Project Team 4 Project Proposal

* Brinda Asuri (bva246), Grayson Merritt (gkm457), Pranav Garg (pg26855), Soham Siddhant Bidyadhar (sb67347), Thomas Starkie (ts43286)

## **Problem:**

### Predicting in-hospital mortality in the ICU is a critical task that can significantly impact patient outcomes, resource allocation, and overall healthcare quality. Early identification of at-risk patients allows for proactive interventions and potentially life-saving decisions. Despite advances in critical care, the number of inpatient deaths remains high. According to data from the National Hospital Discharge Survey (NHDS), over 700,000 patients died in U.S. hospitals in 2010 alone​. This underscores the need for effective predictive tools to better identify patients at risk of mortality.

### Research has shown that patients who die during hospitalization have longer average stays compared to all patients​, suggesting that timely prediction and intervention could reduce length of stay and improve outcomes. Furthermore, significant trends in mortality have been observed across different age groups, with a higher percentage of deaths occurring in older patients, particularly those over 85​. The increasing mortality rates for conditions such as septicemia, which saw a 17% increase in inpatient deaths from 2000 to 2010, further highlight the importance of accurate mortality prediction models​.

### Existing prediction models such as APACHE (Acute Physiology and Chronic Health Evaluation) have been widely used in critical care settings but often fail to generalize across different healthcare systems and geographies. Differences in ICU protocols, resource availability, patient demographics, and local healthcare practices all introduce significant variability. These factors make developing a globally generalizable model for predicting mortality in ICU patients a complex challenge.

### The goal of this project is to use the GOSSIS dataset to build a machine learning model that predicts ICU mortality. The model will focus on generalizing across diverse healthcare settings by utilizing a range of clinical and demographic predictors to identify key factors that correlate with mortality risk. The primary objective is to provide clinicians with actionable insights that can improve patient care in ICUs globally. For example, clinicians will better understand how to identify high risk patients much faster, optimize various resources within the ICU, better personalize treatment plans, and redfine triage protocols with our insights from predictive modeling. On a further note, while GOSISS also has members from Argentina, India, Nepal, Sri Lanka, and Brazil, the only dataset available to us contains US and New Zealand/Australia data.

## **Data:**

The dataset originates from the Global Open Source Severity of Illness Score (GOSSIS) consortium, which pools critical care data from over 380,000 patients admitted to 366 hospitals across Australia, New Zealand, and the United States. The GOSSIS consortium was created to develop a standardized severity of illness score that could generalize across healthcare systems, aiming to overcome calibration and discrimination issues found in more localized, US-centric models. Data was collected from the Australia New Zealand Intensive Care Society (ANZICS) and eICU Collaborative Research Databases, covering 225 hospitals in AU/NZ and 131 hospitals in the US. Patients included were over 16 years old with at least 6 hours of ICU stay and this was their first ICU stay. The data included all ICU admissions discharged in 2014-2015 and the data was collected in the first 24 hours. Our dataset from Kaggle has around 91,000 rows.

The dataset contains a comprehensive list of predictors that represent areas such as patient demographics, vital signs, and clinical history. Some key variables include demographic Info such as Age, BMI, ethnicity, gender, height, and weight. There are some ICU-related details such as ICU admission source, ICU stay type, ICU type, and pre-ICU length of stay (LOS). Additionally, there are patient scores such as APACHE II and APACHE IIIj diagnosis, post-operative status, Glasgow Coma Scale (GCS) scores (e.g., GCS eyes, motor, verbal). Vital signs such as heart rate, blood pressure (systolic/diastolic), respiratory rate, temperature were also measured. Chronic conditions such as diabetes, cirrhosis, immunosuppression, leukemia, solid tumors, and AIDS were taken into account. And finally, the target variable for the model will be ICU mortality, I.E. whether the patient died during their ICU stay.

There are a couple of other things to note about this dataset. The first is that the dataset covers patients from hospitals with differing levels of care and resource availability, making it important to capture the variability in ICU practices. This is a factor we must keep in mind as we build our models. Additionally, given the complexity of capturing ICU records and healthcare in general, there are some missing values in the dataset. Imputation strategies will be employed to handle these gaps, using both statistical methods and any clinical knowledge we research. Things such as APACHE Scores are unfamiliar to our group, so we will have to do some research to understand more of our feature variables in depth. Also of high importance, ICU mortality tends to be lower than survival rates, meaning the dataset is quite imbalanced (~80,000 survived vs ~7,000 passed away). Proper handling of this imbalance is crucial for us to build an accurate and unbiased model. Finally, we would like to add some sort of geolocation data to even further develop what factors can affect patient mortality. However, the dataset appears to not include any location data, only having hospital and ICU id’s. This is probably due to privacy laws like HIPAA, and will affect if we can use any geolocation data.

## **Potential Approaches:**

To address the prediction of ICU mortality, we plan to implement several machine learning models, including logistic regression, random forest, XGBoost, k-nearest neighbors, and a neural network. Logistic regression will help provide insights into feature significance, while random forest and XGBoost can handle non-linear relationships and large datasets effectively. KNN will help us explore distance-based and high-dimensional feature spaces (PCA can also be tested here), and the neural network will capture the more complex patterns we predict we will find.

Given the likely class imbalance in the dataset, we will apply oversampling techniques (we have used SMOTE before but will look into others) to ensure balanced training and scaling of features to normalize variables such as heart rate and blood pressure. This will prevent models like logistic regression from being skewed by differently scaled variables.

Feature engineering will be done to try and uncover more insights. We will also explore interaction terms between vital signs and demographic features to capture more complex relationships that influence patient outcomes.

For validation, we will use stratified k-fold cross-validation to ensure our models perform well. Evaluation metrics will include AUC-ROC, precision, recall, accuracy, and other confusion matrix metrics to assess the model’s ability to distinguish between mortality and survival for patients.

**References:**

<https://www.kaggle.com/datasets/mitishaagarwal/patient/data>

<https://pmc.ncbi.nlm.nih.gov/articles/PMC9233021/>

<https://gossis.mit.edu/>

<https://pubmed.ncbi.nlm.nih.gov/35354159/>

<https://www.cdc.gov/nchs/products/databriefs/db118.htm#:~:text=Key%20findings,-Data%20from%20the&text=The%20number%20of%20inpatient%20hospital,patients%20aged%2085%20and%20over>

Note: PDF associated with the CDC webpage above   
  
**After Completion:**   
[Medium Blog Link](https://medium.com/@pg.garg.pranav/predicting-icu-mortality-using-machine-learning-a-global-perspective-dc18d009afb0)