

Predicting the Effects of Sedative Infusion on Acute Traumatic Brain Injury Patients

Jeffrey Reynolds McCullen

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Computer Science and Applications

Chandan Reddy
Vignesh Subbian
Anuj Karpatne

Date of Defense: April 1st, 2020
Blacksburg, Virginia

Keywords: Data Mining, Data Analytics, eICU Database, Healthcare, Health Analytics,
Machine Learning, ICU, Deep Learning, Neural Networks, Linear Regression, Traumatic
Brain Injury, Classification, Regression, Sedatives

Predicting the Effects of Sedative Infusion on Acute Traumatic Brain Injury Patients

Jeffrey Reynolds McCullen

ABSTRACT

Healthcare analytics has traditionally relied upon linear and logistic regression models to address clinical research questions mostly because they produce highly interpretable results [1, 2]. These results contain valuable statistics such as p-values, coefficients, and odds ratios that provide healthcare professionals with knowledge about the significance of each covariate and exposure for predicting the outcome of interest [1]. Thus, they are often favored over new deep learning models that are generally more accurate but less interpretable and scalable. However, the statistical power of linear and logistic regression is contingent upon satisfying modeling assumptions, which usually requires altering or transforming the data, thereby hindering interpretability. Thus, generalized additive models are useful for overcoming this limitation while still preserving interpretability and accuracy.

The major research question in this work involves investigating whether particular sedative agents (fentanyl, propofol, versed, ativan, and precedex) are associated with different discharge dispositions for patients with acute traumatic brain injury (TBI). To address this, we compare the effectiveness of various models (traditional linear regression (LR), generalized additive models (GAMs), and deep learning) in providing guidance for sedative choice. We evaluated the performance of each model using metrics for accuracy, interpretability, scalability, and generalizability. Our results show that the new deep learning models were the most accurate while the traditional LR and GAM models

maintained better interpretability and scalability. The GAMs provided enhanced interpretability through pairwise interaction heat maps and generalized well to other domains and class distributions since they do not require satisfying the modeling assumptions used in LR. By evaluating the model results, we found that versed was associated with better discharge dispositions while ativan was associated with worse discharge dispositions. We also identified other significant covariates including age, the Northeast region, the Acute Physiology and Chronic Health Evaluation (APACHE) score, Glasgow Coma Scale (GCS), and ethanol level. The versatility of versed may account for its association with better discharge dispositions while ativan may have negative effects when used to facilitate intubation. Additionally, most of the significant covariates pertain to the clinical state of the patient (APACHE, GCS, etc.) whereas most non-significant covariates were demographic (gender, ethnicity, etc.). Though we found that deep learning slightly improved over LR and generalized additive models after fine-tuning the hyperparameters, the deep learning results were less interpretable and therefore not ideal for making the aforementioned clinical insights. However deep learning may be preferable in cases with greater complexity and more data, particularly in situations where interpretability is not as critical. Further research is necessary to validate our findings, investigate alternative modeling approaches, and examine other outcomes and exposures of interest.

Predicting the Effects of Sedative Infusion on Acute Traumatic Brain Injury Patients

Jeffrey Reynolds McCullen

GENERAL AUDIENCE ABSTRACT

Patients with Traumatic Brain Injury (TBI) often require sedative agents to facilitate intubation and prevent further brain injury by reducing anxiety and decreasing level of consciousness. It is important for clinicians to choose the sedative that is most conducive to optimizing patient outcomes. Hence, the purpose of our research is to provide guidance to aid this decision. Additionally, we compare different modeling approaches to provide insights into their relative strengths and weaknesses.

To achieve this goal, we investigated whether the exposure of particular sedatives (fentanyl, propofol, versed, ativan, and precedex) was associated with different hospital discharge locations for patients with TBI. From best to worst, these discharge locations are home, rehabilitation, nursing home, remains hospitalized, and death. Our results show that versed was associated with better discharge locations and ativan was associated with worse discharge locations. The fact that versed is often used for alternative purposes may account for its association with better discharge locations. Further research is necessary to further investigate this and the possible negative effects of using ativan to facilitate intubation. We also found that other variables that influence discharge disposition are age, the Northeast region, and other variables pertaining to the clinical state of the patient (severity of illness metrics, etc.). By comparing the different modeling approaches, we found that the new deep learning methods were difficult to interpret but provided a slight improvement in performance after optimization. Traditional methods such as linear

regression allowed us to interpret the model output and make the aforementioned clinical insights. However, generalized additive models (GAMs) are often more practical because they can better accommodate other class distributions and domains.

Dedication

I dedicate my thesis work to my family and the professors who helped me prepare for graduate studies. I am grateful to my family for providing support and my professors at Lynchburg University for equipping me with the tools and knowledge to succeed in graduate school.

Acknowledgements

I would like to thank Dr. Chandan Reddy for advising me throughout my graduate studies and providing me the opportunity to engage in healthcare research. I would also like to thank Dr. Vignesh Subbian for providing additional guidance and expertise. Finally, I would like to thank Dr. Brandon Foreman and Dr. Johnathan Ratcliff for providing domain expertise, helping to formulate a practical research question, and taking the time to regularly meet and provide constructive feedback.

Table of Contents

Table of Contents	vi
List of Figures	viii
List of Tables	ix
1 Introduction	1
1.1 Electronic Health Records	1
1.2 Problem Statement	2
1.2.1 Evidence for Choice of Sedative Agents	3
1.3 Problem Motivation	4
1.3.1 Research Question	4
1.3.2 Modeling Approaches	5
2 Methods	8
2.1 Relevant Literature	10
2.1.1 Secondary Analysis	11
2.1.2 Sedation in Traumatic Brain Injury	11
2.1.3 Acute Physiology and Chronic Health Evaluation	12
2.2 Data Description	12
2.2.1 The eICU Program	13
2.2.2 Patient Representation	13
2.3 Data Pre-Processing	15
2.3.1 Data Cleaning	15
2.3.2 Data Integration	16
2.3.3 Data Transformation	17
2.4 Data Exploration	18
2.4.1 Descriptive Summary Statistics	18
2.4.2 Graphical Methods	20
2.5 Model Building	23
2.5.1 Linear Regression	24
2.5.2 Deep Learning	26

2.5.3	Regularized Linear Regression	29
2.5.4	Generalized Additive Models	30
2.6	Model Evaluation	31
2.6.1	Accuracy	32
2.6.2	Interpretability.....	34
2.6.3	Scalability	35
2.6.4	Generalizability.....	35
3	Real World Insights	38
3.1	Clinical Implications	38
3.1.1	Effects of Sedatives.....	38
3.1.2	Effects of Demographic and Clinical Covariates.....	40
3.1.3	Sankey Diagram Interpretation	41
3.1.4	Pairwise Interaction using Heat Map.....	42
3.2	Data Science Insights	44
4	Conclusion	46
4.1	Lessons Learned.....	46
4.1.1	Common Problems in Modeling for Healthcare.....	46
4.1.2	Strengths and Weaknesses of Modeling Approaches	48
4.2	Summary	49
4.3	Future Research.....	50
4.3.1	Investigate the Effects of Versed and Fentanyl/Ativan	50
4.3.2	Examine Additional Outcomes and Exposures.....	51
4.3.3	Test Other Modeling Approaches.....	51
	References.....	52
	Supplementary Materials	54
	Feature Description Table.....	54
	Summary Statistics Table	56
	Coefficient Plots.....	58
	GA ² M Results	61

List of Figures

Figure 2.1: The three level hierarchy of patient data used in the eICU-CRD as portrayed in [12]. Each patient has one or more hospital visits and each hospital visit has one or more ICU stays.	14
Figure 2.2: Class distribution for hospital discharge location.	21
Figure 2.3: Boxplot (left) and histogram (right) showing the distribution of hospital ventilation days. The outliers in the boxplot were determined using the 1.5 times IQR rule. Thus, patients with over five ventilation days are outliers.	22
Figure 2.4: Boxplot (left) and histogram (right) showing the distribution of hospital length of stay (days). The outliers in the boxplot were determined using the 1.5 times IQR rule. Thus, patients with a hospital length of stay greater than 20.05 days are outliers.	22
Figure 2.5: Bar chart showing the mortality distribution of patients with TBI.	22
Figure 3.1: Sankey diagrams showing the flow of patients from each sedative group to the different hospital discharge locations. The monotherapy diagram is on the left, the dual therapy diagram is in the middle, and the triple therapy diagram is on the right.	42
Figure 3.2: Heat maps showing the pairwise interactions between the APACHE and GCS scores.	43
Figure 3.3: Heat maps showing the pairwise interactions between propofol and the APACHE and GCS scores.	44

List of Tables

Table 2.1: Mapping of TBI ICD-9-CM codes to a detailed description [11].	9
Table 2.2: Demographic diversity of our training set from the eICU-CRD. For age, we report the mean, sample standard deviation (SD), minimum, maximum, median, and interquartile range (IQR). For the other demographic variables, we report the number of patients and the percent of the total population they represent. The data set is regionally diverse but is not diverse in terms of age, gender, or ethnicity. Most of the patients are white males over 60 years old.....	20
Table 2.3: Accuracy and scalability of the various models for the monotherapy experiment.	32
Table 2.4: Accuracy and scalability of the various models for the dual therapy experiment.	32
Table 2.5: Accuracy and scalability of the various models for the triple therapy experiment.	32
Table 3.1: Final model summary for predicting hospital discharge location using monotherapy.....	39
Table 3.2: Final model summary for predicting hospital discharge location using dual therapy.....	40
Table 3.3: Final model summary for predicting hospital discharge location using triple therapy.....	40

1 Introduction

The uncertainties in healthcare due to lack of effective analytics hinders clinical decision making and practice guidelines, ultimately resulting in worse outcomes for patients and caregivers. In response, healthcare analytics is emerging as a valuable specialization by promising improved clinical care through research made possible by technological advancements [1]. Previously, clinicians used pen and paper to record events and information about patients such as lab results, drug infusions, etc. However, this method of recordkeeping makes it difficult to leverage the vast amounts of valuable information they contain to make new discoveries. Indeed, it would be very time consuming, error prone, and inefficient to effectively sift through and analyze file systems full of patient clinical information. Moreover, this manual analysis would likely be limited in scope and a breach of patient confidentiality.

1.1 Electronic Health Records

The development of electronic health records (EHRs) has largely replaced paper-based systems, which is particularly good for improving recordkeeping. In these systems, record-keepers enter patient information in a computer, making it easier to archive, organize, and query. These records can be de-identified so as to not violate patient confidentiality. Importantly, EHRs also make it possible to mine large amounts of clinical data, which, if properly interpreted and analyzed, leads to more informed

clinical practice guidelines (CPGs) and clinical decision making [1]. Thus, it is necessary for clinical experts, scientists, and software engineers to form multidisciplinary teams to make such contributions.

Unfortunately, EHRs also have problems and challenges that research teams must overcome. They are subject to human mistakes such as data entry errors and incomplete or inaccurate records. Additionally, practices, architectures, and standards differ across hospitals and other health institutions, which makes data interoperability and integration challenging. This can lead to numerous problems, namely biased or inaccurate experimental results. Such bias results from basing results on overly narrow patient cohorts that do not generalize to the greater population whereas inaccuracies result from human mistakes during the data entry or research process. Nevertheless, dedicated teams can overcome these challenges and make clinically relevant insights through a robust data analytics pipeline that includes steps for data exploration and pre-processing.

This process of mining routinely acquired EHR data is termed “secondary analysis” since it deals with patients generally rather than particular individuals in a limited or local setting [1]. In our specific case, we will use EHRs to provide guidelines for sedative choice for patients with traumatic brain injury.

1.2 Problem Statement

The specific problem we address through this research is our lack of understanding of the effect of various sedatives on the injured brain. The purpose of infusing sedatives into patients with TBI is to prevent secondary brain injury, which is further brain damage subsequent to the initial brain injury (the primary injury) but not necessarily caused by it. The negative effects of primary brain injury include

increased intracranial pressure (ICP), decreased cerebral perfusion pressure (CPP), axonal shear injury, intracranial hematoma, and cerebral ischemia [3]. Management of secondary insults such as systemic hypotension, hypoxia (inadequate oxygen in your tissues), hypercapnia (CO₂ retention), and intracranial hypertension is essential for optimizing patient outcomes. Thus, the purpose of sedative infusion is to optimize ventilation to prevent hypoxia and decrease ICP and the cerebral metabolic rate (CMRO₂), which reduces cerebral blood flow (CBF) and cerebral blood volume (CBV). Their administration is useful in cases involving alcohol or drug intoxication, seizure control, and mechanical ventilation (MV) and intubation due to their anti-anxiety and therapeutic effects and ability to decrease level of consciousness. Unfortunately, this often comes at the cost of decreased systemic blood pressure and other negative effects that are predictors for poor outcomes [3].

1.2.1 Evidence for Choice of Sedative Agents

There are numerous sedative agents but, unfortunately, there is limited evidence to determine the most appropriate choice in specific cases. Clinicians and researchers sometimes conduct studies on their primary or local patients that are limited in regional and demographic diversity, resulting in biased results and poor generalizability. Furthermore, multiple sedative agent infusions confound many of the studies, making it difficult to analyze the effects of each individually [3]. The Brain Trauma Foundation recognizes this lack of understanding in their guidelines for sedative infusion, noting that the quality of the body of evidence for current recommendations is low and susceptible to change if a larger study reaches alternative conclusions [3, 4]. Thus, clinical experts often base their decisions on their opinion and experience rather than experimental results and evidence obtained from analysis of electronic health records. Clearly, more research is necessary in order to generate more evidence on the advantages and disadvantages of each sedative agent and the appropriate context for their infusion.

1.3 Problem Motivation

Traumatic brain injury is a major public health concern in the United States. According to the 2014 Center for Disease Control (CDC) TBI Surveillance report, there were 2.87 million cases of TBI emergency department visits, hospitalizations, and deaths in 2014 [5]. They were caused primarily by falls (especially in older people), being hit by or against an object, and motor vehicle crashes. Survivors often suffer from short-term and long-term disabilities including decreased memory capacity, mobility, sensation, and emotional functioning [5]. Clinicians and researchers must mitigate these effects through procedures and guidelines such as those for sedative choice.

Thus, the problem motivation for this research pertains to both clinicians and researchers in data science. In short, it is to use electronic health records to build scalable, generalizable, accurate, and interpretable models to:

1. Understand the effect of various sedatives on the injured brain.
2. Generate evidence on the appropriate choice of sedative agents that optimizes patient outcomes.
3. Understand the strengths and weaknesses of traditional LR, GAM, and deep learning approaches.

