Introductory Machine Learning Benchmarks for Neuro ICU Patients on the eICU Critical Care Dataset

Darwin Agunos

Advancements in using medical data have led to the development of several different scoring systems. Some scoring systems have different specific use cases (for example the Glasgow Coma Scale (GCS) while others are for generic towards all ICU patients (APACHE, APS, MPM, etc..). These generic scoring systems are used to assess disease severity and are routinely used to predict patient outcomes (length of stay estimation and mortality probability). The world's most widely used severity of illness scoring systems today are the APACHE IV/IVa scoring systems which were last developed around 2006-2008. This work examines the most recent APACHE scoring systems and benchmarks its performance on Neuro ICU Patients specific to the eICU Critical Care Dataset. We also construct simple linear regression and logistic regression models and compare their performances.

1. Introduction

With both the increased availability of medical data and advancements in machine learning, the medical industries capabilities to address a wider range of healthcare problems (e.g. clinical drug trial predictions, malignant cell detection, etc...). Critical care is a subspecialty of medicine that has benefitted from these advancements as critical care is an especially data-intensive field. The continuous monitoring of patients in Intensive Care Units generate large streams of data that can be used as inputs for machine learning algorithms. More specifically, increased statistical tooling and better data collection techniques have led to better estimates of patient risk assessment when entering an Intensive Care Unit (ICU).

ICUs provide care for severely ill patients who require invasive life-saving treatment. ICUs patients are monitored scrupulously to detect physiological changes associated with a deteriorating illness that would require reassessment of current treatment. Scrupulous observations are facilitated by beside monitors which outputs a large continuous stream of data, however small portions of this data are archived and even less data is utilized efficiently [1]. In most cases, commercial clinical databases have been designed to document clinical activity for reporting, liability, and billing reasons rather than the development of risk-assessment models. As such, those garnering an increasing excitement of using medical data for "Big Data" analytics should understand the limitations of current medical databases and the changes that need to occur to enter an era of "precision medicine" [2].

Advancements in digital health technologies, among other reasons, have led to the integration of well-networked critical care telemedicine (tele-ICU) systems across the United States. These systems provide the ability to generate large-scale remote monitoring data for critically ill patients. TeleICU's allow for large amounts of data to be collected and streamed for real-time monitoring for ICU teams.

Philips Healthcare, a major vendor of ICU equipment and services, provides a teleICU service known as the eICU program. Phillips Healthcare, in partnership with the MIT laboratory for Computational Physiology, constructed the eICU Collaborative Research Database [3]. The eICU Collaborative Research Database holds high granularity data for over 200,000 patient admissions to ICUs spanning 208 hospitals in the United States. The database is deidentified, and includes vital sign measurements, care plan documentation, severity of illness measures, diagnosis information, treatment information, hospital admission information and more.

The Acute Physiology, Age and Chronic Health Evaluation (APACHE) scoring system is a severity-of-disease classification system, one of the several available ICU scoring systems [4]. The APACHE score is a method for predicting hospital mortality among critically ill adults and length of stay (LOS) predictions. It is applied within 24 hours of a patient admission to an ICU.

Data	Median [IQR], Mean (std) or Number (%)
Age, years (median [IQR])	63.00 [50.00,75.00]
Unit length of stay, days (median [IQR])	1.86 [1.02,3.60]
Hospital length of stay, days (median [IQR])	5.13 [2.85,9.44]
Admission height, cm (mean (std))	169.56 (12.41)
Admission weight, kg (mean (std))	82.89 (25.96)
Hospital Region (n (%))	
Midwest	5227 (52.39)
South	3136 (31.43)
West	1614 (16.18)
Gender (n (%))	
Male	5157 (51.69)
Female	4816 (48.27)
Ethnicity (n (%))	
African American	1319 (13.22)
Asian	160 (1.60)
Caucasian	7521 (75.38)
Hispanic	332 (3.33)
Native American	64 (0.64)
Other/Unknown	581 (5.82)
Hospital discharge year (n (%))	
2014	4852 (48.63)
2015	5125 (51.37)
Status at unit discharge (n (%))	
Alive	9492 (95.14)
Expired	485 (4.86)
Status at hospital discharge (n (%))	
Alive	9138 (91.59)
Expired	839 (8.41)

Table 1. **Demographic of the 9,997 ICU unit admissions in the database.** Note that there is a 1:1 ratio for patient-to-unit admission. Since patients can have multiple unit admissions over time only the first record of admission for a patient was taken.

