**Capstone Project Proposal**

**Medical AI Research**

Jenny Tsai & Jichong Wu

February 7, 2024

* **OBJECTIVES:**

Predicting Postoperative Blood Transfusions for Coronary Artery Bypass Graft Patient

* **DATA SOURCE:**

The data was downloaded from Participant Use Data File (PUF) on the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP). In this project, we are looking at the data from 2018 to 2020, which has a total of 4953 observations and at least 274 features across three datasets.

| **Year** | **# of rows** | **# of columns** |
| --- | --- | --- |
| 2020 | 1493 | 276 |
| 2019 | 1639 | 274 |
| 2018 | 1821 | 274 |
| **TOTAL** | **4953** | **274** |

Among the 274 features, 20 features pertaining to the current project will be entered to our models for further analysis.

* 1. Sex: Male, female
  2. Race: White, black, other
  3. Body mass index (BMI) (using height and weight)
  4. INOUT: inpatient, outpatient
  5. Age
  6. ANESTHES: general, regional, other
  7. DIABETES
  8. SMOKE
  9. DYSPNEA
  10. FNSTATUS2: Functional health status prior to surgery
  11. HXCOPD: History of severe COPD
  12. ASCITES
  13. HXCHF: Heart failure in 30 days before surgery
  14. HYPERMED: Hypertension requiring medication
  15. DIALYSIS
  16. DISCANCR: Disseminated cancer
  17. STEROID: Immune suppressive therapy
  18. WTLOSS: Malnourishment
  19. BLEEDIS: Bleeding disorders
  20. TRANFUS: Preop Transfusion of >= 1 unit of whole/packed RBCs in 72 hours prior to surgery

The target variable of this project is *OTHBLEED,* predicting whether or not the patient needs blood transfusion after surgery (binary). Target can be further categorized into intraoperative vs. postoperative vs. no transfusion, therefore can be transformed into a 3-class variable when needed.

* **APPROACH:**
  + **Exploratory Data Analysis**
    - Histogram, box plots, and bar plots to examine basic insights and trends
    - QQ plots and correlation matrix to examine correlation relationship between features
  + **Feature Selection/Engineering**

We are looking to try out various methods (e.g., chi-square test, decision trees, feature selection packages) and use different subset of selected features for modeling. Features can be selected from the pre-select set of 20 features and from the whole datasets (274 features).

* + **Modeling**

We are looking to try out at least three classification models to see which one(s) performs the best. From simple to complex, we will include logistic regression (traditional statistics method), Multi-Layer Perceptron (MLP, shallow neural network), and Fully Connected Neural Network (FNN) in this project. Other classification models can be also considered including decision tree, random forest, SVM, and KNN. As for the performance index, we will look at accuracy and f1 score jointly as the target variable is not imbalanced.

* **TIMELINE:**

| **Week** | **Progress** |
| --- | --- |
| 1/23 | Choose research project, write analysis plan & timeline, get dataset |
| 1/30 | Data preprocessing & EDA, set up github repo |
| 2/6 | Feature selection & Logistic regression |
| 2/13 | MLP |
| 2/20 | FNN |
| 2/27 | Discuss the preliminary results |
| 3/5 | Discuss the preliminary results |
| 3/12 | Spring break – prepare for presentation |
| 3/19 | Preliminary presentation |
| 3/26 | Improve model based on feedback |
| 4/2 | Set up GUI for visualization |
| 4/9 | Write up paper |
| 4/16 | Wrap up paper and prepare for presentation |
| 4/23 | Final presentation and paper submission |