1.3.1 Research Question

To address this problem, we formulated a research question to identify the patient cohort, exposures of interest, outcome of interest, and other covariates. In short, the specific research question we address is as follows: *in the population of patients with TBI, is the exposure of particular sedatives associated with a different hospital discharge location than in the control group when controlling for severity of illness,*

alcohol abuse, and intubation? We narrowed the patient cohort to patients in the intensive care unit (ICU) with TBI. The exposures of interest are various sedatives, namely fentanyl, propofol, versed, ativan, and precedex. The outcome of interest is the hospital discharge location, which we ranked from worst to best. These locations are death, remains hospitalized, nursing home, rehabilitation, and home respectively. The other covariates are demographic characteristics and clinical variables. The demographic covariates include age, gender, ethnicity, and region. Clinical covariates include severity of illness and level of consciousness metrics (Glasgow Coma Scale (GCS), APACHE, etc.), ethanol levels, and intubation status. Including these covariates allows us to control for severity of illness and ethanol intoxication (alcohol abuse).

1.3.2 Modeling Approaches

Healthcare analytics has traditionally relied upon linear and logistic regression models to produce interpretable results to address these sort of research questions [1, 2]. For instance, the results of a stepwise multiple linear regression model provides valuable statistics including measures for statistical significance, practical significance, uncertainty and precision, error rate, and overall model fit. Similarly, logistic regression produces valuable statistics such as the odds ratio [1].

While the success of new and innovative deep learning techniques has led to its adoption in various domains, the use of these traditional approaches is sometimes still preferred in healthcare analytics due to their ability to produce the aforementioned highly interpretable results [1]. Thus, deep learning is often not the ideal solution for many clinical questions.

Deep learning is a machine learning technique that involves creating a neural network consisting of multiple layers that learn abstract features from the data through a process called backpropagation to minimize the loss. This involves feeding samples forward through the network, making a prediction, backpropagating the error, and adjusting weights to minimize the error by learning abstract features for classification and regression. Though the network requires sufficient time and data to train, the resulting model is usually more accurate than traditional models due to their superior ability to learn non-linear relationships. Unfortunately, since the network learns these features through backpropagation using the partial derivative of the loss function with respect to the weights, domain experts such as healthcare professionals often find them difficult to interpret. They do not produce the valuable statistics found in the traditional models and are described as “black box” models because the meaning of the internal calculations leading to the prediction is difficult to decipher, though there have been numerous attempts to make them more transparent [6, 7].

There are also variations on traditional linear regression methods in our analysis such as regularized forms of linear regression and generalized additive models (GAMs). Regularization in machine learning models is used to improve performance by penalizing model complexity to improve generalizability, thereby improving its performance on the test set [8]. Excessive model complexity results in overfitting to the training data, resulting in worse performance on the test data set due to the model’s lack of generalizability. Generalized additive models use smoother functions to provide greater generalizability and learn non-linear relationships [9]. Researchers have enhanced these models to allow for pairwise interactions between the variables, which allows for further insights into the significance of interacting features [10].

Data science researchers are interested in understanding the relative strengths and weaknesses of these different modeling approaches. Hence, we are motivated to identify relative strengths and weaknesses of each modeling approach in the context of our research question. We provide a more detailed overview of each modeling approach and how we used it for our research question in the model building section.

We then compare traditional LR, generalized additive, and deep learning models to address our healthcare research question, which is to investigate the effects of sedative infusion on patients with acute TBI.

2 Methods

We formed a team of software and biomedical engineers, researchers, and clinical experts to formalize and investigate our research question using a multistep data analytics pipeline. This pipeline included steps for reviewing relevant literature, data description, data pre-processing, exploratory data analytics (EDA), model building, and model evaluation. The software engineer role primarily involved processing the data, building and evaluating the models, and presenting the results. The data science researchers and clinical experts were primarily responsible for providing guidance, analyzing the results, and providing insights into the different modeling approaches and the clinical significance of the results. However, these roles overlapped and the data analytics pipeline required frequent revisiting of previous steps as we continued to learn more about the data and ask further questions. We began by formalizing the research question to define the patient cohort, exposures of interest, and outcomes of interest.

Since our goal was to understand the relationship between sedatives and the injured brain, we narrowed the patient cohort to only patients with acute traumatic brain injury. We identified patient diagnoses using the International Classification of Diseases, Ninth Edition, Clinical Modification (ICD-9-CM) codes and classified a TBI diagnosis using the established Center for Disease Control and Prevention (CDC) ICD-9-CM TBI surveillance codes [11]. See Table 2.1 for a list of these codes with descriptions. We established exclusion criteria to prevent confounding and bias. This initially only included chronic

alcohol use (alcoholism) but exploratory data analytics helped us determine other possible confounders such as severity of illness.

TBI ICD-9-CM Codes	Description
800.0 – 801.9	Fracture of the vault or base of the skull
803.0 – 804.9	Other and unqualified and multiple fractures of the skull
850.1 – 850.5, 850.9, 851.0 – 854.1	Intracranial injury, including concussion, contusion, laceration, and hemorrhage
959.01	Head injury, unspecified

Table 2.1: Mapping of TBI ICD-9-CM codes to a detailed description [11].

The exposures of interest are the various sedative agents chosen due to their prevalence in treating TBI based on the clinical experience and expertise of the clinicians on our team. Specifically, we were interested in eight particular sedative agents: fentanyl, propofol, versed, ketamine, ativan, pentobarbital, precedex, and alprazolam. However, we excluded ketamine, pentobarbital, and alprazolam due to insufficient data. See [3] for details on some of these sedatives.

We identified several outcomes of interest for a patient with TBI. The outcome of interest we focused on for this project was the discharge location; however, we considered hospital mortality, medical ventilation days, and length of stay for future research. We grouped the discharge locations into five categories. From worst to best, these categories are death, remains critically ill (hospitalized on the floor or in the ICU), nursing home, rehabilitation, and home. Exploratory data analytics was useful for identifying confounding variables so we could control for them, which is especially important for predicting mortality and length of stay because they are particularly susceptible to confounding [1]. Such confounders include chronic alcohol use (alcoholism), ethanol serum level, and severity of illness.

Our data analytics pipeline began with reviewing the relevant literature to inform the subsequent steps. Subsequently, we chose to use the Philips Hospital to Home electronic ICU (eICU) program database to obtain our dataset [12, 13]. The rationale for this choice is that it provides greater patient diversity (patients came from across the United States), more structured data (no free text entries), and more TBI patients than other possible choices such as Medical Information Mart for Intensive Care III (MIMIC-III) [1, 14, 15]. Thus, it provided us with a regionally diverse patient cohort with sufficient TBI patients so we could build fair models that do not have a bias towards a particular region. During data pre-processing, we cleaned, transformed, and integrated the data so that the resulting data set had TBI patients with variables indicating their sedative use, demographics, clinical state, and other covariates of interest. We built models for numerous experiments using interpretable methods (linear regression and generalized additive models) so we could determine the influence of each of the variables on the outcome of interest (discharge location) and compared them with new deep learning approaches. Finally, we validated the models using five-fold cross validation. We evaluated the models using metrics for accuracy (root mean squared error), scalability (training and testing time), interpretability, and generalizability. Subsequent sections provide more detailed descriptions of each step.

2.1 Relevant Literature

The relevant literature includes studies relating to healthcare data analytics and the use of sedatives for treating patients with TBI.

2.1.1 Secondary Analysis

The guide for Secondary Analysis of Electronic Health Records developed by the Massachusetts Institute of Technology (MIT) Critical Data [1], an international group of researchers and clinicians who support research to improve the healthcare system, guided our research process. The guide informed our understanding of the current landscape of the healthcare system in terms of clinical guidelines, challenges, and limitations. The guide also outlined the process of formalizing a healthcare research question and investigating it using the data analytics pipeline. Specific case studies using real electronic health care data provided concrete examples of the best practices to conduct effective research.

2.1.2 Sedation in Traumatic Brain Injury

Prior studies related to our particular research question of understanding the relationship between sedatives and the injured brain provided a useful overview of various sedatives and their advantages and disadvantages to provide guidance on when to use them [3]. Ketamine or midazolam (versed) plus fentanyl are recommended as induction agents in hemodynamically unstable patients (patients with an unstable blood flow) patients. For hemodynamically stable patients (patients with a stable blood flow), thiopentone or propofol are recommended, both with the addition of fentanyl. As a maintenance agent, propofol and fentanyl are recommended. The advantage of propofol is that it has a rapid onset of sedation; however, midazolam is suitable in cases where propofol would require high dosages, in hypotensive patients, and for prolonged sedation. Shorter acting agents such as remifentanyl and dexmedetomidine (precedex) have the advantages of faster recovery of consciousness and neurological assessment. However, the dosage might be more important than the agent choice [3].

2.1.3 Acute Physiology and Chronic Health Evaluation

Since effectively addressing our research question requires controlling for severity of illness, we reviewed the literature relevant to severity of illness scores, namely the fourth version of the Acute Physiology and Chronic Health Evaluation (APACHE-IVa) [16]. This APACHE score improved upon previous versions of the scoring system which began in 1981 [17, 18]. It is calculated to predict patient outcomes (MV days, mortality, and length of stay) using the Acute Physiology Score (APS), which consists of 12 physiologic variables, age, and chronic health status in 56 disease groups. These include variables such as the Glasgow Coma Scale (GCS), mean blood pressure, intubated status, ventilation status, etc. To perform the calculation, the worst measurement in terms of psychological derangement during the “APACHE day”, which is the first ICU day, is used. Research has shown that this score has good discrimination and calibration [16].

2.2 Data Description

We used the Philips Electronic Intensive Care Unit Collaborative Research Database (eICU-CRD) to obtain our data set [12]. While we considered other databases such as the Medical Information Mart for Intensive Care III (MIMIC-III) database, we favored the eICU-CRD because of its structured clinical data, greater regional diversity, and greater quantity of patients with TBI. Unlike MIMIC-III, the eICU-CRD has data from ICUs across the United States [12], making its regional data diverse and better for evaluating model fairness. Additionally, its structured clinical data make it easier to extract relevant features and its larger quantity of TBI patients makes it better for training the models.

2.2.1 The eICU Program

The eICU Research Institute (eRI) created the eICU-CRD using data from the eICU program, a telehealth ICU (teleICU) program maintained by a major vendor of ICU equipment and services called Philips Health. They use this program to supplement the bedside team with bedside monitors that detect physiologic changes and deteriorating illness for patients in ICUs across the United States to reduce mortality, length of stay, and cost of care. Clinicians and caregivers can access this data through the eCareManager information management system, allowing them to track care plans and promptly respond to sudden deterioration with the appropriate treatment regime [12].

The eICU-CRD is a subset of the data maintained by the eRI from the eICU program between the years 2014 and 2015. They stratified the subset by hospital so the number of patients from each hospital is proportional to that of the full data set. The eRI formed the eICU-CRD in partnership with MIT's Laboratory for Computational Phycology, which previously released the MIMIC-III database, and released it under the Health Insurance Portability and Accountability Act (HIPPA) safe harbor provision to de-identify the data. This process involved removing ICU and hospital identifiers and protected health information (PHI) such as ages over 89, social security numbers, phone numbers, addresses, etc. By doing so, they safely make the data public so researchers can use it for analysis to improve clinical decision making and patient monitoring [12].

2.2.2 Patient Representation

The database contains 31 comma separated value (CSV) files that researchers can upload to a relational database, which we did using PostgreSQL relational database [19]. These files represent tables (lab

results, clinical notes, patients, etc.) linked together using identifiers at each level of the three level hierarchy of patient tracking information. The first level is the patient level, which consists of a unique patient identifier for each patient. The second level is the hospital level, which consists of a unique identifier for each hospital visit (patient health system stay). A single patient can have many hospital visits. The third level is the unit level, which consists of a unique identifier for each ICU stay. A single hospital visit can have many ICU stays. In short, the eICU-CRD represents a patient as a sequence of hospital visits that consist of a sequence of ICU stays. See Figure 2.1 for a visual representation of this hierarchy [12].

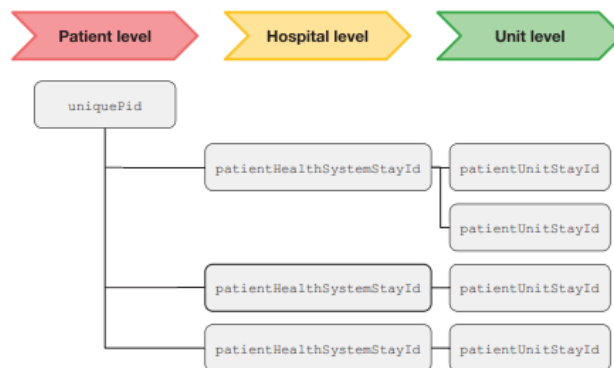


Figure 2.1: The three level hierarchy of patient data used in the eICU-CRD as portrayed in [12]. Each patient has one or more hospital visits and each hospital visit has one or more ICU stays.

Thus, our data set from the eICU-CRD consists of regionally diverse patients from ICUs across the United States from 2014 to 2015. It contains tables useful for extracting relevant features such as demographics, medical history, and diagnoses. These tables are linked together using a three level hierarchy of patient data [12].

2.3 Data Pre-Processing

Data pre-processing was necessary to obtain our desired data structure and samples from the eICU-CRD. It involved data cleaning, data integration, and data transformation. Data cleaning involved removing noisy or redundant data and accounting for data entry errors. Data integration involved linking together various tables and sources of data. Data transformation involved engineering new features and translating existing features into a format suitable for analysis [1].

Our desired data structure was a table in which the rows represent unique patients with TBI and the columns represent the relevant features, namely the exposures of interest, covariates, and outcomes of interest. The exposures of interest are the various sedatives: fentanyl, propofol, versed, ketamine, ativan, pentobarbital, precedex, and alprozolam. However, we later excluded ketamine, pentobarbital, and alprozolam due to insufficient data, as previously mentioned. The covariates we used are demographics (age, gender, ethnicity, and region) and clinical state features (Glasgow Coma Scale (GCS), Acute Physiology Score (APS), APACHE score, ethanol intoxication status, diagnoses, and intubated status). The outcomes of interest are the hospital discharge location (death, remains hospitalized, nursing home, rehabilitation, and home), hospital length of stay, hospital ventilation days, and mortality. For this research project, we focused on predicting the hospital discharge location.