These predictions, aggregated across many patients, can be used to benchmark hospitals, and subsequently identify the policies that work in favor towards producing better patient outcomes. The APACHE IV scoring system require a set of parameters to make these predictions. These parameters include physiologic measurements, previous patient medical history/treatment history and admission diagnosis for the current stay. APACHE IVa is the most recent scoring system, developed in 2006 and has been described as having the highest discrimination of any other adult risk adjustment model (SAPS 3, SOFA, MPM III).

In this report we describe the use of the APACHE IV/APACHE IVa scoring system to create risk-assessment prediction models for patients in Neurological Intensive Care Units. This work will serve as a baseline for future projects within the domain.

2. Methodology

The data for this study is a subset of the 200,859 patient unit encounters collected from the eICU Collaborative Research Database. It is comprised of 9,997 unique patients admitted to the Neurological ICU from 2014 and 2015. Table 1 provides

APACHE Admission Diagnosis	Number (%)
Cerebrovascular Accident (CVA)	1846 (18.50)
Intracranial Hemorrhage/Hematoma	887 (8.89)
Neoplasm-Cranial, Surgery for	540 (5.41)
Head Only Trauma	463 (4.64)
Seizures (primary-no structural brain disease)	421 (4.22)
Subdural Hematoma	406 (4.07)
Neoplasm-Neurologic	272 (2.73)
Subarachnoid Hemorrhage/Intracranial Aneurysm	211 (2.11)
Laminectomy/Spinal Cord Decompressions (excluding malignancies)	201 (2.01)
Neurologic surgery, other	194 (1.94)

Table 2. **Most frequent categories of APACHE diagnosis using clinically meaningful groups defined in the code repository** [5].

demographics of this patient cohort while Table 2 highlights the top 10 most frequent admission diagnoses for this group as coded by trained eICU clinicians using the APACHE IV diagnosis system. To avoid counting more than one hospital outcome for any one patient our analysis only includes the first recorded patient admission to the Neurological ICU.

To predict mortality and length of stay (LOS) predictions we will be using the APACHE Score as our single feature. The APACHE Score requires a set of parameters to make these predictions. Creating mortality and LOS prediction models using these other features directly will be explored in a later paper. Exploratory plots detailing relationships between APACHE Score and patient outcomes and more are available in Appendix A.

For mortality and LOS predictions statistical models we used Logistic Regression and Linear Regression, respectively. To track the success of our models we employed three different metrics for our regression models (R^2, Mean Squared Error, Mean Absolute Error) and six different metrics for our classification model (Area Under Receiver Operating Curve (AUROC), Area Under Precision Recall Curve (AUPRC), Specificity, Sensitivity, Positive Predictive Value (PPV), Negative Predictive Value (NPV). Model performance was measured against the APACHE IV / APACHE IVa models' metric scores for each respective task (mortality/length of stay).

3. Experiments

All experimentation was performed in Python 3.7.6 (using various libraries) and is available in my Github [7].

3.1. Data Pipeline

For this study, the only feature we are using for our predictions is the APACHE Score. For missing values, we drop all rows. Missing data accounted for 0.008% of all data.

The remaining 99.992% of data (9892 instances) was split into training and testing sets with a test data size of 20%. This test data is not used in this study. For the training data we performed both Repeated K-Fold Cross Validation / Repeated Stratified K-Fold Cross Validation [6] using the former CV strategy on our regression task and latter CV strategy on our classification task. Let *n_splits* denote the number of folds and *n_repeats* denote the number of repeats. For our models we used *n_splits = 10* and *n_repeats = 100* totaling *1000* different data splits for each model. Metrics for each respective task were documented every split. For our results we report the average metric score as well as the standard deviation. The standard deviation score lets us know the stability of our model.

