.0, 0.0]	remove due to >50 percent null
hist_oth_drug_don		Y	[1.0, 0.0, nan]	keep
history_mi_don		Y	[0.0, 1.0, nan]	remove due to >95 percent zero
hypertens_dur_don	if "hist_hypertens_don" equals 0, set to 1	N		keep
index2	Created by Brent: This variable indicates the number of liver transplants the patient has ever had previously	N		keep
init_age		N		keep
init_albumin		N		keep
init_ascites		N		keep
init_bilirubin		N		keep
init_bmi_calc		N		keep
init_dialysis_prior_week		Y	[nan, 0.0, 1.0]	remove due to >95 percent zero
init_enceph	replaced 4 with null	N		keep

init_hgt_cm		N		keep
init_inr		N		keep
init_meld_peld_lab_score		N		keep
init_serum_creat		N		keep
init_serum_sodium		N		keep
init_stat	if equal to 6010 or 6011, set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
init_wgt_kg		N		keep
inotrop_support_don		Y	[1.0, 0.0, nan]	keep
inr_tx		N		keep
insulin_dep_don	if not equal to 1, set to 1; otherwise, 0	Y	[1.0, 0.0]	keep
insulin_don		Y	[1.0, 0.0, nan]	keep
life_sup_tcr		Y	[0.0, 1.0, nan]	remove due to >95 percent zero
life_sup_trr		Y	[0.0, 1.0]	keep
lityp	if equal to 20, set to 1; else 0 if not null	Y	[1.0, 0.0, nan]	keep
macro_fat_li_don	if "macro_fat_li_don" < 30, set to 1; if "li_biopsy"=1 AND "macro_fat_li_don" is null, set to 1; if "macro_fat_li_don" >= 30, set to 2; if "li_biopsy" is null OR "li_biopsy" equals 0, set to 0	Y	[0.0, 1.0, 2.0]	keep
malig		Y	[0.0, 1.0, nan]	keep
malig_tcr		Y	[0.0, 1.0, nan]	keep
malig_type	if not in [4096 to 8192] and not null, set to 0; otherwise 1	Y	[nan, 1.0, 0.0]	remove due to >50 percent null
med_cond_trr		Y	[3.0, 2.0, 1.0]	keep
meld_diff_reason_cd_1	if "meld_diff_reason_cd" == 15 or 16, set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
meld_diff_reason_cd_2	if "meld_diff_reason_cd" ==8, set to 1	Y	[0.0, 1.0]	keep
meld_peld_lab_score		N		keep
micro_fat_li_don	see "macro_*" for transformations	Y	[0.0, 2.0, 1.0]	keep
non_hrt_don		Y	[0.0, 1.0, nan]	keep
num_prev_tx		N		keep
on_vent_trr		Y	[0.0, 1.0]	keep
oth_life_sup_tcr		Y	[0.0, 1.0]	remove due to >95 percent zero
oth_life_sup_trr		Y	[0.0, 1.0]	remove due to >95 percent zero

ph_don		N		keep
portal_vein_tcr		Y	[nan, 0.0, 1.0]	remove due to >95 percent zero
portal_vein_trr		Y	[0.0, 1.0, nan]	keep
prev_ab_surg_tcr		Y	[0.0, 1.0, nan]	keep
prev_ab_surg_trr		Y	[0.0, 1.0, nan]	keep
prev_tx		Y	[0.0, 1.0]	keep
pri_payment_tcr		Y	[1.0, 0.0, nan]	keep
pri_payment_trr		Y	[1.0, 0.0]	keep
protein_urine		Y	[1.0, 0.0, nan]	keep
prvtxdif	filled null with 0	N		keep
pt_diuretics_don		Y	[1.0, 0.0, nan]	keep
pt_oth_don		Y	[1.0, 0.0, nan]	keep
pt_steroids_don		Y	[1.0, 0.0, nan]	keep
pt_t3_don		Y	[0.0, nan, 1.0]	remove due to >95 percent zero
pt_t4_don		Y	[1.0, 0.0, nan]	keep
recov_out_us		Y	[0.0, 1.0]	remove due to >95 percent zero
resuscit_dur	filled null with 0	N		keep
sgot_don		N		keep
sgpt_don		N		keep
share_ty	if 3 or 4, set to 0; if 5 or 6, set to 1	Y	[0.0, 1.0]	remove due to >95 percent zero
tattoos		Y	[1.0, 0.0, nan]	keep
tbili_don		N		keep
tbili_tx		N		keep
tipss_tcr		Y	[nan, 0.0, 1.0]	keep
tipss_trr		Y	[0.0, 1.0, nan]	keep
vasodil_don		Y	[0.0, 1.0, nan]	keep
vdrl_don		Y	[0.0, 1.0, nan]	remove due to >95 percent zero
ventilator_tcr		Y	[0.0, 1.0]	remove due to >95 percent zero
warm_isch_tm_don	filled null with 0	N		remove due to >95 percent zero
wgt_kg_calc		N		keep
wgt_kg_don_calc		N		keep
wgt_kg_tcr		N		keep
work_income_tcr		Y	[nan, 1.0, 0.0]	keep
work_income_trr		Y	[0.0, nan, 1.0]	keep

Appendix B. Description of HCUP features for the GANN model

HCUP Category ID	HCUP Description
1	Incision and excision of CNS
3	Laminectomy; excision intervertebral disc
9	Other OR therapeutic nervous system procedures
10	Thyroidectomy; partial or complete
12	Other therapeutic endocrine procedures
33	Other OR therapeutic procedures on nose; mouth and pharynx

37	Diagnostic bronchoscopy and biopsy of bronchus
42	Other OR therapeutic procedures on respiratory system
43	Heart valve procedures
48	Insertion; revision; replacement; removal of cardiac pacemaker or cardioverter/defibrillator
61	Other OR procedures on vessels other than head and neck
67	Other therapeutic procedures; hemic and lymphatic system
70	Upper gastrointestinal endoscopy; biopsy
76	Colonoscopy and biopsy
78	Colorectal resection
80	Appendectomy
82	Endoscopic retrograde cannulation of pancreas (ERCP)
84	Cholecystectomy and common duct exploration
86	Other hernia repair
99	Other OR gastrointestinal therapeutic procedures
104	Nephrectomy; partial or complete
105	Kidney transplant
114	Open prostatectomy
124	Hysterectomy; abdominal and vaginal
126	Abortion (termination of pregnancy)
146	Treatment; fracture or dislocation of hip and femur
152	Arthroplasty knee
153	Hip replacement; total and partial
158	Spinal fusion
160	Other therapeutic procedures on muscles and tendons
161	Other OR therapeutic procedures on bone
172	Skin graft
225	Conversion of cardiac rhythm