2.3.1 Data Cleaning

There were numerous data cleaning steps necessary to obtain this data set from the eICU-CRD. Since we only wanted to consider the first ICU stay so our results would not rely on later information, we needed to discard hospital visits that did not have a recorded first ICU stay. We also needed to remove

data entry errors such as negative values for length of stay, GCS, and ventilation days. Finally, we removed non-TBI patients to narrow the patient cohort to include only patients in the ICU with TBI.

2.3.2 Data Integration

The data integration steps involved combining tables and other sources of data together to obtain the final data set [1]. The patient table provided us with the patient tracking information, gender, ethnicity, age, and discharge location. The infusion drug table provided us with the names of the drugs the patient received and we used it to determine if he or she received a particular sedative. Similarly, we used the medication table to determine if a patient had an order for a particular sedative. We assumed that a clinician infused the patient with the ordered sedative because without this assumption there would be many intubated and ventilated patients who did not receive any sedative, which is unlikely. We used the diagnosis table to narrow the patient cohort to TBI patients using the International Classification of Diseases, Ninth Edition, Clinical Modification (ICD-9-CM) codes. We classified a patient as having TBI using the list of codes provided by the Center for Disease Control ICD-9-CM TBI surveillance report [11]. See table 4.1 for a listing and description of these codes. We used the hospital table to determine the regions of the hospitals that admitted the patients. The lab table was useful for determining the ethanol level of the patients to determine if he or she met the criteria for ethanol intoxication, which we defined as having an ethanol level greater than 80 mg/dL within 12 hours of the hospital admission time. We also sourced an ICD-9-CM table from the Observational Medical Outcomes Partnership (OMOP) in order to augment the eICU-CRD with a description of the ICD-9-CM codes for TBI and alcohol-related diagnoses [20].

2.3.3 Data Transformation

The data transformation steps involved feature engineering, aggregations, translations, and normalization [1]. To normalize the continuous variables in the data set, we subtracted the mean (demeaned) and divided by the standard deviation. This centers and scales the variables so they each have a mean of zero and standard deviation of one. The purpose of this normalization is to prevent variables with large values from dominating the results by using the same standardized unit for all the continuous variables. We also created binary flags for each experiment:

- Monotherapy binary flags indicate whether the patient received only one exposure within one week of the hospital admission time for the following drugs: fentanyl, propofol, versed, ativan, and precedex.
- Dual therapy flags indicate whether the patient received only the first or second exposure within one week of hospital admission time for the following drug pairs: fentanyl/propofol, fentanyl/versed, fentanyl/ativan, and fentanyl/precedex.
- Triple therapy flags indicate whether the patient received only the first, second, or third exposure within one week of hospital admission time for the following drug groups: fentanyl/propofol/versed, fentanyl/propofol/ativan, fentanyl/precedex/versed, and fentanyl/precedex/ativan.

Additionally, we translated the age field in order to make it numeric instead of categorical and created binary flags to indicate a TBI diagnosis and its priority (primary or secondary). Finally, we engineered a feature for ethanol intoxication using the lab results and ranked the discharge locations by referencing the clinicians on the team.

After this pre-processing, our data set was complete and ready for data exploration and model building and evaluation. We provide a detailed description of each feature of interest in the feature description section of the supplementary materials.

2.4 Data Exploration

The data exploration step involves summarizing the data to provide overall context. It allowed us to detect outliers, visualize potential relationships between the sedatives and outcomes of interest, perform feature selection, and develop parsimonious models to help satisfy modeling assumptions [1].

Exploratory data analytics (EDA) can be graphical or non-graphical. Graphical methods are preferred so domain experts can more readily identify patterns and trends and detect anomalies that may not be evident from a tabular structure. However, non-graphical methods can sometimes provide more precise measurements and information [1]. We performed EDA on the training set only so we could use it in the model formulation process without biasing the model by using the test set.

2.4.1 Descriptive Summary Statistics

We began with a descriptive statistics table that partitions the patients into categories (survived, died, or unknown). For each feature, we expressed its characteristics using a set of parameters in each of these categories. These parameters depend upon the data type of the feature. We one hot encoded each categorical feature into multiple binary features and provided its group count and frequency (percent) in each category. For example, we expanded the region feature into four mutually exclusive binary features: South, West, Midwest, and Northeast. The flag is set for exactly one of these new features to

indicate the region of the hospital the patient visited. For quantitative features, we show the central tendency (mean, median, mode), its spread (sample standard deviation, interquartile range, minimum and maximum), and its distribution (skewness and kurtosis) for each category. See the summary statistics section of the supplementary materials for the full tabulation.

By reviewing these summary statistics, we noted differences in patient baseline characteristics to identify possible confounding variables. These variables are not exposures of interests (sedatives) but nonetheless contribute to the outcome of interest (hospital discharge condition). For example, the APACHE severity of illness score is noticeably higher among the cohort of patients who died compared to those that survived. In fact, the median APACHE score of the patients who died (87) is nearly twice the median score of the patients who survived (44). We also noted the values of injury-severity related variables and found that the patients who died had a shorter length of stay, greater ventilator days, higher age, and a lower GCS score. Clearly, severity of illness and other clinical variables accounts for some of the variance in the hospital discharge condition and we must control for them using adjustment methods (multivariable linear regression, analysis on survivors alone as a subgroup etc.) [1].

These statistics also provide context to our data set that we use to determine possible sources of bias (e.g. demographic bias) that influence the external validity (generalizability) of the model. For example, our data set consists largely of white males over 50 years old so models developed using this data set may not generalize well to cohorts of patients who are younger or of a different age or ethnicity. Hence, while the eICU-CRD provides fair regional diversity, it has biases for age, gender, and ethnicity. See Table 2.2 for a detailed overview of these demographics.

Age	
Mean $\pm$ SD	60.9 $\pm$ 22
Min - Max	11 – 90+
Median (IQR)	65 (45 – 80)
Gender	
Male	1731 (60.1%)
Female	1137 (39.5%)
Other/Unknown	9 (0.3%)
Ethnicity	
African American	198 (6.8%)
Hispanic	130 (4.5%)
Native American	23 (0.7%)
Asian	63 (2.1%)
Caucasian	2272 (78.9%)
Other/Unknown	191 (6.6%)
Region	
South	725 (25.1%)
West	701 (24.3%)
Midwest	981 (34%)
Northeast	286 (9.9%)
Total	2877 (100%)

Table 2.2: Demographic diversity of our training set from the eICU-CRD. For age, we report the mean, sample standard deviation (SD), minimum, maximum, median, and interquartile range (IQR). For the other demographic variables, we report the number of patients and the percent of the total population they represent. The data set is regionally diverse but is not diverse in terms of age, gender, or ethnicity. Most of the patients are white males over 60 years old.

2.4.2 Graphical Methods

The graphical methods we employed involved using bar charts, box plots, and histograms to provide an overview of the class distribution. This helps with satisfying modeling assumptions for the traditional models in the subsequent model building step.

We used a bar chart as a univariate graphical EDA technique to provide an overview of the class distribution for our primary outcome of interest (hospital discharge location). Most of the patients went home after discharge and few remained hospitalized. This unequal class distribution is evident in Figure 2.2. We consider this distribution to select the appropriate model for statistical analysis. Such an imbalance can make it difficult to satisfy modeling assumptions for traditional models such as linear regression. Failing to satisfy these assumptions makes the results less reliable.

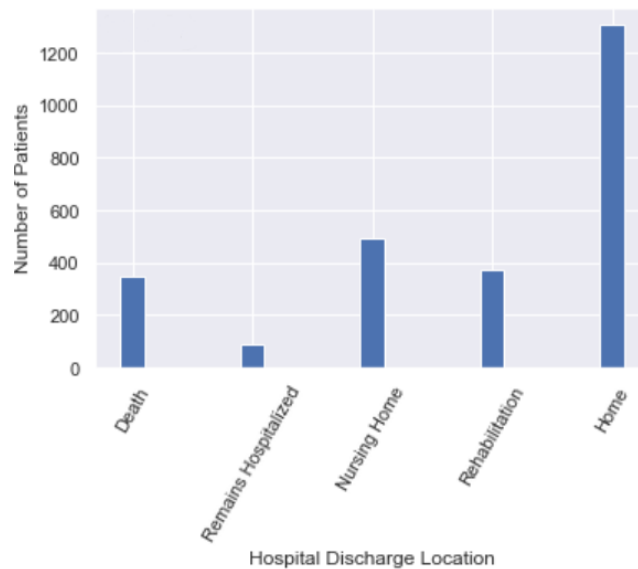


Figure 2.2: Class distribution for hospital discharge location.

We also used univariate graphical EDA techniques on other outcomes of interest (length of stay, mortality, and ventilation days). We used boxplots and histograms for quantitative outcomes and bar charts for categorical outcomes. The high positive skew of the ventilation days (Figure 2.3) and length of stay (Figure 2.4) are evident in the visualizations. Similarly, the unequal class distribution for mortality is evident in Figure 2.5. We considered these distributions to select the appropriate model for statistical analysis.

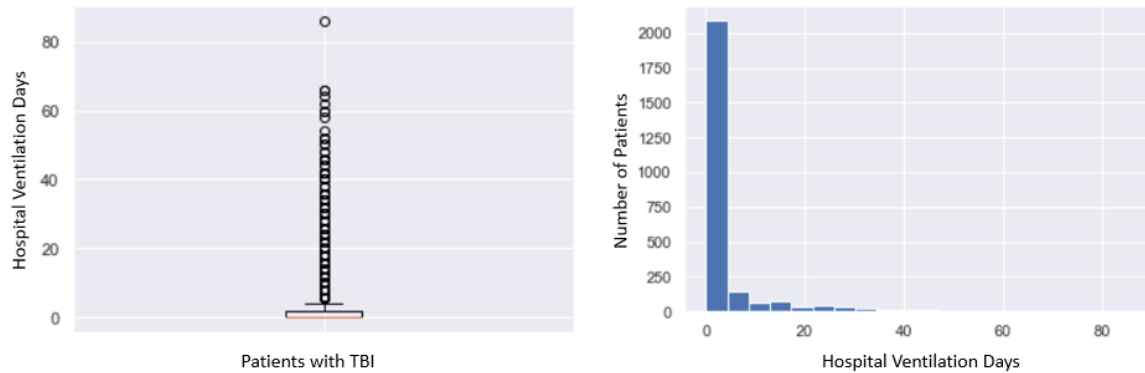


Figure 2.3: Boxplot (left) and histogram (right) showing the distribution of hospital ventilation days. The outliers in the boxplot were determined using the 1.5 times IQR rule. Thus, patients with over five ventilation days are outliers.

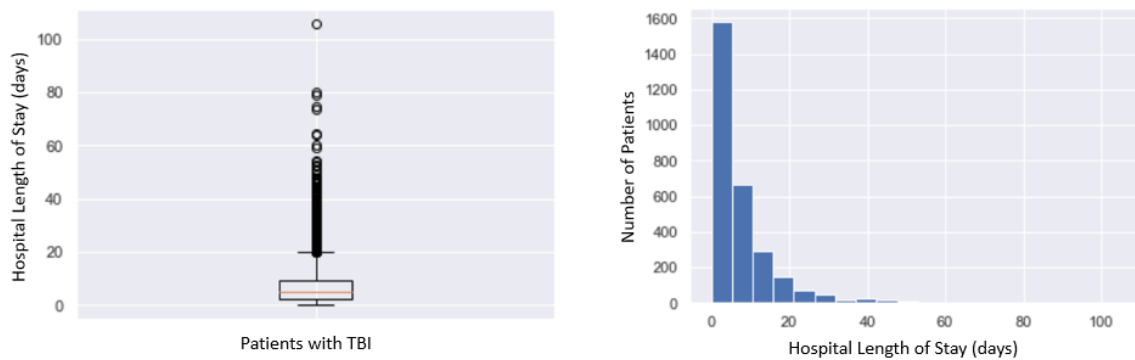


Figure 2.4: Boxplot (left) and histogram (right) showing the distribution of hospital length of stay (days). The outliers in the boxplot were determined using the 1.5 times IQR rule. Thus, patients with a hospital length of stay greater than 20.05 days are outliers.

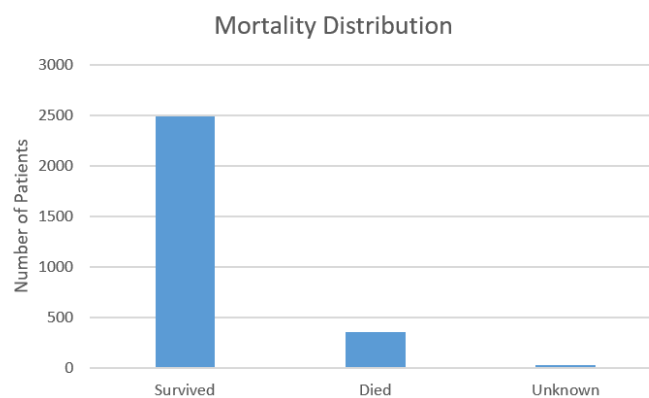


Figure 2.5: Bar chart showing the mortality distribution of patients with TBI.

Hence, the steps in EDA were valuable for multiple steps in the data analytics pipeline. It was useful for providing context to the analysis, identifying possible confounding variables, and understanding the distribution, spread, and central tendency of each feature. This allowed us to perform feature selection and transformation and provided guidance for selecting the appropriate model for statistical analysis. It also allowed us to contextualize our findings so we could make real world insights considering the external validity (generalizability) of the models with respect to possible sources of bias.

2.5 Model Building

The model building step involved creating traditional, GAM, and new deep learning models to address our research question. These models have hyperparameters that underwent fine-tuning using a validation set in the subsequent model evaluation step.

We built one of each type of model for three independent experiments, one for each kind of therapy (monotherapy, dual therapy, and triple therapy). The monotherapy experiment used flags to indicate if the patient received only one of the sedatives early in his or her first ICU stay (within one week of ICU admission time). The dual therapy experiment used flags to indicate if the patient received at least one of two particular sedatives. The triple therapy experiment used flags to indicate if the patient received at least one of three particular sedatives. Thus, there were three of each type of model, one for each experiment.

2.5.1 Linear Regression

The traditional model we used was linear regression (LR) [21]. The objective of LR is to predict the dependent variable y given the independent variables x_0 to x_N by minimizing the loss using a linear equation fitted to the training data. In least-squares regression, this involves minimizing the sum of squares of the residuals (the vertical distance between the prediction $E(y)$ and the actual value of y) by fitting the coefficient β for each variable x_i . The hypothesis equation for standard linear regression has the form

$$E(y) = \sum_{i=1}^N \beta_i x_i$$

where $E(y)$ is the expected value of y for the sample x , N is the number of variables in the sample, x_i is the value of the i^{th} variable of the sample x , and β_i is the coefficient. The loss function for standard linear regression has the form

$$L(\beta) = \sum_{j=1}^M (E(y)_j - y_j)^2$$

where $L(\beta)$ is the loss for the parameter vector β , M is the number of samples, $E(y)_j$ is the expected value of y for the j^{th} sample, and y_j is the actual value. The goal of LR is to find the parameters β that minimize the loss.