This study serves to create baseline measurements. All metadata (training/validation/test sets) are saved and will be used to create better prediction models in future papers

		Metrics for Model Validation Set		
Task	**Model**	***R^2***	**MSE (Day^2)**	**MAE (Day)**
ICU LOS Prediction Scores	**APACHE Version IV**	0.143	15.716	2.358
	APACHE Version IVa	0.111	16.295	2.577
	Simple Linear Regression	0.045 ± 0.017	17.470 ± 2.919	2.520 ± 0.105
Hospital LOS Prediction Scores	**APACHE Version IV**	0.075	82.598	5.383
	APACHE Version IVa	0.042	85.592	5.704
	Simple Linear Regression	0.028 ± 0.016	120.644 ± 59.847	5.594 ± 0.309

Table 3. **Length of Stay model prediction metric scores**. Regression models are matched up against APACHE IV / APACHE IVa predictions. Regression models were cross-validated *1000* times. We report the mean metric score and the standard deviation over all *1000* iterations. *Note that standard deviation scores are not available for the APACHE prediction models.

		Metrics for Model Validation Set					
Task	**Model**	**AUROC**	**AUPRC**	**Sens.**	**Spec.**	**PPV**	**NPV**
ICU Mortality Prediction Scores	**APACHE Version IV**	0.904	0.414	0.358	0.982	0.502	0.968
	APACHE Version IVa	0.907	0.427	0.313	0.987	0.553	0.966
	Simple Logistic Regression	0.883 ± 0.028	0.385 ± 0.070	0.165 ± 0.056	0.992 ± 0.003	0.515 ± 0.130	0.960 ± 0.003
Hospital Mortality Prediction Scores	**APACHE Version IV**	0.878	0.473	0.396	0.968	0.532	0.946
	APACHE Version IVa	0.879	0.472	0.354	0.975	0.56	0.943
	Simple Logistic Regression	0.855 ± 0.024	0.434 ± 0.053	0.205 ± 0.046	0.988 ± 0.004	0.618 ± 0.096	0.933 ± 0.004

Table 4. **Mortality model prediction metric scores**. Classification models are matched up against APACHE IV / APACHE IVa predictions. Classification models were cross-validated *1000* times. We report the mean metric score and the standard deviation over all *1000* iterations. *Note that standard deviation scores are not available for the APACHE prediction models.

4. Results & Discussion

In this section we discuss our results and comparisons to the APACHE IV / APACHE IVa predictions. The APACHE IV / APACHE IVa models used logistic regression for mortality predictions [3]. It is unknown what was used for LOS predictions.

In our models, we establish baseline data using the simplest machine learning models. For LOS predictions we developed a linear regression model. For mortality predictions we developed a logistic regression.

4.1. Length of Stay Predictions

Length of stay model prediction metric scores are shown in Table 3. Note that APACHE IV performs better in general than APACHE IVa for ICU LOS prediction and Hospital LOS Prediction. Our simple linear regression model performs notably worse all metrics board except for MAE. We note that our model has less mean error for MAE than APACHE IVa for both ICU LOS and Hospital LOS predictions but could still perform worse given the standard deviation of our model. Looking at our simple linear regression's standard deviation for each metric we can conclude that there is little variability in most cases.

Our results conclude that a simple linear regression LOS prediction model using the APACHE score yields worse notably worse performance than the APACHE IV / APACHE IVa prediction models.

4.2. Mortality Predictions

Mortality model prediction metric scores are shown in Table 4. This is an imbalanced classification task (which can be seen in Appendix A Figure 1) where patients who died are the minority by a large margin (statement applicable to both ICU and hospital mortality).

Note that APACHE IV performs better than APACHE IVa perform better in different areas and vice versa. More specifically, in the case of ICU mortality predictions, we expect the APACHE IV prediction model to identify more patients who will die (higher sensitivity scores). However, this comes with the tradeoff of more false positives (lower specificity scores).

Given our baseline we see that our models perform worse across all metrics except for specificity for both ICU and hospital mortality predictions (the proportion of true negatives that are correctly identified as such). This is also to be expected as there is a huge overlap in APACHE score for patients who survived and died (Appendix A Figure 3).

Our results conclude that a simple logistic regression mortality prediction model using the APACHE score yields worse notably worse performance than the APACHE IV / APACHE IVa prediction models.

