Capstone Project Proposal
Medical AI Research

Jenny Tsai, Jichong Wu, & Puneet Gupta

February 7, 2024

- **OBJECTIVES:**

Predicting Postoperative Blood Transfusions for Coronary Artery Bypass Graft Patient

Coronary Artery Bypass Graft (CABG) is a common cardiac surgery but continues to have many associated risks, including needing perioperative blood transfusions. Previous research has shown that blood transfusion during CABG surgery is associated with an increased risk for mortality after surgery. Specially, post-operative blood transfusion after CABG is associated with higher odds of readmission and heart failure within 30-days.

To lower the risk of mortality after surgery, there is a need to develop models that preoperatively predict which patients will need an intra-operative or post-operative blood transfusion. This will not only help to improve patient selection and patient education, but also physician preoperative awareness and perioperative guidelines for CABG patients. Therefore, the goal of this project is to explore different approaches and find the models that can best make predictions, including feature selection/engineering, classical statistical models, and neural networks.

- **DATASET:**

The data was downloaded from the 2020 Participant Use Data File (PUF) on the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP), provided by Dr. Gupta. In this project, we focus on the data from 2018 to 2020, which has a total of 4953 observations and 275 variables across three datasets.

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

Among the 275 features, 20 were identified to be most relevant to the current project (see the list below). The target variable is ***OTHBLEED*** (Occurrences Bleeding Transfusions), predicting whether or not the patient needs blood transfusion after surgery (binary variable). Target can be further categorized into intraoperative vs. postoperative vs. no transfusion, therefore can be transformed into a 3-class variable when needed.With different feature selection/engineering strategies, these features will be entered into our

models to predict the target variable, and we will compare the performance with each other as well as with the benchmarks from previous research.

<u>20 Pre-selected Features:</u>

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. TRANSFUS: Preop Transfusion of >= 1 unit of whole/packed RBCs in 72 hours prior to surgery

- **APPROACH:**
    - **Data preprocessing and Exploratory Data Analysis**
        - Performing data cleaning to combine the datasets, identify missing data, and impute missing data.
        - Histogram, box plots, and bar plots to examine basic insights and trends
        - QQ plots and correlation matrix to examine correlation relationship among features
    - **Build the baseline models and compare with the benchmark results (from Dr. Gupta)**
        - Using the clean dataset to run the following models and comparing the results with each other and with the benchmark from previous research
            - SVM
            - Decision tree

        - XGBoost
        - CatBoost
    - **Model improvements through feature selection / feature engineering**
        - Using various feature selection methods and feature engineering packages (e.g. *TPOT, featurewiz, featuretools, etc.*) to create different (sub)sets of features for modeling.
        - Features can be the pre-select set of 20 features or the ~275 features from the whole datasets
    - **Advanced modeling**
        - Using different sets of features and synthetic data generator to construct new models
        - Run the newly constructed datasets with CNN, Transformer, and other Deep Neural Networks to improve the results.

    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	Set up the meeting mechanism, develop modular functions and code pipelines for data preprocessing and baseline models
2/13	Conduct literature review
2/20	Feature selection/feature engineering
2/27	Model improvement with CNN, Transformer, Deep Neural Networks
3/5	Discuss the preliminary results
3/12	Spring break – prepare for presentation
3/19	Preliminary presentation
3/26	Improve model and the work based on feedback
4/2	Build UI/Dashboard for visualization / demo
4/9	Write up paper
4/16	Complete the paper and the poster, and prepare for presentation
4/23	Final presentation and paper submission