This method is widely used in healthcare because it produces interpretable results such that clinicians and data science researchers can readily identify the factors that led to the prediction. Thus, in addition to making a prediction, the linear regression model produces valuable statistical measurements for each feature. The p -value measures its statistical significance, the coefficients measure its practical

significance, and the standard error and confidence intervals measure uncertainty and precision. Additionally, the adjusted R-squared measures the overall model fit and the root mean squared error (RMSE) measures the error. In our case, the dependent variable we are predicting is the hospital discharge location and the independent variables are the sedatives and demographic and clinical state covariates. We will use the statistical measurements from the output to interpret our results in the real world insights section.

We built the linear regression models using a forward stepwise multiple regression process [22] using the R programming language [23] with RStudio [24]. Multiple regression means that we used multiple covariates and forward stepwise means that we iteratively added covariates to our model for numerous steps until we arrived at our final model. The formula at the first step only included the exposures of interest (sedatives) as the independent variables and the outcome of interest (hospital discharge location) as the dependent variable. After fitting this model, we recorded the step number, formula used, significant independent variables as determined by the p-value, and the adjusted R-squared. For the next step, we updated the formula by removing non-significant independent variables and adding one of the other covariates to the independent variables in the formula. Subsequently, we fitted the new model and recorded the results in a similar manner. We continued this process until we had no covariates left. Finally, we tried adding in variables that we previously dropped out due to being insignificant and keep them if they were significant in the new model. The final model consisted of only the exposures of interest and the significant covariates. The significant covariates for each experiment were age, the Northeast region, the APACHE score, verbal and motor GCS values, and ethanol intoxication. Finally, we tested the final model to assess if the linear model assumptions were satisfied using the global test on four degrees of freedom using the 0.05 significance level, which tests for satisfying the following

assumptions: global stat, skewness, kurtosis, link function, and heteroscedasticity. For each model, only the heteroscedasticity modeling assumptions were satisfied.

2.5.2 Deep Learning

The new deep learning model we used was a feedforward multiple layer perceptron (MLP) [25], which typically produces more accurate results at the cost of interpretability and scalability [2]. This model uses a neural network architecture. The neural network consists of an input layer, hidden layers, and an output layer. Each layer consists of neurons that represent features. The neurons in the input layer are the features of the data set, which feed into the next layer. The neurons of the hidden layers receive the outputs from the neurons in the previous layer as inputs. Subsequently, they compute their output by performing a dot product of the inputs with their associated weights (parameters) and applying an activation function to the result to introduce non-linearity. The output layer makes the final prediction. Thus, each neuron has the form

$$\alpha_i^l = g\left(\sum_{i=1}^N \beta_i x_i\right)$$

where α_i^l is the output value of the i^{th} neuron in layer l , $\sum_{i=1}^N \beta_i x_i$ is the linear combination of the inputs x_i with their associated weights β_i , and g is the activation function. The feedforward process involves feeding samples (or batches of samples) into the network and making predictions at the output layer. The loss function computes the error using the actual and predicted values in a manner similar to linear regression. The network learns by backpropagating the error and adjusting the weights using the delta

rule, which involves taking the derivative of the loss function with respect to the weights and multiplying by a learning rate. The delta rule has the form

$$\Delta\beta_i = -\eta \frac{\partial L}{\partial \beta_i}$$

where $\Delta\beta_i$ is the change in the weight β_i , η is the learning rate, and $\frac{\partial L}{\partial \beta_i}$ is the partial derivative of the loss function L with respect to the weight β_i . This training repeats over many passes through the training data set (epochs) to decrease the error gradually by descending the error landscape in this algorithm called gradient descent. Eventually, the algorithm converges (the error ceases to decrease) and the model is ready for evaluation. In the resulting model, the neurons in the hidden layers represent abstract features that that model learned from the data to reduce the error. Unfortunately, these features are difficult for domain experts to interpret. The model is essentially a “black box”, meaning it is difficult to identify the factors that led to the output since it does not produce interpretable statistical measures such as those in the linear regression output. However, the accuracy of the MLP is generally better because of its ability to discriminate using more complicated (non-linear) functions, though it may require careful construction of the neural network architecture and hyperparameter selection and fine-tuning. Hyperparameters refer to parameters that we tune during the training process to optimize the model. For example, the dropout rate is useful for introducing regularization into the neural network [26]. It involves randomly dropping neurons during the training process to prevent the neurons from co-adapting (relying on other neurons). The learning rate is another hyperparameter that determines the step size for moving in the direction opposite of the gradient when descending the error landscape to minimize the loss. Other hyperparameters include the batch size, number of epochs, number of neurons in each layer, and number of layers.

We built the deep learning model using the full feature set and optimized the hyperparameters and model architecture using randomized search [27] and manual fine-tuning. Randomized search involves randomly sampling parameter values from a distribution and choosing the combination of values that resulted in the best estimator. We performed randomized search over the batch size, dropout rate, number of epochs, number of units (neurons) in the hidden layers, number of hidden layers, and the learning rate. For the batch size, we used a distribution of random integers from one to 128. For the dropout rate, we used a uniform distribution from zero to 0.9. For the number of epochs, we used a distribution of random integers from 10 to 30. For the number of units, we used a distribution of random integers from two to 10. For the number of hidden layers, we used a distribution of random integers from one to five. For the learning rate, we used a uniform distribution from 0.0001 to 0.3. Finally, we manually set the activation functions to rectified linear unit (ReLU), the optimizer to the adam method, and the loss function to the mean squared error. We chose ReLU activation functions [28] because they allow the model to learn non-linear relationships and do not suffer from the vanishing gradient problem. The ReLU activation function has the form

$$\text{ReLU}(z) = \max(0, z)$$

where z is short for $\sum_{i=1}^N \beta_i x_i$ in the equation for neuron output above. We chose the adam optimization method [29] because it uses an adaptive learning rate with a decay factor to prevent the step size from becoming too small. It also uses the concept of “momentum” to prevent getting stuck in saddle points and local minimums in the error landscape. We performed the randomized search using fivefold cross validation over 10 iterations (random selections from the parameter distributions). The randomized search found the same parameters for each best estimator. It found that the optimal batch size was 66, the optimal dropout rate was 0.129, the optimal number of epochs was 10, the optimal number of units in the hidden layers was nine, the optimal number of hidden layers was one, and the optimal learning

rate was 0.156. We implemented this model using the Keras library [30] for the Python programming language [31] and performed the randomized search using tools from the Scikit-learn library [32].

2.5.3 Regularized Linear Regression

In addition to standard linear regression, we also used linear regression with regularization. In particular, we used ridge, lasso, and elastic net regression. Each form of regularization introduces a penalty term λ to the coefficients in the loss function to penalize model complexity and prevent overfitting. In lasso regression [33], we multiply the penalty term by the summation of the absolute values of the coefficients, which has the effect of shrinking the less important feature coefficient values to zero and is useful for feature selection. The new loss function for lasso regression has the form

$$L(\beta) = \sum_{j=1}^M (E(y)_j - y_j)^2 + \lambda \sum_{i=1}^N |\beta_j|$$

where $L(\beta)$ is the loss for the parameter vector β , M is the number of samples, $E(y)_j$ is the expected value of y for the j th sample, y_j is the actual value, λ is the penalty term, and N is the number of features in each sample. In ridge regression [34], we multiply the penalty term by the summation of the squared values of the coefficients, which causes the less important feature coefficient values to shrink but not necessarily to zero. The new loss function for ridge regression has the form

$$L(\beta) = \sum_{j=1}^M (E(y)_j - y_j)^2 + \lambda \sum_{i=1}^N \beta_j^2$$

In elastic net regression [35, 36], we use both the lasso and ridge enhancements. The new loss function for elastic net regression has the form

$$L(\beta) = \sum_{j=1}^M (E(y)_j - y_j)^2 + \lambda_1 \sum_{i=1}^N |\beta_i| + \lambda_2 \sum_{i=1}^N \beta_i^2$$

In each case, the goal is to find the parameters β that minimize the loss. We implemented all regularized linear regression models using classes from the Scikit-learn library for Python [37, 38, 39].

We built the regularized linear regression models using the full feature set to allow the model to learn the important features and set the coefficient values accordingly. We performed randomized search for selecting the hyperparameter penalty term λ using fivefold cross validation over 10 iterations (random selections from the parameter distributions). The randomized search found that the optimal penalty term λ value was 0.02 in each case.

2.5.4 Generalized Additive Models

Finally, we built a standard and pairwise generalized additive model to test if it can maintain reasonable interpretability without sacrificing the accuracy typically found in deep learning models. These models use smoother functions over the variables to introduce non-linearity and thereby increase accuracy. However, they maintain interpretability because they produce statistical metrics for each variable while not requiring satisfying the modeling assumptions of LR, thereby providing greater generalizability. Furthermore, we can plot each independent variable with the dependent variable on a two-dimensional plot [9]. The hypothesis equation for standard GAMs have the form

$$g(E(y)) = \sum_{i=1}^N f_i(x_i)$$

where $E(y)$ is the expected value of y for the sample x , N is the number of variables in the sample, x_i is the value of the i th variable of the sample x , and f_i is the smoother function.

Researchers have also successfully allowed for pairwise interactions to increase accuracy while maintaining interpretability by producing a heat map illustrating the effect of the interaction on the dependent variable. The researchers called GAMs that allow for these pairwise interactions GA^2Ms . The hypothesis equation for GA^2Ms has the form

$$g(E(y)) = \sum_{i=1}^N f_i(x_i) + \sum_{j=1}^N f_{i,j}(x_i, x_j)$$

where f_i is the smoother function and the second term represents the pairwise interactions.

Thus, we built GAM and GA^2M models for our analysis to determine if they can sufficiently preserve both interpretability and accuracy for our research question [10]. We built a GAM model for each experiment (mono, dual, and triple therapy) but only built the GA^2M model for the monotherapy experiment because the model learns interaction terms and so does not require manually combining sedatives into therapy groups. We implemented the GAM using R and the GA^2M using the Machine Learning Tool Kit Library [40] for the Java programming language [41].

2.6 Model Evaluation

After model building, we evaluated the models using metrics for accuracy, interpretability, generalizability, and scalability. After conducting each experiment, we compared the models to identify

their relative strengths and weaknesses. See tables Table 2.3, Table 2.4, and Table 2.5 for a tabular comparison of the models for each experiment.

Models	Linear Regression	Deep Learning	Generalized Additive Model	GA ² M	Elastic Net	Lasso	Ridge
RMSE	1.0758	1.05	1.0734	1.08	1.08	1.08	1.08
Training Time	0	~1s per epoch	0	166s	0	0	0
Test Time	0	~1s	0	0	0	0	0

Table 2.3: Accuracy and scalability of the various models for the monotherapy experiment.

Models	Linear Regression	Deep Learning	Generalized Additive Model	Elastic Net	Lasso	Ridge
RMSE	1.0738	1.06	1.0734	1.08	1.08	1.08
Training Time	0	~1s per epoch	0	0	0	0
Test Time	0	~1s	0	0	0	0

Table 2.4: Accuracy and scalability of the various models for the dual therapy experiment.

Models	Linear Regression	Deep Learning	Generalized Additive Model	Elastic Net	Lasso	Ridge
RMSE	1.0734	1.0725	1.0734	1.08	1.08	1.08
Training Time	0	~1s per epoch	0	0	0	0
Test Time	0	~1s	0	0	0	0

Table 2.5: Accuracy and scalability of the various models for the triple therapy experiment.

2.6.1 Accuracy

We evaluated the accuracy of the models by estimating their true predictive power using unseen data. This provided us with insights into the ability of the models to make good predictions on other populations. We did this by partitioning our data set of 3836 unique patients into training and testing sets. The training set consisted of 75% of the patients and the testing set contained the remaining 25%.

We used the training set for the model building and exploratory data analytics steps. Its purpose was to provide data for the model to fit its parameters in the model formulation process. We used fivefold cross validation [42] to provide an unbiased evaluation of the model fit on the training data. This evaluation helped prevent overfitting and guided us in fine-tuning the hyperparameters to optimize the model. We reserved the testing set for the model evaluation step. Its purpose was to make predictions on unseen data after fitting the model during the training process. The specific metric we used to compare the accuracy of the models was the root mean squared error (RMSE).

Our results show that the accuracy of the models differed slightly between the different approaches and experiments. In each case, the deep learning model outperformed the GAM model which outperformed linear regression. However, this difference was marginal and added model complexity and reduced interpretability and scalability. We believe that the lack of greater performance enhancements in the more complex models due to the simplicity of the data set, which consists of the exposures and outcome of interest and the covariates, many of which are binary variables. We also observed that adding regularization to the linear regression models did not result in performance improvements, suggesting that our linear regression model was not overfitting. However, this is likely due to the fact that we used a forward stepwise process and manually kept only significant variables for the final model. On the other hand, the regularized LR models used the full feature set and were able to learn the most significant features. The GA²M model also did not result in performance enhancements, though it provided additional interpretability through pairwise interaction heat maps.

2.6.2 Interpretability

Evaluating the interpretability of the models involved presenting a summary of the final models and their results to the domain experts on the team. For example, the linear regression summary included the covariates along with their p-values, coefficients, confidence intervals, and standard errors. Additionally, it included the adjusted R-squared and the root mean squared error. The domain experts described their ability to understand the summary and make meaningful real world insights.

Our discussion showed that the linear regression, GAM, and the GA²M models produced the most interpretable results, which is what we expected. Each linear regression model showed each covariate and exposure along with a measure for the statistical significance (p-value), practical significance (coefficient), and uncertainty and precision (standard error and confidence interval). They also produced a measure for the overall model fit (adjusted R-squared) and the error (root mean squared error). The GA²M model produced pairwise interaction heat maps that provided insights into the interaction between the covariates and exposures. See the real world insights section and tables for the full results and the insights they provide. This interpretability led to real world insights for clinicians, whose problem motivation was to have guidance for sedative choice for patients with TBI, as previously discussed. We discuss these insights in the subsequent real world insights section below. As expected, the deep learning model was the least interpretable. It did not produce valuable statistics for each covariate and exposure or provide a measure for the overall model fit. Its results showed the error for each epoch it trained on and the final prediction.