5. Conclusions, Limitations and Future Work

In this work we constructed LOS and mortality prediction models for patients in the Neuro ICU and compared them to the APACHE IV / APACHE IVa prediction models. We conclude that constructing models based on the single feature APACHE Score performs notably worse for both prediction tasks.

For future work we will create prediction models that instead use the features that predicted the APACHE Score and levy different machine learning algorithms.

References

1. Celi, L. A., Mark, R. G., Stone, D. J. & Montgomery, R. A. "Big Data" in the Intensive Care Unit: Closing the Data Loop. Am J Respir Crit Care Med 187, 1157–1160 (2013).
2. Johnson, A. E., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D. A., & Clifford, G. D. (2016). Machine Learning and Decision Support in Critical Care. Proceedings of the IEEE. Institute of Electrical and Electronics Engineers, 104(2), 444–466. https://doi.org/10.1109/JPROC.2015.2501978
3. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG and Badawi O. Scientific Data (2018). DOI: http://dx.doi.org/10.1038/sdata.2018.178.
4. Zimmerman, Jack & Kramer, Andrew & Mcnair, Doug & Malila, Fern. (2006). Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today???s critically ill patients. Critical care medicine. 34. 1297-310. 10.1097/01.CCM.0000215112.84523.F0.
5. Pollard, T. J. et al. MIT-LCP/eicu-code: eICU-CRD Code Repository. Zenodo https://doi.org/10.5281/zenodo.1249016 (2018).
6. Berrar, Daniel. (2018). Cross-Validation. 10.1016/B978-0-12-809633-8.20349-X.
7. https://github.com/darwin-a/Programming_Projects/tree/master/Python/Data%20Science/eICU/Project%201

Appendix A: Exploratory Plots

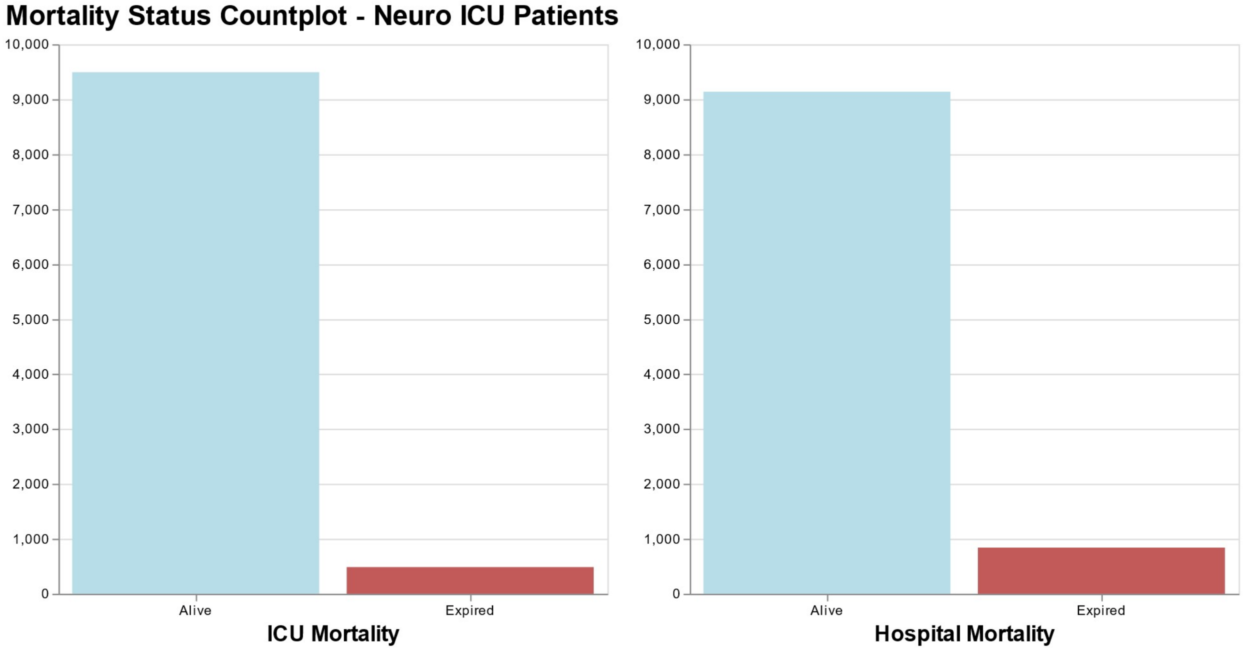

Figure 1. **Mortality Status of the Neuro ICU Patients by Admission**. Left graph indicates mortality status coming out of the ICU. Right graph indicates mortality status at the end of patient admission stay.

