# Predicting Blood Transfusions for Coronary Artery Bypass Graft Patients using Deep Neural Networks and Synthetic Data

**Hsiao-Tien Tsai1 • Jichong Wu1 • Puneet Gupta2 • Eric R Heinz2 • Amir Jafari1**

## Abstract

Coronary Artery Bypass Graft (CABG) is a common cardiac surgery but continues to have many associated risks, including the need for blood transfusions. Previous research has shown that blood transfusion during CABG surgery is associated with an increased risk for infection and mortality. The current study aims to use modern techniques, such as deep neural networks and data synthesis, to develop models that can best predict the need for blood transfusion among CABG patients. Results show that neural networks with synthetic data generated by DataSynthesizer has the best performance. Implications of results and future directions were discussed.

**Keywords** deep neural network • data synthesis • blood transfusion • cardiac surgery

✉Hsiao-Tien Tsai  
[jennytsai@gwu.edu](mailto:jennytsai@gwu.edu)

Jichong Wu  
[jwu36@gwu.edu](mailto:jwu36@gwu.edu)

Puneet Gupta MD  
[gupta14@gwu.edu](mailto:gupta14@gwu.edu)

Eric R Heinz MD PhD  
[eheinz@mfa.gwu.edu](mailto:eheinz@mfa.gwu.edu)

Amir Jafari PhD  
[ajafari@gwu.edu](mailto:ajafari@gwu.edu)

1 Data Science, George Washington University, DC, United States

2 Department of Anesthesiology and Critical Care Medicine, George Washington University School of Medicine and Health Sciences, DC, United States

## Introduction

Coronary Artery Bypass Graft (CABG) is a common cardiac surgery but continues to have many associated risks, including infection, major bleeding, and the need for blood transfusion [1] [2] [3]. Blood transfusion is a life-saving intervention in patient care, however, previous research has shown that blood transfusion during CABG surgery is associated with an increased risk for infection and mortality after surgery [1]. Specially, post-operative blood transfusion after CABG is associated with higher odds of readmission and heart failure within 30-days [4].

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 research is to explore modern data analysis techniques and find the models that can best make predictions, such as deep neural networks and synthetic data generation.

## Related Work

### Cardiac Surgery and Blood Transfusion

Research have been conducted to investigate factors that can help to predict major bleeding [5] and the need for red blood cell transfusion after cardiac surgery [3]. In one of the most relevant studies [6], the researchers employed machine learning models to predict perioperative allogeneic blood transfusion for cardiac patients, where the best model (Random Forest) showed good performance (AUC 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 = 3,782), thus the results may not be generalizable to other samples with more diverse demographic backgrounds 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 has been reported to be associated with many known risks regardless of volume. 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 using more advanced techniques and sophisticated models. For example, synthetic data generation techniques have been widely used to train and test neural networks, especially when the data has privacy concerns such in domains such as healthcare and employment [2].

To address this research gap, the current research used the national medical database in the U.S. with a large sample size of over 13,500 data points. Additionally, we predicted the need for blood transfusion regardless of volume. Lastly, we experimented with various modern data analytic approaches to optimize the performance, including deep neural networks and synthetic data generation.

### Neural Networks

ADD LIT REVIEW HERE

### Bayesian Networks and Data Synthesizer

A Bayesian network is a graphical model of the joint probability distribution for a set of variables. The attributes are called *nodes* in the graph, and a conditional relationship between any two attributes is represented as an *edge* between the two nodes. Nodes and edges construct a Bayesian network, and multiple Bayesian networks can be averaged to form a Bayesian model [7]. Bayesian networks are typically used to draw probabilistic inference about one attribute in the network given the values of other attributes, and therefore are suitable to be used for missing data imputation [8] as well as synthetic data generation [2] [9].

Data synthesizer is a tool that takes a dataset as input and generates a structurally and statistically similar synthetic dataset using Bayesian Networks [10]. DataSynthesizer consists of three modules — DataDescriber, DataGenerator and ModelInspector. DataDescriber collects the user-provided information about data, such as data types and correlations between attributes, and produces a data summary, adding noise to the distributions to preserve privacy. DataGenerator samples from the summary computed by DataDescriber and outputs synthetic data. ModelInspector provides statistics and plots for the users to visually inspect the similarity between the real data and the synthetic data.