2.6.3 Scalability

Evaluating the scalability of the models was necessary to determine their ability to perform efficiently, particularly on large amounts of data. This is especially important in healthcare because electronic health records often contain large quantities of data. We measured the scalability of the models by recording the training and testing times.

Our results showed that the linear regression and generalized additive models were the most scalable in terms of training and testing time. Both models did not take any significant amount of time to train or test, though the manual forward stepwise process took about 30 to 45 minutes per model. On the other hand, the deep learning model took about a second to train and test per epoch. However, making a prediction for one sample did not take a significant amount of time after we built and optimized the model. Nevertheless, building and optimizing the deep learning model took a significant amount of time. Since it was more complicated, we had to learn a deep learning framework and preprocess our data accordingly. Additionally, the randomized search process for optimizing the model architecture and tuning hyperparameters took about 30 minutes per model. We also performed randomized search for the regularized linear regression models, however, it did not take as long because there were fewer hyperparameters. The GA²M model took about 166 seconds to build due to the time it took to create the pairwise interactions.

2.6.4 Generalizability

Evaluating the generalizability of the models was necessary to determine the ability of the models to generalize to other domains and class distributions. We measured this for each model by assessing the

difficulty of satisfying modeling assumptions and processing requirements while maintaining interpretability.

We found that the generalized additive models were the most generalizable. They were able to maintain interpretability without requiring the modeling assumptions of linear regression. The linear regression models did not generalize well due to the difficulty of satisfying the modeling assumptions, which is often a consequence of the imbalanced class distribution. While there are strategies to address this (log transformations of the dependent variable, etc.), they usually hinder the interpretability of the model. The deep learning models also did not generalize well because it required a significant amount of preprocessing, searching, and development, as previously discussed. Thus, applying it to different domains would require additional data cleaning, integration, and transformations.

Finally, we compared the models to identify their relative strengths and weaknesses. By doing so, we provide guidance into the relevant factors to consider when choosing a model. In our case, the standard linear regression models and the generalized additive models are the best choice because of their greater interpretability and scalability. The slight increased accuracy of deep learning are not worth the loss of interpretability and scalability and additional model complexity. However, the improvements in accuracy may be greater with more data and more important for other research questions. Choosing between linear regression and the generalized additive model depends upon whether linear regression modeling assumptions can be fully satisfied. Linear regression was simpler but the reliability of the statistical results (p-values, etc.) is contingent upon satisfying modeling assumptions, making it less generalizable. Though the GA²M model did not improve accuracy in our case, it provided additional interpretability through the pairwise interaction heat maps. Additionally, it is likely to improve

performance with a more complicated data set with more samples and features. However, it took more time to train and create the heat maps for each interaction than standard GAMs and LR, which is significant when scalability is a concern. Adding regularization to linear regression prevented overfitting and allowed the model to learn the significant features so the manual forward stepwise process was not necessary. However, it added an additional hyperparameter that required fine-tuning and, in our case, did not improve performance because our linear regression model was not overfitting, likely because the forward stepwise process filtered out the less important features.

Thus, we evaluated the models individually using metrics for accuracy, interpretability, scalability, and generalizability. Subsequently, we compared them to identify their relative strengths and weaknesses and found that, in our case, linear regression and the generalized additive models were preferable to deep learning. The accuracy of the models did not differ significantly. The linear regression and the GAM models were the easiest to interpret. The GA^2M model enhanced this interpretability through the pairwise interaction heat maps. Both the linear regression and the standard generalized additive models were scalable. However, the GA^2M model does not scale as well due to additional calculations required for the pairwise interactions. The generalized additive models were able to generalize well to other domains without the significant preprocessing necessary to satisfy modeling assumptions or build and optimize the model.

3 Real World Insights

The real world insights of this research project are for clinical experts and data science researchers. These include insights into the effects of various sedatives and covariates on patients with TBI and the strengths and weaknesses of different modeling approaches in the healthcare domain.

3.1 Clinical Implications

For clinical experts, we used the results from linear regression and the generalized additive models to provide insights into the effects of the sedatives and demographic and clinical state covariates. We also provide interpretations for Sankey diagrams and pairwise interaction heat maps from our results.

3.1.1 Effects of Sedatives

The insights into the effects of sedatives on patients with TBI provides guidance on when to use them. Currently, there is limited guidance to reference to determine the best sedative choice in specific cases [3]. However, our research provides some insights into the advantages and disadvantages of each sedative agent to help guide clinicians in making such decisions. Specifically, we found that versed was associated with better discharge dispositions (95% CI: 0 to 1.80, $p = 0.04$). One possible explanation is that versed truly is associated with better discharge dispositions. However, the clinical experts we worked with suggested that another possible explanation is that versed is a more versatile sedative. While clinicians typically use fentanyl, propofol, and precedex just to facilitate intubation, they

sometimes use versed for other purposes such as to still the patient for CT and MRI scans. Thus, the better outcomes associated with versed may be due to this increased versatility and not its use for sedation. We also found that the fentanyl/ativan dual therapy was associated with worse discharge dispositions (95% CI: -0.17 to 0.01, $p = 0.04$). Since the other dual therapies also included fentanyl, this may indicate that ativan has negative effects. The other sedatives and therapies do not have a significant impact on patient discharge dispositions according to our analysis. See tables Table 3.1, Table 3.2, and Table 3.3 for the linear regression results that led to these insights. We provide the coefficient plots illustrating the 95% confidence interval for each variable in each experiment in the supplementary materials section. We believe future research is necessary to confirm these results. In particular, future research should investigate whether the better outcomes for versed were due to its alternative uses or its use for sedation. Additionally, researchers should further investigate the possible negative effects of ativan.

Covariates and Exposures	Coefficient	Std. Error	95% CI	P-value
Fentanyl	-0.27	0.21	-0.69 – 0.13	0.19
Propofol	-0.13	0.09	-0.32 – 0.05	0.17
Versed	0.90	0.45	0.00 – 1.80	0.04
Ativan	0.02	0.78	-1.51 – 1.56	0.97
Precedex	-0.04	0.18	-0.40 – 0.30	0.79
Age	-0.28	0.02	-0.33 – -0.22	< 0.001
Northeast Region	-0.27	0.08	-0.44 – -0.11	< 0.001
APACHE Score	-0.41	0.04	-0.50 – -0.33	< 0.001
GCS Verbal Score	0.24	0.03	0.17 – 0.32	< 0.001
GCS Motor Score	0.24	0.04	0.15 – 0.33	< 0.001
Ethanol Intoxication	0.19	0.08	0.03 – 0.35	0.01

Adjusted R-squared: 0.394

RMSE: 1.0758

Table 3.1: Final model summary for predicting hospital discharge location using monotherapy.

Covariates and Exposures	Coefficient	Std. Error	95% CI	P-value
Fentanyl/Propofol	-0.00	0.07	-0.15 – 0.15	0.98
Fentanyl/Versed	0.42	0.23	-0.04 – 0.89	0.07
Fentanyl/Ativan	-0.59	0.29	-1.17 – -0.01	0.04
Fentanyl/Precedex	-0.06	0.17	-0.40 – 0.27	0.71
Age	-0.28	0.02	-0.34 – -0.22	< 0.001
Northeast Region	-0.28	0.08	-0.44 – -0.11	< 0.001
APACHE Score	-0.41	0.04	-0.49 – -0.33	< 0.001
GCS Verbal Score	0.26	0.03	0.18 – 0.34	< 0.001
GCS Motor Score	0.24	0.04	0.15 – 0.32	< 0.001
Ethanol Intoxication	0.19	0.08	0.03 – 0.35	0.01

Adjusted R-squared: 0.3938

RMSE: 1.0738

Table 3.2: Final model summary for predicting hospital discharge location using dual therapy.

Covariates and Exposures	Coefficient	Std. Error	95% CI	P-value
Fentanyl/Propofol/Versed	-0.09	0.18	-0.46 – 0.27	0.61
Fentanyl/Propofol/Ativan	0.06	0.19	-0.30 – 0.44	0.71
Fentanyl/Precedex/Versed	0.36	0.27	-0.18 – 0.90	0.19
Fentanyl/Precedex/Ativan	-0.49	0.29	-1.07 – 0.08	0.09
Age	-0.28	0.02	-0.34 – -0.22	< 0.001
Northeast Region	-0.26	0.08	-0.43 – -0.10	0.001
APACHE Score	-0.41	0.04	-0.50 – -0.33	< 0.001
GCS Verbal Score	0.25	0.04	0.17 – 0.33	< 0.001
GCS Motor Score	0.24	0.04	0.15 – 0.33	< 0.001
Ethanol Intoxication	0.19	0.08	0.03 – 0.35	0.01

Adjusted R-squared: 0.3932

RMSE: 1.0734

Table 3.3: Final model summary for predicting hospital discharge location using triple therapy.

3.1.2 Effects of Demographic and Clinical Covariates

In addition to these insights regarding the effects of the exposures, we also made insights into the effects of demographic and clinical state covariates. The significant demographic covariates were age and the Northeast region. Older age and the Northeast region were both associated with worse discharge dispositions. We believe that age is truly associated with worse discharge dispositions, however, the

findings for the Northeast region may be due to chance since it was underrepresented compared to the other regions. The significant clinical state factors that negatively influenced discharge dispositions are high APACHE scores and low GCS motor and verbal scores. These results confirm that these severity of illness and level of consciousness measurements have a significant impact on patient discharge locations and need to be controlled for using methods such as multivariate linear regression. It is also insightful to note the non-significant covariates. According to our analysis, the demographic covariates that did not have a significant impact are gender, ethnicity, and the non-Northeast regions. The non-significant clinical state factors were intubation, GCS eyes value, and the priority of the TBI diagnosis (primary or secondary).

3.1.3 Sankey Diagram Interpretation

We created Sankey diagrams for each experiment to aid in our clinical interpretation. These diagrams illustrate our linear regression results showing that versed is associated with better discharge dispositions while fentanyl/ativan is associated with worse discharge dispositions. However, by observing these diagrams, we found that both versed and ativan had fewer samples compared to the other sedatives, which further demonstrates the need for more samples and additional research. We also observed that precedex tends to be associated with better outcomes compared to propofol and fentanyl. This association may be explained by the fact that precedex is more often used on patients with less severe illnesses compared to fentanyl and propofol. However, our linear regression results did not find the difference statistically significant, possibly due to lack of sufficient samples. See Figure 3.1 for the Sankey diagram for each experiment.

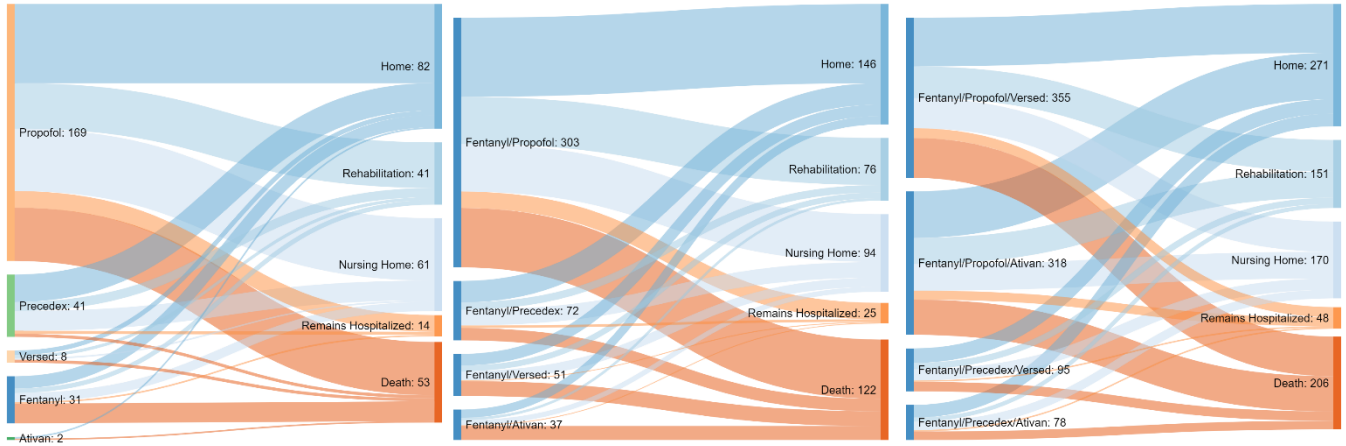


Figure 3.1: Sankey diagrams showing the flow of patients from each sedative group to the different hospital discharge locations. The monotherapy diagram is on the left, the dual therapy diagram is in the middle, and the triple therapy diagram is on the right.

3.1.4 Pairwise Interaction using Heat Map

The GA^2M model provided us with heat maps illustrating pairwise interactions with residual sum of square (RSS) values that measure the strength of the interaction. Among the strongest interactions were between the GCS values (verbal, eyes, and motor) and the APACHE score. By viewing the GCS verbal vs. APACHE heat map, we observe that worse discharge locations were associated with higher APACHE and GCS verbal scores. The GCS eyes vs. APACHE heat map shows that worse discharge locations were associated with higher APACHE and lower GCS eye scores. The GCS motor vs APACHE heat map shows that worse discharge locations were associated with higher APACHE and medium GCS motor scores. Thus, high APACHE scores are consistently associated with worse discharge locations when paired with GCS scores. However, the GCS association varied depending on the category (verbal, eyes, or motor). Based on our knowledge of the domain, we believed that higher GCS scores were generally worse for patient outcomes. Thus, this variance may be due to confounding factors that the APACHE score includes in its calculation, making it a more stable and reliable predictor

of discharge disposition. Another possible interpretation is that a low GCS score does not necessarily correlate to better discharge dispositions because patients with a lower GCS score (between three and six) tend to be unconscious and remain so during their stay. On the other hand, patients with a higher GCS score (between seven and 10) tend to remain conscious. Thus, while GCS provides a good measure of a patient's level of consciousness, it may not be as reliable as severity of illness metrics that consider additional factors such as APACHE, at least when it comes to predicting discharge disposition. See Figure 3.2 for the heat maps illustrating the interaction between the APACHE and GCS scores. Another interesting finding is that propofol had the strongest interactions with the GCS and APACHE scores among the various sedatives. Patients who received propofol and had a high APACHE score tended to have worse discharge locations than those who did not. However, its association when paired with the GCS scores also varied depending on the category. See Figure 3.3 for the heat maps illustrating the interaction between propofol and the APACHE and GCS scores. See the GA²M results section of the supplementary materials for additional pairwise interaction heat maps.