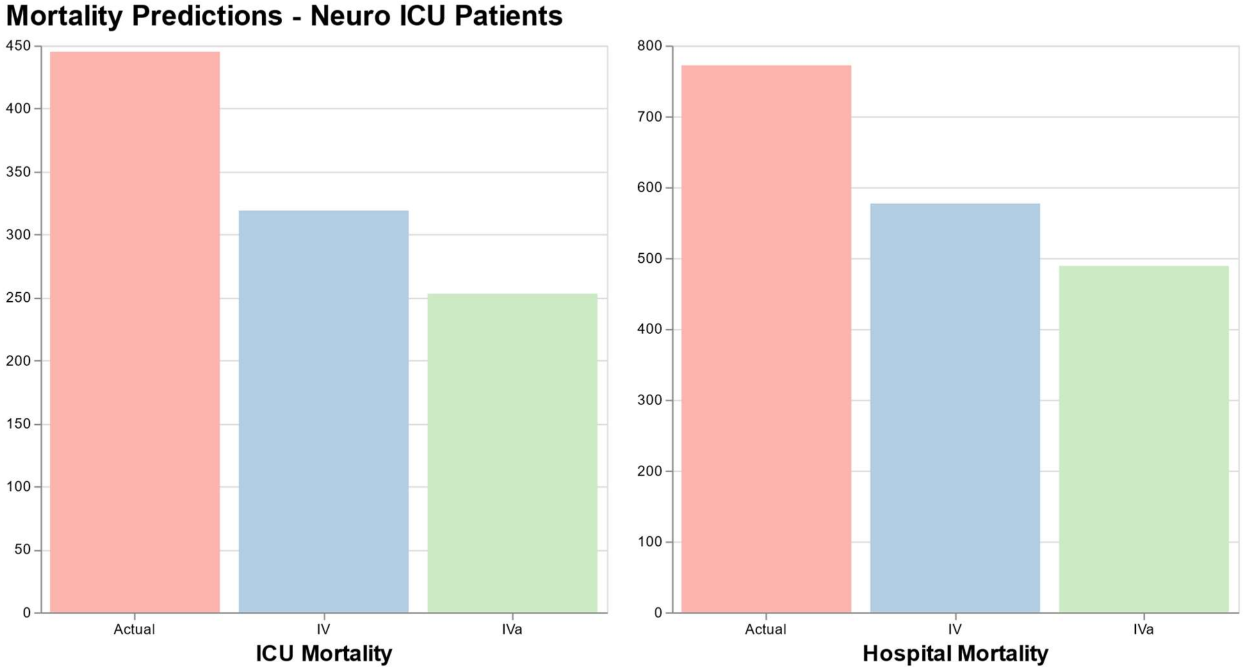

Figure 2. **Mortality Predictions of the Neuro ICU Patients**. Left graph compares APACHE IV/APACHE IVa ICU mortality predictions to actual recorded ICU mortality predictions. Right graph compares hospital mortality predictions.