To define the correlations between attributes in the dataset, DataSynthesizer can operate in one of three modes. In correlated attribute mode, a differentially private Bayesian network [11] is used to capture the correlation structure between attributes, then draw samples from this model to construct the result dataset. Independent attribute mode can be used when there is insufficient data to derive a reasonable correlated model. In this mode, a histogram is created for each attribute, noise is added to the histogram to achieve differential privacy, and then samples are drawn for each attribute. Finally, for cases of extremely sensitive data, one can use random mode that simply generates type-consistent random values for each attribute [10].

In the current research, we used correlated attribute mode as factors that can help to predict blood transfusion are often correlated. When correlated attribute mode is chosen, DataDescriber runs the GreedyBayes algorithm to construct Bayesian networks (BN) to model correlated attributes (see Table 1).

**Table 1 A white and black text with black text

Description automatically generated**Algorithm for GreedayBayes.

In the GreedyBayes algorithm, a Bayesian network N is constructed from input dataset D, attributes A, and the maximum number of parents node k, which defaults to 4. V is the set of visited attributes, and Π is a subset of V that will become parents of node X if added to N. Which attributes Π are selected as parents of X is determined greedily by maximizing mutual information (X , Π). The Bayesian networks constructed in this algorithm gives the sampling order for generating attribute values. When constructing noisy conditioned distributions, Lap(4(d−k)) is injected to preserve privacy, where d is the number of attributes, k is the maximum number of parents of a node, and n is the number of tuples in the input dataset.

## 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). In the current research, we focus on the data from 2015 to 2022, which has a total of 13,534 observations and 296 variables across eight datasets.

First, we built a data preprocessing pipeline to clean the raw data, including imputation (mean for numeric variables and most frequent values for categorical variables), standardization, and encoding (e.g., one-hot encoding for categorical variables). Secondly, preprocessed data was reshaped before entering each neural network. Among the 296 variables, 41 features were identified as most relevant to the current study. The target variable is Occurrences Bleeding Transfusions, which is a binary variable predicting whether the patient needs blood transfusion during or after surgery. The 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 analysis strategies, these features were entered into our models to predict the target variable, and we compared the performance with each other as well as with the benchmarks from previous research.

### Exploratory Data Analysis (EDA)

Among the 13,534 patients in the eight-year combined dataset, nearly 80% are male. The mean age is 65.73 with a standard deviation of 9.82. As for ethnicity composition (see Table 1), nearly half of the patients are white (48%) though over a third did not report their ethnicity (44%). Body Mass Index (BMI) were calculated based on HEIGHT and WEIGHT, indicating the signs of overweight with a mean BMI of 29.26 and a standard deviation of 5.76.

**Table 1** Ethnicity Composition of CABG Patients.

|  |  |
| --- | --- |
| Ethnicity | Counts |
| White | 6,488 |
| Unknown/Not Reported | 5,918 |
| Black or African American | 606 |
| Asian | 381 |
| Some Other Race | 60 |
| American Indian or Alaska Native | 38 |
| Native Hawaiian or Pacific Islander | 36 |
| Native Hawaiian or Other Pacific Islander | 7 |

Among all CABG patients, around half of the patients had blood transfusion (52.8%) and the other half did not (see Figure 1), therefore the binary target variable Bleeding Occurrence is balanced. If further broken down into intra- vs. postop-blood transfusion, we can see that most of the patients who received blood transfusion had it *during* the surgery (86.5%) and only 13.5% had blood transfusion *after* the surgery.

A graph of blue bars

Description automatically generated with medium confidence

**Fig. 1** Bleeding Occurrence Breakdown.

### Analysis Strategy

In this study, we aimed to employ two deep neural networks – Fully-Connected Neural Networks (FNN) and Convolutional Neural Networks (CNN) to predict the need for perioperative blood transfusions for CABG patients. Additionally, we used two approaches to generate synthetic data to train these neural networks – DataSynthesizer and REaLTabFormer (Realistic Relational and Tabular Data using Transformers). Data Synthesizer is based off Bayesian Networks, which are probabilistic graphical models that represent probabilistic relationship between variables. While REaLTabFormer uses a sequence-to-sequence (Seq2Seq) model for generating synthetic relational datasets and uses GPT-2 for non-relational tabular data.