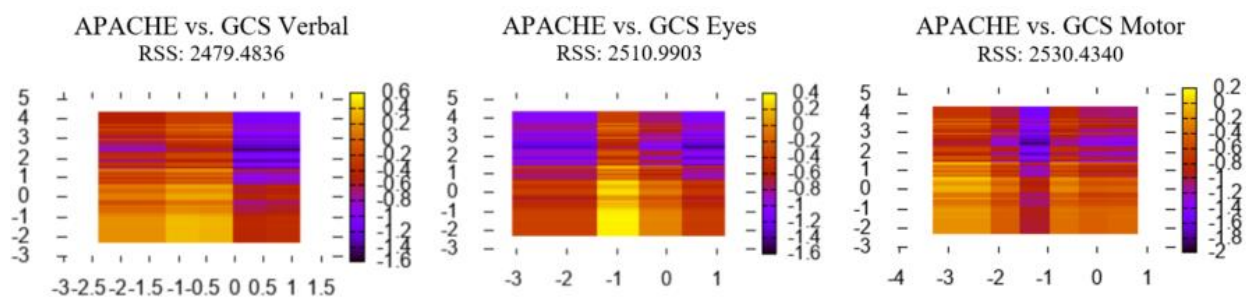


Figure 3.2: Heat maps showing the pairwise interactions between the APACHE and GCS scores.

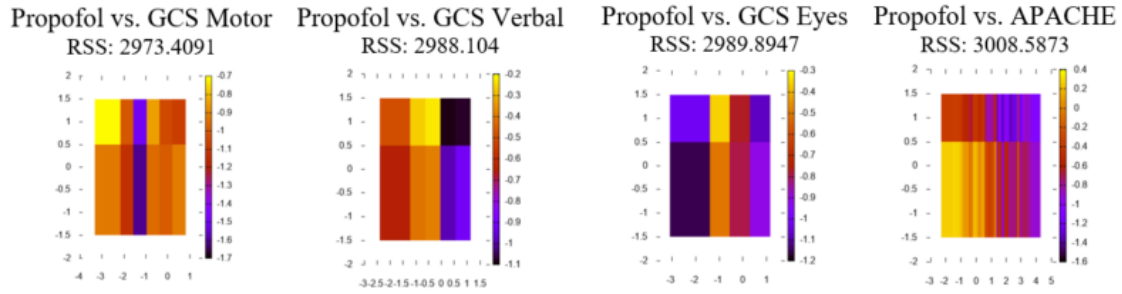


Figure 3.3: Heat maps showing the pairwise interactions between propofol and the APACHE and GCS scores.

3.2 Data Science Insights

For data science researchers, we made insights into the various modeling approaches in the healthcare domain. We compared and contrasted traditional linear regression methods with new deep learning methods and the generalized additive models. In our case, we found that the deep learning models achieved the best RMSE using the full feature set after fine-tuning the hyperparameters, though the other models closely followed it. However, linear regression and the generalized additive models produced the most interpretable results. The generalized additive models produced the most generalizable results since they did not require significant preprocessing to satisfy modeling assumptions or prepare it for deep learning. The GA^2M model provided greater interpretability than the standard GAM due to the pairwise interaction heat maps. Both linear regression and the generalized additive models were efficient and more scalable than deep learning in terms of training and testing time. However, standard linear regression often involves a time-consuming forward stepwise process, though more complex models like deep learning involve a time-consuming process of fine-tuning hyperparameters. Thus, deep learning does not guarantee significant improvements in accuracy, particularly on parsimonious models, and often comes at the cost of interpretability, generalizability, and scalability. Thus, the best model

choice is contingent upon the priority of these evaluation metrics and the complexity of the data set and research question. Hence, it is imperative for data science researchers to understand the tradeoffs between accuracy, interpretability, generalizability, and scalability in the context of their research question so they can more effectively build and evaluate their models.

Hence, the real world insights pertain to both data science researchers and clinical experts. The insights for clinical experts provide guidance into sedative agent selection for patients with TBI and the effects of each exposure and covariate on patient outcomes. The insights for data science researchers help them understand the tradeoff between accuracy, interpretability, generalizability, and scalability for models used in the healthcare domain. Thus, this research provides actionable insights for both our specific research question and healthcare analytics generally. By doing so, we seek to contribute to our understanding of using models to perform clinical predictive tasks to improve the quality of clinical care.

4 Conclusion

In conclusion, we made many valuable insights into different modeling approaches and the effects of sedatives and other variables on patients with TBI. Nevertheless, there are opportunities for future research based on the lessons we learned.

4.1 Lessons Learned

We learned many lessons throughout the course of this research and there are numerous things we would have done differently given the opportunity.

4.1.1 Common Problems in Modeling for Healthcare

By reviewing the relevant literature and previous work, we learned about some of the common problems in modeling for healthcare analytics. First, we learned about the problem of class imbalance, which makes it difficult to satisfy modeling assumptions and account for different costs of misclassification errors. For example, the cost of falsely classifying a patient as healthy can lead to premature discharge, which is harmful for the patient and can result in expensive unscheduled readmissions. Second, we learned about the challenge of capturing temporal information such as the sequence of lab results or clinical visits over time to provide information about disease progression. This leads to representing a patient as a sequence of hospital and ICU visits, which is evident in the eICU-CRD patient tracking

system previously discussed in the data description section. Third, we learned about the challenge of effectively using both human-derived and machine-derived features to make predictions to ensure that models consider both the knowledge and experience of clinical experts and information recorded in EHRs [43].

These problems were more evident as we encountered them firsthand to address our research question, particularly when we applied the deep learning approach and evaluated its strengths and weaknesses compared to traditional methods. The strengths of deep learning are that it typically improves accuracy by learning abstract features from the data itself rather than relying on educated assumptions from domain experts. Thus, it learns machine-derived features at the hidden layers while using human-derived features at the input layer. However, this comes at the cost of scalability and interpretability since the deep learning models require more data and time to train and the factors influencing the prediction are difficult to decipher.

During the deep learning model formulation process, we learned about the challenge of finding the optimal neural network architecture and hyperparameter values, namely through manual fine-tuning and randomized search. Initially, we formulated the model manually using a deep architecture with many hidden units. However, we realized that this strategy was not reducing the error so we decided to investigate alternatives. Thus, we found two alternative methods: grid search and randomized search. Grid search allows you to specify the parameters you want to search through and returns the combination of parameters resulting in the best estimator. On the other hand, randomized search allows you to specify a distribution to select from for each parameter and returns the combination of parameters resulting in the best estimator. Ultimately, we decided to use the randomized search method so we

would not rely on our assumptions to guide the parameter selection process, instead specifying a distribution to select from at random. We found that the best estimator used a much simpler model than the model we constructed manually, consisting of a single hidden layer and nine hidden units. Thus, we learned that parsimonious models are sometimes preferable and that overly complex models often hinder performance in terms of both accuracy and scalability.

4.1.2 Strengths and Weaknesses of Modeling Approaches

We also learned about the challenges and strengths and weaknesses of the different modeling approaches. In the deep learning approach, it is challenging to interpret the output and find the optimal architecture and hyperparameter values. On the other hand, linear regression was interpretable but it was challenging to satisfy all the modeling assumptions and it tended to be slightly less accurate. The GAM models, however, provided a good balance of interpretability and accuracy and did not require satisfying the modeling assumptions.

Having learned these lessons, there are numerous things we would have done differently given the opportunity. Some of these things include doing a literature review earlier, evaluating other outcomes of interest such as length of stay and ventilation days, and testing other techniques to improve the models. Doing the literature review earlier would have exposed us to the three aforementioned challenges in healthcare analytics and guided us in testing other techniques from related work to improve the models. For example, [43] found that using an LSTM to capture temporal information significantly improved model performance. This was also tested in the RETAIN model, which uses a reverse time attention mechanism to mimic physician practice by giving greater weight to more recent visits [2]. Thus, rather than only consider the first ICU stay for our deep learning model; we could consider the full sequence of

visits. However, the drawback of doing this is that it is less clinically useful because clinicians want to be able to predict the effects of the sedatives early in the ICU stay before the later events in subsequent visits occur. Nevertheless, testing this method would provide insights into the effects of the sedatives at later periods and may confirm the effectiveness of including temporal information. Another method we found in the literature that we could have used was representing all the EHR data in a standardized format and allowing the model to learn machine-derived features from it. Researchers in [44] applied this approach by representing patients as a temporal sequence of events using the FHIR specifications without discarding potentially useful information. Thus, rather than relying on domain experts to identify useful features, they included the entire sequence of events and allowed the model to weigh their importance using an attention mechanism. This resulted in greater interpretability because domain experts could identify the importance of each event. It also improved the generalizability of the model because they represented the data in a standardized format in accordance with the FHIR specifications, though this would increase the complexity of the model and possibly hinder its scalability. Another strategy we would consider is using different loss functions to account for different costs of misclassification errors to more accurately represent the true costs associated with misclassification. Finally, we would have built models for other outcomes of interest including length of hospital stay and ventilation days to provide more insights into the effect of the sedatives on these outcomes.

4.2 Summary

Our research provides insights for clinicians and data science researchers by investigating the effects of sedatives on patients with TBI and evaluating the strengths and weaknesses of deep learning and traditional modeling approaches. We used TBI patient data from ICUs across the U.S. to investigate the

effect of sedative exposures (fentanyl, propofol, versed, ativan, and precedex) with respect to hospital discharge disposition. Our results show that versed was associated with better discharge dispositions while fentanyl/ativan was associated with worse discharge dispositions. Other significant covariates included Northeast region, age APACHE, GCS, and ethanol level. However, future research will provide greater insights into our research question by testing our results and alternative strategies from previous work and examining other outcomes and exposures. Thus, we will continue to investigate our research question in order to provide guidance for sedative choice and increase our understanding of the strengths and weaknesses of deep learning and traditional modeling approaches for healthcare analytics.

4.3 Future Research

Future research is necessary to validate our results, investigate other outcomes and exposures of interest, and test alternative strategies such as those we learned about in previous work. Additionally, external validation using another data set would be useful for further evaluating the scalability and generalizability of the models.

4.3.1 Investigate the Effects of Versed and Fentanyl/Ativan

To validate our results, future research should investigate the effects of versed and fentanyl/ativan. Since we found that versed was associated with better discharge dispositions, future research is necessary to determine if it truly causes better discharge dispositions. Other possible explanations for this association are chance and the fact that clinicians sometimes use it for alternative purposes such as to still the patient for CT and MRI scans. Since we found that fentanyl/ativan was associated with worse discharge dispositions, future research is necessary to investigate the possible negative effects of ativan. We

suspect that ativan may lead to worse discharge dispositions because the other dual therapies with fentanyl did not significantly influence patient outcomes.

4.3.2 Examine Additional Outcomes and Exposures

In addition to testing our findings, future research should also examine other outcomes and exposures of interest. For example, investigating outcomes such as length of stay and ventilation days would provide a more comprehensive overview of the effects of each exposure and additional guidance into sedative choice. Similarly, investigating other exposures such as the ones we dropped due to insufficient data (ketamine, pentobarbital, and alprazolam) would provide valuable additions to this guidance. In addition to other exposures, the dosage is worth considering since previous work suggests that it may be more important than sedative choice [3].

4.3.3 Test Other Modeling Approaches

Future research is also necessary to incorporate other modeling approaches and techniques from previous research on our research question. Specifically, these techniques include using LSTMs to capture temporal sequential information, using all EHR data points rather than rely on experts to identify a subset of features, and addressing the class imbalance problem using an alternative loss function to represent the true cost of misclassification.

References

- [1] Massachusetts Institute of Technology Critical Data, Secondary Analysis of Electronic Health Records, Cambridge, MA, United States: Springer Nature, 2016.
- [2] E. Choi et al. "RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism," *Neural Information Processing Systems*, 2016.
- [3] O. Flower and S. Hellings, "Sedation in Traumatic Brain Injury," *Emergency Medicine International*, 2012.
- [4] Brain Trauma Foundation, "Guidelines for the Management of Severe Traumatic Brain Injury," 2016.
- [5] Centers for Disease Control and Prevention, "Surveillance Report of Traumatic Brain Injury-related Emergency Department Visits, Hospitalizations, and Deaths," 2014.
- [6] V. Buhmester et al. "Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey," *arXiv:1911.12116 [cs.AI]*, 2019.
- [7] S. Vashishth et al. "Attention Interpretability Across NLP Tasks," *arXiv:1909.11218 [cs.CL]*, 2019.
- [8] J. Kukacka et al. "Regularization for Deep Learning: A Taxonomy," *arXiv:1710.10686 [cs.LG]*, 2017.
- [9] Y. Lou et al. "Intelligible Models for Classification and Regression," *ACM SIGKDD International Conference Knowledge Discovery and Data Mining*.
- [10] Y. Lou et al. "Accurate intelligible Models with Pairwise Interactions," *ACM SIGKDD International Conference Knowledge Discovery and Data Mining*, 2013.
- [11] Centers for Disease Control and Prevention, "Supplemental Content CDC ICD-9-CM TBI Surveillance Codes".
- [12] T. J. Pollard et al. "Data Descriptor: The eICU Collaborative Research Database, a freely available multi-center database for critical care research.," *Scientific Data*, 2018.
- [13] MIT Laboratory for Computation Physiology, "eICU Collaborative Research Database," MIT, [Online]. Available: <https://eicu-crd.mit.edu/>.
- [14] A. Johnson et al. "MIMIC-III, a freely accessible critical care database," *Scientific Data*, 2016.
- [15] MIT Lab for Computation Physiology, "MIMIC," MIT, [Online]. Available: <https://mimic.physionet.org/>.
- [16] J. E. Zimmerman et al. "Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients.," *Society of Critical Medicine and Lippincott Williams & Wilkins*, 2016.
- [17] Knaus et al. "APACHE II: a severity of disease classification system," *Critical Care Medicine*, 1985.
- [18] Knaus et al. "The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults," *Chest*, 1991.
- [19] "PostgreSQL: The World's Most Advanced Open Source Relational Database," PostgreSQL Global Development Group, [Online]. Available: <https://www.postgresql.org/>.
- [20] OMOP, "OMOP Common Data Model," [Online]. Available: <https://www.ohdsi.org/data-standardization/the-common-data-model/>.