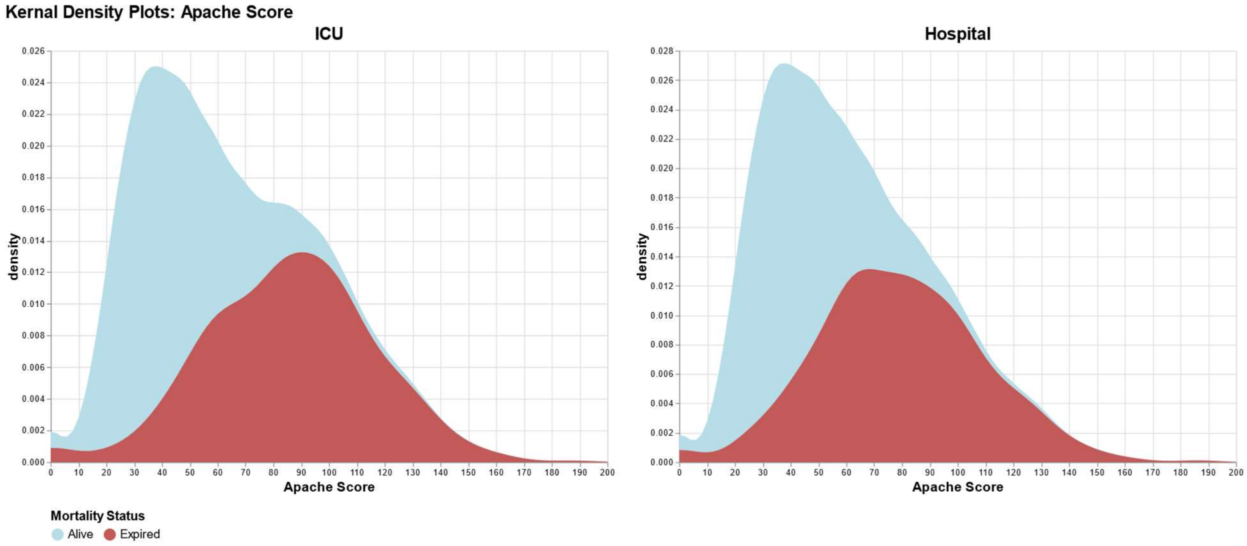

Figure 3. APACHE Score Distribution by Mortality Status. Left graph indicates APACHE score distribution by mortality status coming out of the ICU. Right graph indicates mortality status at the end of patient admission stay.

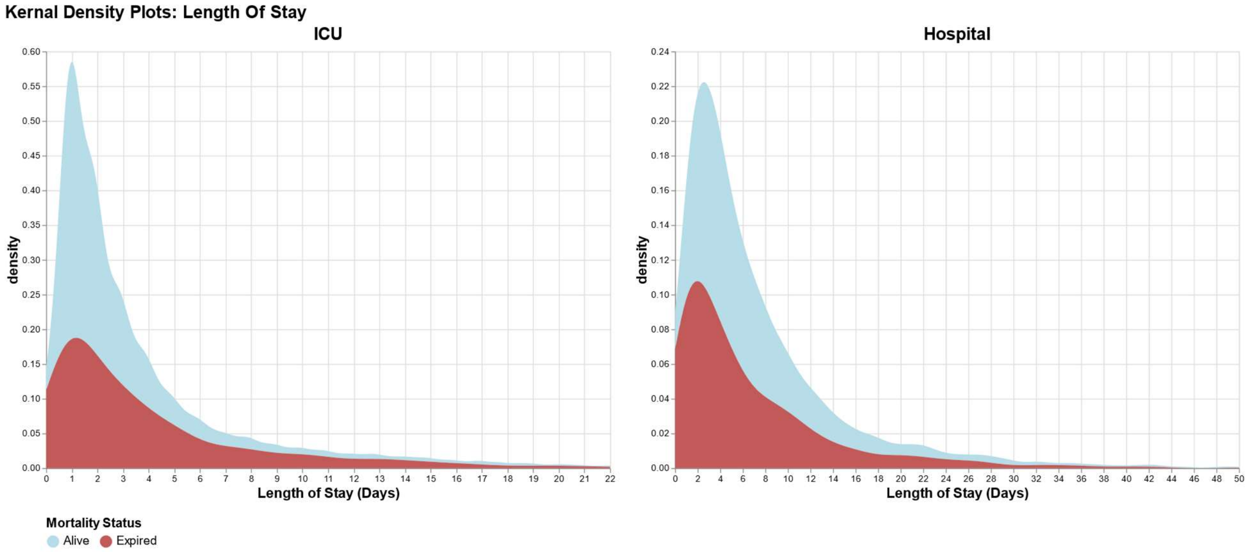

Figure 4. Length of Stay (LOS) Distribution by Mortality Status. Left graph indicates LOS distribution by mortality status coming out of the ICU. Right graph indicates LOS status at the end of patient admission stay.