

Data Science Program

Capstone Report - Spring 2024

Predicting Blood Transfusions for Coronary Artery Bypass Graft Patient

Jenny Hsiao-Tien Tsai,

Jichong Wu,

Puneet Gupta

supervised by

Amir Jafari

Abstract

A strong abstract sums up your work in very few sentences: (i) state the problem you are addressing; (ii) say why it's an interesting problem, and which issues are hard to tackle; (iii) give your approach towards solving the problem; (iv) say Why and how well your approach solves the problem.

Contents

[1 Introduction 3](#_Toc91714543)

[2 Problem Statement 3](#_Toc91714544)

[3 Related Work 3](#_Toc91714545)

[4 Solution and Methodology 3](#_Toc91714546)

[5 Results and Discussion 4](#_Toc91714547)

[5.1 Experimentation protocol 5](#_Toc91714548)

[5.2 Data tables 5](#_Toc91714549)

[5.3 Graphs 5](#_Toc91714550)

[6 Discussion 6](#_Toc91714551)

[7 Conclusion 7](#_Toc91714552)

[8 Bibliography 7](#_Toc91714553)

# Introduction

Coronary Artery Bypass Graft (CABG) is a common cardiac surgery but continues to have many associated risks, including major bleeding which might need 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.

# Problem Statement

This project aims to explore various approaches that can best predict whether or not a CABG patient will need blood transfusions, including traditional statistical/tree-based models, feature selection, feature engineering, synthetic data generation, and deep neural networks.

# Related Work

Tschoellitsch et al. (2022) employed machine learning models to predict perioperative allogeneic blood transfusion for cardiac patients. The best model (Random Forest) showed good performance (RUC ranged from .76 - .86), however, the study has several limitations. For example, the data was from a single adult cardiac surgery center in Austria with a relatively small sample size (N = 3782), thus the results may not be generalizable to other samples with different demographics or nationalities. Moreover, the studies only predicted allogeneic blood transfusion (i.e., transfusion of more than 10 units of packed red blood cells (pRBC)), while blood transfusion regardless of volume has been associated with many known risks. Lastly, the study only tested the basic machine learning models (e.g., tree-based models), and it is likely that the performance can be significantly improved with more advanced techniques and neutral networks.

To address this research gap, the current project will use the national medical database in the U.S. with a large sample size of over 8,000 datapoints. Additionally, we will predict blood transfusion regardless of volume. Lastly, we will experiment with various approaches in order to optimize the performance, including feature selection, feature engineering, synthetic data generation, and deep neural networks.

# Solution and Methodology

*Data Source and Data Preprocessing*

The data was downloaded from the Participant Use Data File (PUF) on the American College of Surgeons National Surgical Quality Improvement Program ([ACS NSQIP](https://www.facs.org/quality-programs/data-and-registries/acs-nsqip/)). In this project, we focus on the data from 2018 to 2022, which has a total of 8,587 observations and around 294 variables across five datasets.

After data preprocessing, including imputation (mean for numeric variables and most frequent values for categorical variables), and standardization, and recoding, the dataset with 41 features identified as most relevant to the current study was served as our baseline data. 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.

# Results and Discussion

The results section details your metrics and experiments for the assessment of your solution. It

then provides experimental validation for your approach with visual aids such as data tables and



Figure 2: Try to guess what this gure illustrates; I double-dare you.

graphs. In particular, it allows you to compare your idea with other approaches you've tested,

for example solutions you've mentioned in your related work section.

## Experimentation protocol

It is of the utmost importance to describe how you came up with the measurements and results

that support your evaluation.

## Data tables

Every data table should be numbered, have a brief description as its title, and specify the units

used.

As an example, Table 1 compares the average latencies of native application calls to networked

services. The experiments were conducted on an Apple MacBook Air 2010 with a CPU speed of

1.4GHz and a bus speed of 800MHz. Each data point is a mean over 20 instances of each call,

after discarding both the lowest and the highest measurement.



## Graphs

Graphs are often the most important information in your report; you should design and plot

them with great care. A graph contains a lot of information in a short space. Graphs should be

numbered and have a title. Their axes should be labelled, with the quantities and units specified.

Make sure that individual data points (your measurements) stand out clearly. And of course,



Figure 3: Probability of including [k] faulty/malicious nodes in the service

always associate your graph with text that explains your results, and outlines the conclusions you

draw from these results.

For example, Figure 3 compares the efficiency of three different service architectures in eliminating adversarial behaviors. Every data point gives the probability that k faulty/malicious nodes managed to participate in a computation that involves 32 nodes. In the absence of at least one reliable node (k = 32), the failure will go undetected; but the results show that this case is extremely unlikely, regardless of the architecture. The most significant result pertains to k = 16: the reliable nodes detect the failure, but cannot reach a majority to recover. The graph shows that the CORPS 5% architecture is much more resilient than the DHT 30% architecture, by a magnitude of 1011.

# Discussion

The discussion section focuses on the main challenges/issues you had to overcome during the project. Outline what your approach does better than the ones you mentioned in your related work, and explain why. Do the same with issues where other solutions outperform your own. Are there limitations to your approach? If so, what would you recommend towards removing/mitigating them? Given the experience you've gathered working on this project, are there other approaches that you feel are worth exploring?

# Conclusion

Give a clear, short, and informative summary of all your important results. Answer the initial

question(s) or respond to what you wanted to do, as stated in your introduction. It can be a

short table or a list, and possibly one or two short comments or explanations.

Target a reader who may not have time to read t [2]he whole report yet, but needs the results or

the conclusions immediately. This is a typical situation in real life. Some readers will read your

introduction and skip to your conclusion first, and read the whole report only later (if at all).

You may also draw perspectives. What's missing? In what directions could your work be

extended?

# Bibliography

|  |  |
| --- | --- |
| [1] | R. Guerraoui, "Genuine atomic multicast in asynchronous distributed systems," *Theoretical,* 2001. |
| [2] | V. P. a. L. Iftode, "Byzantine fault tolerant public key authentication in peer-to-peer," *Computer Networks,,* 2006.. |
| [3] | J. R. Douceur, *IPTPS '01.,* 2001. |