- [21] "Linear Models," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares.
- [22] M. Wahi, Composer, *Healthcare Analytics: Regression in R*. [Sound Recording]. LinkedIn. 2017.
- [23] "The R Project for Statistical Computing," GNU, [Online]. Available: <https://www.r-project.org/>.
- [24] "RStudio," RStudio Inc., [Online]. Available: <https://rstudio.com/>.
- [25] "Neural network models (supervised)," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/neural_networks_supervised.html#multi-layer-perceptron.
- [26] N. Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from," *Journal of Machine Learning Research*, 2014.
- [27] "Tuning the hyper-parameters of an estimator," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/grid_search.html#randomized-parameter-optimization.
- [28] X. Glorot et al. "Deep Sparse Rectifier Neural Networks," *NIPS*, 2001.
- [29] D. Kingma et al. "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs.LG]*, 2014.
- [30] "Keras: The Python Deep Learning library," [Online]. Available: <https://keras.io/>.
- [31] Python Software Foundation, [Online]. Available: <https://www.python.org/>.
- [32] "Randomized Search CV," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html.
- [33] "Linear Models: Lasso," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/linear_model.html#lasso.
- [34] "Linear Models: Ridge regression and classification," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression-and-classification.
- [35] H. Zou et al. "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society*, 2005.
- [36] "Linear Models: Elastic-Net," Scikit-learn. [Online].
- [37] "Lasso," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html.
- [38] "Ridge," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html.
- [39] "Elastic Net," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html.
- [40] Y. Lou et al. "MLTK: Machine Learning Tool Kit," [Online]. Available: <http://www.cs.cornell.edu/~yinlou/projects/mltk/>.
- [41] "Java Software," Sun Microsystems, [Online]. Available: <https://www.oracle.com/java/>.
- [42] "Cross-validation: evaluating estimator performance," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html.
- [43] A. Ashfaq et al. "Readmission prediction using deep learning on electronic health records," *Journal of Biomedical Informatics*, 2019.
- [44] A. Rajkomar et al. "Scalable and accurate deep learning with electronic health records," *Nature Partner Journals*, 2018.

Supplementary Materials

Feature Description Table

The feature description table below groups features into categories and maps them to a detailed description to provide a comprehensive overview of all the features we considered. The categories are identifiers, exposures of interest (sedatives), outcomes of interest, demographics, and clinical state variables. Each category has a set of features and a corresponding description that includes its precise definition and data type.

Feature	Description
Identifiers	
Unique PID	The unique patient identifier.
Patient unit stay id	The unique ICU stay identifier.
Patient health system stay id	The unique hospital visit identifier
Exposures of Interest*	
Fentanyl	Binary flags indicating exposure to fentanyl.
Propofol	Binary flags indicating exposure to propofol.
Versed	Binary flags indicating exposure to versed.
Ketamine	Binary flags indicating exposure to ketamine.
Ativan	Binary flags indicating exposure to ativan.
Pentobarbital	Binary flags indicating exposure to pentobarbital.
Precedex	Binary flags indicating exposure to precedex.
Alprozolam	Binary flags indicating exposure to alprozolam.
Sedated	Binary flags indicating exposure to any of the above sedatives.
Monotherapy**	
Fentanyl	Binary flag indicating exposure to only fentanyl within one week of the ICU admission time.
Propofol	Binary flag indicating exposure to only propofol within one week of the ICU admission time.
Versed	Binary flag indicating exposure to only versed within one week of the ICU admission time.
Ativan	Binary flag indicating exposure to only ativan within one week of the ICU admission time.
Precedex	Binary flag indicating exposure to only precedex within one week of the ICU admission time.
Monotherapy	Categorical string indicating which of the above monotherapies the patient received.
Dual Therapy**	
Fentanyl or Propofol	Binary flag indicating exposure to only fentanyl or propofol within one week of the ICU admission time.
Fentanyl or Versed	Binary flag indicating exposure to only fentanyl or versed within one week of the ICU admission time.
Fentanyl or Ativan	Binary flag indicating exposure to only fentanyl or ativan within one week of the ICU admission time.
Fentanyl or Precedex	Binary flag indicating exposure to only fentanyl or precedex within one week of the ICU admission time.
Dual Therapy	Categorical string indicating which of the above dual therapies the patient received.
Triple Therapy**	
Fentanyl, propofol, or versed	Binary flag indicating exposure to only fentanyl, propofol, or versed within one week of the ICU admission time.
Fentanyl, propofol, or ativan	Binary flag indicating exposure to only fentanyl, propofol, or ativan within one week of the ICU

	admission time.
Fentanyl, premedex, or versed	Binary flag indicating exposure to only fentanyl, premedex, or versed within one week of the ICU admission time.
Fentanyl, premedex, or ativan	Binary flag indicating exposure to only fentanyl, premedex, or ativan within one week of the ICU admission time.
Triple Therapy	Categorical string indicating which of the above triple therapies the patient received.
Outcomes of Interest	
Ranked Discharge Location	Numeric value ranking the discharge location from worst to best. The discharge locations from worst to best are death, remains hospitalized, nursing home, rehabilitation, and home. There is a ranked discharge location feature provided at both the ICU and hospital levels.
Categorical Ranked Discharge Location	Categorical string indicating the discharge location (death, remains hospitalized, nursing home, rehabilitation, or home).
Ventilation Days	Numeric value indicating the number of days the patient spent on ventilation. There is a ventilation days feature provided at both the ICU and hospital levels.
Expired	Binary flag indicating whether the patient expired (died) at the time of discharge. There is an expired feature provided at both the ICU and hospital levels.
Length of Stay	Numeric value indicating the number of days the patient was in the ICU or hospital. There is a length of stay feature provided at both the ICU and hospital levels.
Demographics	
Age	Numeric value indicating the age of the patient. If the patient is over 89 years old, this will have a value of 90.
Gender	Categorical string indicating the patient's gender (male, female, or other/unknown)
Ethnicity	Categorical string indicating the patient's ethnicity (African American, Hispanic, Native American, Asian, Caucasian, or other/unknown)
Region	Categorical string indicating the region of the hospital (South, West, Midwest, or Northeast).
Clinical	
TBI	Binary flag indicating whether the patient received a TBI diagnoses. An additional flag feature indicates whether the TBI diagnoses is primary.
GCS	The Glasgow Coma Scale score for the patient using the worst values from the APACHE day (the first day in the ICU). There is a separate feature for motor, eyes, verbal, and total.
APS Score	The Acute Physiology Score of the patient used to calculate the APACHE score.
APACHE Score	The APACHE IVa Score of the patient used to control for severity of illness.
Intubated	Binary flag indicating the intubated status of the patient during the APACHE day (the first day in the ICU).
Alcohol	Binary flag indicating whether the patient received an alcohol-related diagnosis. An additional flag feature indicates whether this diagnosis is primary.
Ethanol Intoxication	Binary flag indicating whether the patient had an ethanol level greater than 80 mg/dL within 12 hours of the hospital admission time.
Ventilated	Binary flag indicating whether the patient received ventilation. There is a ventilated feature provided at both the ICU and hospital levels.

*** Each exposure of interest has three binary flags. The first indicates whether it the patient received it during the ICU stay. The second indicates whether he or she received it during the hospital visit. The third indicates whether he or she received it within one week of the ICU admission time.**

**** Each mono, dual, and triple therapy only uses a single binary flag for each sedative or sedative group to indicate whether he or she received *only one* or more of them within one week of the ICU admission time and not any of the other sedatives. This is to avoid confounding.**

Summary Statistics Table

The summary statistics table below groups features into categories and shows summary statistics for each feature that allowed us to perform preliminary exploratory data analysis. The categories are ICU level exposures of interest (sedatives), monotherapy, dual therapy, triple therapy, outcomes of interest, demographics, and clinical state. Each category contains a set of features. For a detailed description and precise definition of each feature, see the feature description table above. See the table description below for a detailed description for further guidance on how to interpret the results.

Feature	Survived	Died	Unknown	Total
ICU Level Exposures of Interest				
Fentanyl	827 (82.4%)	164 (16.3%)	12 (1.1%)	1003 (34.8%)
Propofol	548 (79.3%)	131 (18.9%)	12 (1.7%)	691 (24%)
Versed	335 (79.7%)	79 (18.8%)	6 (1.4%)	420 (14.5%)
Ketamine	4 (66.6%)	2 (33.3%)	0	6 (0.2%)
Ativan	534 (79.4%)	128 (19%)	10 (1.4%)	627 (23.3%)
Pentobarbital	0	0	0	0
Precedex	135 (86.5%)	17 (10.8%)	4 (2.5%)	156 (5.4%)
Alprozolam	58 (95%)	2 (3.2%)	1 (1.6%)	61 (2.1%)
Sedated	1223 (82.9%)	235 (15.9%)	16 (1%)	1474 (51.72%)
Monotherapy				
Fentanyl	22 (62.8%)	13 (37.1%)	0	35 (1.2%)
Propofol	157 (80.5%)	35 (17.9%)	3 (1.5%)	195 (6.7%)
Versed	8 (80%)	2 (20%)	0	10 (0.3%)
Ativan	1 (50%)	1 (50%)	0	2 (0%)
Precedex	41 (93.1%)	2 (4.5%)	1 (2.2%)	44 (1.5%)
Dual Therapy				
Fentanyl or Propofol	280 (78.2%)	73 (20.3%)	5 (1.3%)	358 (12.4%)
Fentanyl or Versed	40 (68.9%)	18 (31%)	0	58 (2%)
Fentanyl or Ativan	24 (58.5%)	17 (41.4%)	0	41 (1.4%)
Fentanyl or Precedex	63 (79.7%)	15 (18.9%)	1 (1.2%)	79 (2.7%)
Triple Therapy				
Fentanyl, propofol, or versed	321 (77.3%)	89 (21.4%)	5 (1.2%)	415 (14.4%)
Fentanyl, propofol, or ativan	289 (77.4%)	79 (21.1%)	5 (1.3%)	373 (12.9%)
Fentanyl, precedex, or versed	83 (79%)	21 (20%)	1 (0%)	105 (3.6%)
Fentanyl, precedex, or ativan	65 (76.4%)	19 (22.3%)	1 (1%)	85 (2.9%)
Outcomes of Interest				
Hospital Length of Stay (days)				
N (Nmiss)	2493 (1)	354 (0)	29 (0)	2876 (1)
Mean ± SD	8 ± 9.1	5.1 ± 6.1	10.8 ± 9.1	7.6 ± 8.9
Min - Max	0.1 – 105.7	0 – 39.5	0 – 33.2	0 – 105.7
Median (IQR)	4.9 (2.6 – 9.7)	2.6 (1.1 – 7)	7.9 (3.9 – 16.7)	4.7 (2.3 – 9.4)
Skewness	3.15	2.29	1.04	3.14
Kurtosis	15.36	6.31	0.03	15.54

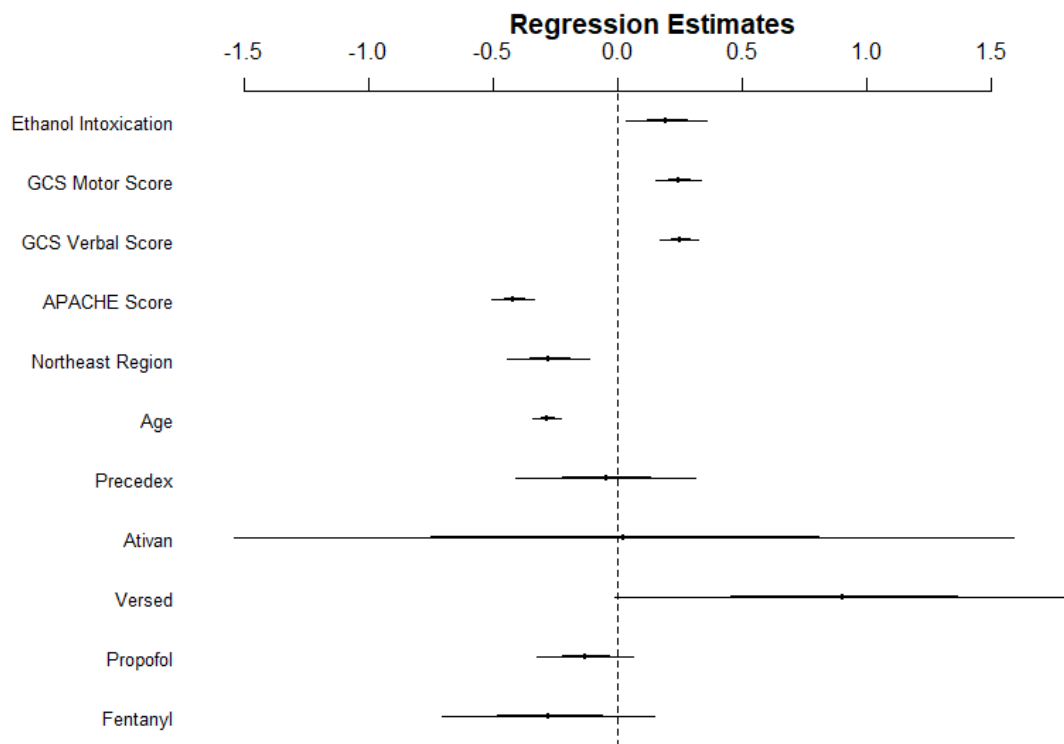
First ICU Stay Length of Stay (days)				
N (Nmiss)	2494 (0)	354 (0)	29 (0)	2877 (0)
Mean ± SD	3.6 ± 5.21	3.8 ± 4.8	6.8 ± 7.4	3.7 ± 5.2
Min - Max	0 – 58.7	0 – 32.4	0.3 – 27.2	0 – 58.7
Median (IQR)	1.7 (0.9 – 3.7)	1.9 (0.8 – 4.9)	3.2 (1.3 – 10.8)	1.7 (0.9 – 3.8)
Skewness	3.46	2.34	1.36	3.30
Kurtosis	16.81	6.83	1.02	15.41
First ICU Stay Ventilation Days				
N (Nmiss)	2217 (277)	315 (39)	0 (29)	2532 (345)
Mean ± SD	1.5 ± 4.3	3 ± 3.7	N/A	1.7 ± 4.2
Min - Max	0 – 43	0 – 22	N/A	0 – 43
Median (IQR)	0 (0 – 1)	2 (1 – 4)	N/A	0 (0 – 1)
Skewness	3.96	2.21	N/A	3.72
Kurtosis	18.55	5.79	N/A	16.86
Hospital Ventilation Days				
N (Nmiss)	2217 (277)	315 (39)	0 (29)	2532 (345)
Mean ± SD	3.2 ± 8.7	6.2 ± 7.4	N/A	3.6 ± 8.6
Min - Max	0 – 86	0 – 44	N/A	0 – 86
Median (IQR)	0 (0 – 2)	4 (2 – 8)	N/A	0 (0 – 2)
Skewness	3.91	2.19	N/A	3.67
Kurtosis	18.02	5.74	N/A	16.42
Demographics				
Age				
N (Nmiss)	2494 (0)	354 (0)	29 (0)	2877 (0)
Mean ± SD	60.4 ± 21.9	64.3 ± 21.7	57.2 ± 24.9	60.9 ± 22
Min - Max	11 – 90+	14 – 90+	16 – 90+	11 – 90+
Median (IQR)	64 (43.2 – 79)	70 (52 – 82)	63 (38 – 78)	65 (45 – 80)
Skewness	-0.46	-0.75	-0.35	-0.49
Kurtosis	-0.97	-0.49	-1.23	-0.93
Gender				
Male	1487 (85.9%)	227 (13.1%)	17 (0%)	1731 (60.1%)
Female	1003 (88.2%)	122 (10.7%)	12 (1%)	1137 (39.5%)
Other/Unknown	4 (44.4%)	5 (55.5%)	0	9 (0.3%)
Ethnicity				
African American	169 (85.3%)	24 (12.1%)	5 (2.5%)	198 (6.8%)
Hispanic	114 (87.6%)	15 (11.5%)	1 (0%)	130 (4.5%)
Native American	23 (100%)	0	0	23 (0.7%)
Asian	52 (82.5%)	9 (14.2%)	2 (3.1%)	63 (2.1%)
Caucasian	1974 (86.8%)	277 (12.1%)	21 (0%)	2272 (78.9%)
Other/Unknown	162 (84.8%)	29 (15.1%)	0	191 (6.6%)
Region				
South	615 (84.8%)	106 (14.6%)	4 (0%)	725 (25.1%)
West	623 (88.8%)	68 (9.7%)	10 (1%)	701 (24.3%)
Midwest	862 (87.8%)	107 (10.9%)	12 (1%)	981 (34%)
Northeast	227 (79.3%)	59 (20.6%)	0	286 (9.9%)
Clinical Features				
TBI Primary	1628 (86.6%)	231 (12.2%)	20 (1%)	1879 (65.3%)
GCS				
N (Nmiss)	2404 (90)	348 (6)	27 (2)	2779 (98)
Mean ± SD	12.5 ± 3.6	6.3 ± 4.3	10.8 ± 4.4	11.7 ± 4.2
Min - Max	3 – 15	3 – 15	3 – 15	3 – 15
Median (IQR)	14 (12 – 15)	4 (3 – 9)	13 (7.5 – 15)	14 (9 – 15)
Skewness	-1.48	0.93	-0.62	-1.08
Kurtosis	0.94	-0.69	-1.05	-0.37
APACHE Score				
N (Nmiss)	2203 (291)	312 (42)	0 (29)	2515 (362)
Mean ± SD	47.4 ± 20.4	86.3 ± 27.7	N/A	52.2 ± 25.0
Min - Max	0 – 139	25 – 160	N/A	0 – 160