In each type of neural network, we designed different models (e.g., different number of layers, activation functions, loss functions, etc.) and tested them with the original dataset, then we re-ran the models with synthetic datasets from DataSynthesizer and REaLTabFormer and compared the results to see which combination yields the best performance.

## Results and Discussion

### Fully-Connected Neural Networks (FNNs)

In FNNs, we designed eight models varying in complexity (5-layer vs. 7-layer with more neurons), optimizers (SGD vs. Adam), output activation functions (sigmoid vs. softmax) and their corresponding loss functions (binary cross entropy vs. categorical cross entropy) to see which one(s) makes the best predictions. To evaluate the model performance, we looked at metrics across accuracy, f1 score, area under the curve (AUC), rooted mean squared error (rMSE). The best model(s) are determined jointly by f1 score and accuracy score with rMSE and AUC as supplementary benchmarks.

**4.1.1 FNNs with Original Data**

Results from FNN with the original dataset were shown in Table 2. Accuracy scores and f1 scores were landed in the range from .68 to .72, with a lowest rMSE of .46 and a highest AUC of .78. The best model was the five-layer design with SGD as optimizer and softmax as output activation function.

**Table 2** FNN results with original dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | accuracy | f1\_score | rMSE | AUC |
| FNN-5layer-SGD-sigmoid | 0.7215 | 0.6888 | 0.4794 | 0.7764 |
| FNN-5layer-Adam-sigmoid | 0.7218 | 0.6961 | 0.4928 | 0.7756 |
| FNN-7layer-SGD-sigmoid | 0.7174 | 0.6954 | 0.4850 | 0.7741 |
| FNN-7layer-Adam-sigmoid | 0.7174 | 0.6782 | 0.5735 | 0.7752 |
| FNN-5layer-SGD-softmax | 0.7229 | 0.7200 | 0.4643 | 0.7782 |
| FNN-5layer-Adam-softmax | 0.7218 | 0.7179 | 0.4975 | 0.7815 |
| FNN-7layer-SGD-softmax | 0.7181 | 0.7150 | 0.4780 | 0.7780 |
| FNN-7layer-Adam-softmax | 0.7185 | 0.7118 | 0.5587 | 0.7766 |

**4.1.2 FNNs with Synthetic Data from REaLTabFormer**

The eight models showed slightly improved performance with the synthetic data from REaLTabFormer (see Table 3). The accuracy scores and f1 scores ranged from .72 to .74, with a lowest rMSE of .45 and a highest AUC of .80. The best models, once again, were the 5-layer model with SGD using either sigmoid or softmax function.

**Table 3** FNN results with synthetic dataset from REaLTabFormer.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | accuracy | f1\_score | rMSE | AUC |
| FNN-5layer-SGD-sigmoid | 0.7362 | 0.7420 | 0.4789 | 0.8006 |
| FNN-5layer-Adam-sigmoid | 0.7359 | 0.7394 | 0.4650 | 0.8007 |
| FNN-7layer-SGD-sigmoid | 0.7303 | 0.7392 | 0.4819 | 0.7939 |
| FNN-7layer-Adam-sigmoid | 0.7351 | 0.7315 | 0.5577 | 0.7972 |
| FNN-5layer-SGD-softmax | 0.7381 | 0.7380 | 0.4506 | 0.8008 |
| FNN-5layer-Adam-softmax | 0.7355 | 0.7355 | 0.4698 | 0.7999 |
| FNN-7layer-SGD-softmax | 0.7362 | 0.7362 | 0.4787 | 0.7995 |
| FNN-7layer-Adam-softmax | 0.7303 | 0.7286 | 0.5575 | 0.7934 |

**4.1.3 FNNs with Synthetic Data from DataSynthesizer**

Finally, results from FNN with the synthetic data from DataSynthesizer showed significantly improved performance (see Table 4). The accuracy scores and f1 scores ranged from .86 to .87, with a lowest rMSE of .42 and a highest AUC of .94. The best models were the 5-layer SGD model with sigmoid function and the 7-layer model SGD model with softmax function. However, although the 5-layer SGD model with sigmoid function had high accuracy and f1 score, its rMSE was the highest across all models across three datasets (rMSE = .67). An interesting observation was found with DataSynthesizer synthetic data that rMSE were relatively high with SGD optimizer across the board (rMSE =.55~.67) compared with same design with Adam optimizer (rMSE =.42~.45).