Median (IQR)	44 (33 – 58)	87 (65 – 1060)	N/A	47 (34 – 64)
Skewness	0.94	0.08	N/A	1.05
Kurtosis	1.14	-0.61	N/A	1.04
Intubated	339 (69.3%)	146 (29.8%)	4 (0.8%)	489 (16.9%)
Alcohol	121 (92.3%)	9 (6.8%)	1 (0.7%)	131 (4.5%)
Ethanol Intoxication	233 (88.9%)	28 (10.6%)	1 (0%)	262 (9.1%)
Ventilated	573 (69.8%)	247 (30.1%)	0	820 (28.5%)
Total	2494 (86.6%)	354 (12.3%)	29 (1.0%)	2877 (100%)

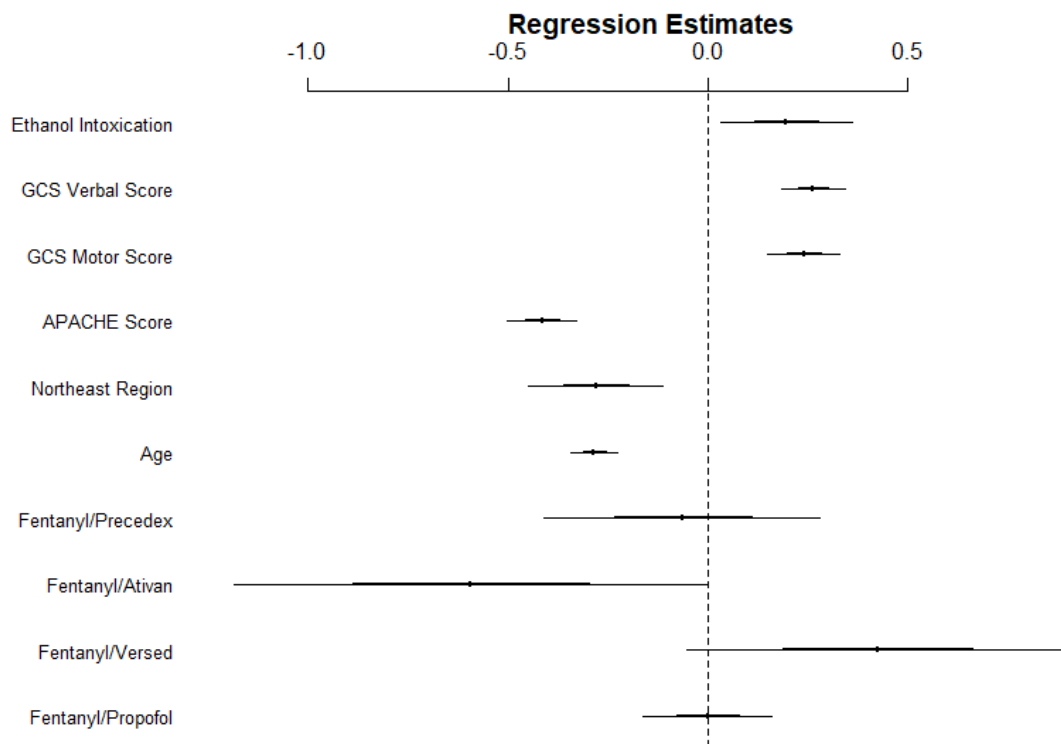
- We report all values by considering only the first ICU stay unless otherwise stated. For example, the flags indicating whether a patient received a sedative is set if it the patient received at some point during the first ICU stay. If it the patient received at some point after that during the hospital stay, then the flag will not be set. Similarly for the TBI and alcohol diagnoses, apache score, GCS score, and intubated status.
- For binary flags, the number in the survived died, and unknown columns indicate the number of patients.
- For continuous and numeric variables, the number indicates the specified value.
 - N: Number of patients who had the variable recorded.
 - Nmiss: Number of patients who did not have the variable recorded
 - SD: Sample population standard deviation
 - IQR: Interquartile range

Coefficient Plots

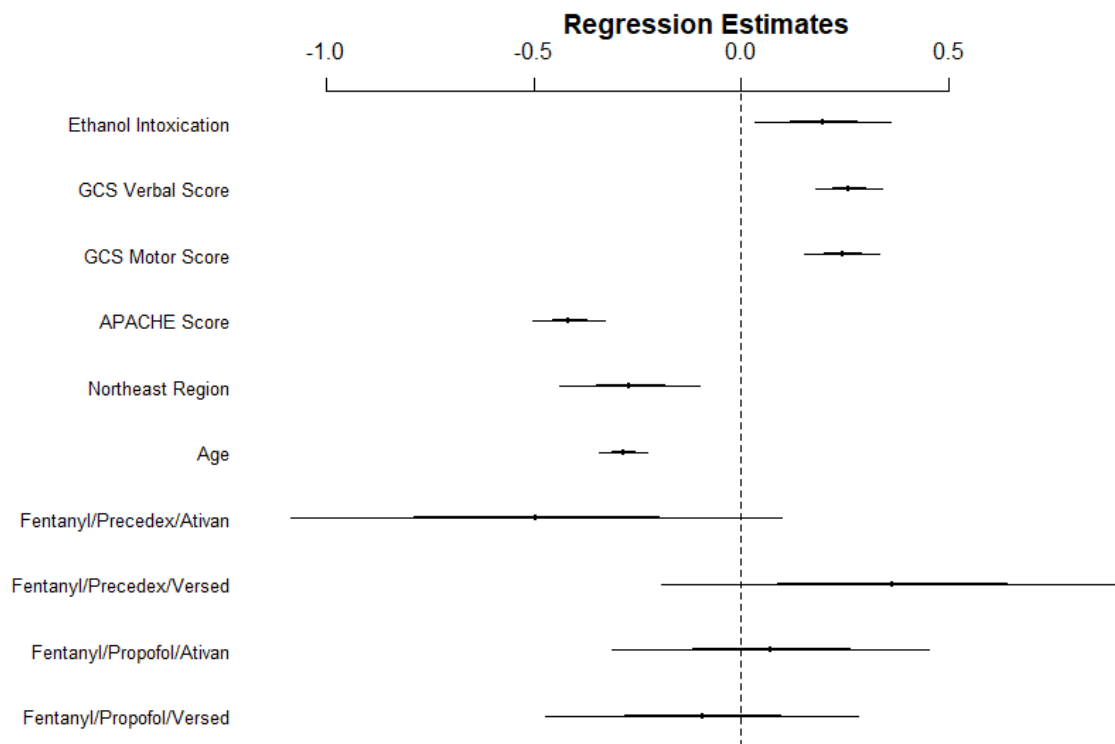
Below are the coefficient plots for the linear regression models for each experiment.



Coefficient plot for the monotherapy experiment showing the coefficient and 95% confidence interval for each variable.



Coefficient plot for the dual therapy experiment showing the coefficient and 95% confidence interval for each variable.



Coefficient plot for the triple therapy experiment showing the coefficient and 95% confidence interval for each variable.

GA²M Results

RMSE: 1.08

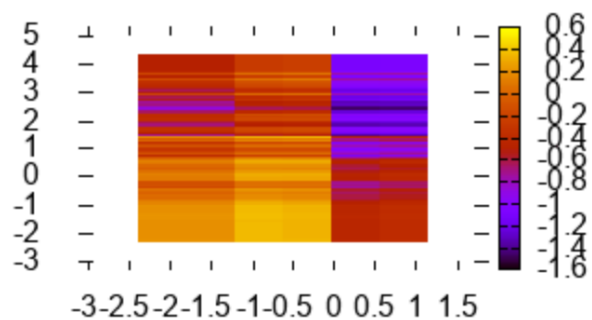
Training Time: 166.764s

The plots below show the strongest pairwise interactions and their residual sum of squares (RSS).

Stronger pairwise interactions have a lower RSS.

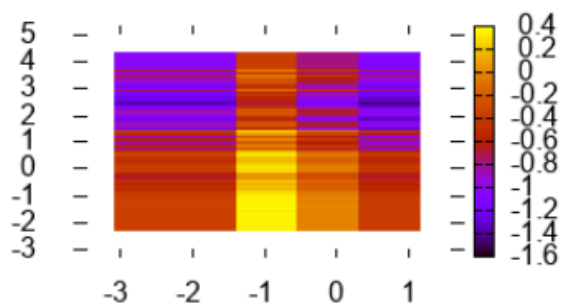
APACHE vs. GCS Verbal

RSS: 2479.4836



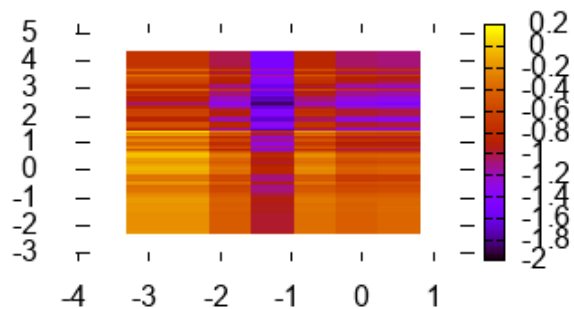
APACHE vs. GCS Eyes

RSS: 2510.9903

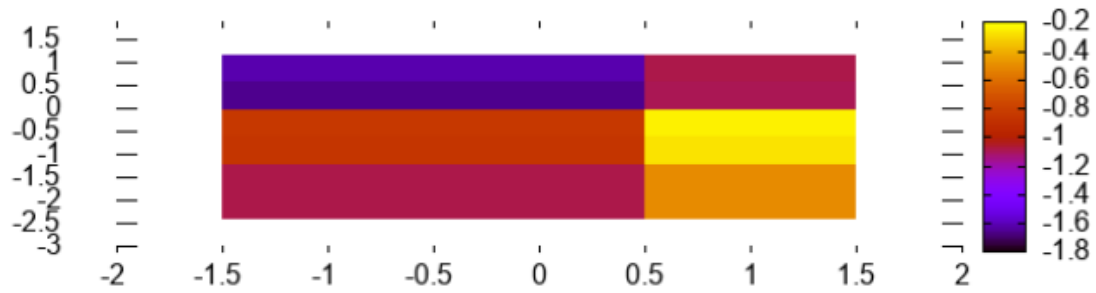


APACHE vs. GCS Motor

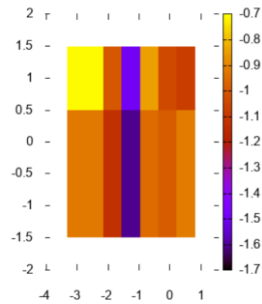
RSS: 2530.4340



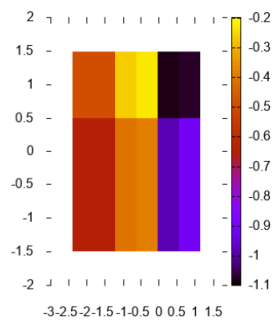
GCS Verbal vs. Ethanol Intoxication
RSS: 2770.2428



Propofol vs. GCS Motor
RSS: 2973.4091

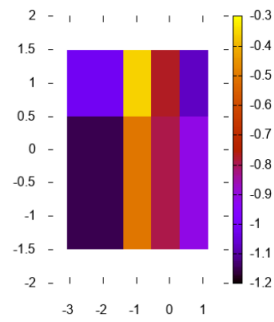


Propofol vs. GCS Verbal
RSS: 2988.104



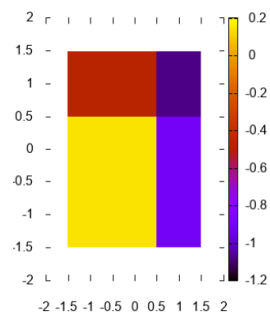
Propofol vs. GCS Eyes

RSS: 2989.8947



Propofol vs. Ventilated

RSS: 3005.7438



Propofol vs. APACHE

RSS: 3008.5873