**Table 4** FNN results with synthetic dataset from DataSynthesizer.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | accuracy | f1\_score | rMSE | AUC |
| FNN-5layer-SGD-sigmoid | 0.8718 | 0.8776 | 0.6790 | 0.9386 |
| FNN-5layer-Adam-sigmoid | 0.8689 | 0.8772 | 0.4340 | 0.9474 |
| FNN-7layer-SGD-sigmoid | 0.8685 | 0.8774 | 0.5853 | 0.9358 |
| FNN-7layer-Adam-sigmoid | 0.8626 | 0.8712 | 0.4222 | 0.9456 |
| FNN-5layer-SGD-softmax | 0.8751 | 0.8749 | 0.6119 | 0.9421 |
| FNN-5layer-Adam-softmax | 0.8696 | 0.8696 | 0.4286 | 0.9495 |
| FNN-7layer-SGD-softmax | 0.8762 | 0.8762 | 0.5596 | 0.9401 |
| FNN-7layer-Adam-softmax | 0.8670 | 0.8667 | 0.4513 | 0.9474 |

### Convolutional Neural Networks (CNNs)

We also compared other neural network model such as CNN. In CNN, we transformed the 2-dimensional tabular dataset to 4-dimensional to meet the data format required by CNN. Two datasets derived from different synthetic data generation techniques (REaLTabFormer and DataSynthesizer) were used to run the CNN model and the accuracy was far lower than the FNN models (see Table 5).

**Table 5** CNN results with across different preprocessed datasets.

|  |  |  |
| --- | --- | --- |
| Model | dataset | accuracy |
| CNN-2D-2layer-noPooling-ReLU | synthetic dataset 2015-2022 from REaLTabFormer | 0.6681 |
|  | synthetic dataset 2015-2022 from DataSynthesizer | 0.5785 |

## Conclusion

The current research aimed to employ modern techniques to develop models that can best predict the need for blood transfusions among CABG patients. In some cases, deep neural networks combined with data synthesis techniques have shown to significantly improve model performance. Especially in FNNs, regardless of model complexity and design, models trained with synthetic data generated from DataSynthesizer had best performance across the board, with f1 score and accuracy ranged from .86 to .87, with a lowest rMSE of .42 and a highest AUC of .94. Future research should look into different methodologies to generate synthetic data for training and developing models, both tree-based models and deep neural networks, that can help inform guidelines for major high-risk surgeries.

# References

|  |  |
| --- | --- |
| [1] | K. Horvath, M. Acker, H. Chang and E. Bagiella et al, "Blood transfusion and infection after cardiac surgery," *The Annals of Thoracic Surgery,* vol. 95, no. 6, pp. 2194-2201, June 2013. |
| [2] | D. Kaur, M. Sobiesk, S. Patil, J. Liu, P. Bhagat, A. Gupta and N. Markuzon, "Application of Bayesian networks to generate synthetic health data," *Journal of the American Medical Informatics Association,* vol. 28, no. 4, pp. 801-811. doi: 10.1093/jamia/ocaa303, March 2021. |
| [3] | Q. Li, H. Lv, Y. Chen, J. Shen, J. Shi and C. Zhou, "Development and validation of a machine learning prediction model for perioperative red blood cell transfusions in cardiac surgery," *International Journal of Medical Informatics,* vol. 184, p. 105343, April 2024. |
| [4] | S. Mufarrih, F. Mahmood, N. Qureshi, R. Yunus and R. Matyal et al, "Timing of blood transfusions and 30-day patient outcomes after coronary artery bypass graft surgery," *Journal of Cardiothoracic and Vascular Anesthesia,* vol. 37, no. 3, pp. 382-391, doi: 10.1053/j.jvca.2022.11.029, Mar 2023. |
| [5] | Y. Gao, X. Liu, L. Wang, S. Wang, Y. Yu, Y. Ding, J. Wang and H. Ao, "Machine learning algorithms to predict major bleeding after isolated coronary artery bypass grafting," *Frontiers in Cardiovascular Medicine,* vol. 9, p. doi: 10.3389/fcvm.2022.881881, July 2022. |
| [6] | T. Tschoellitsch, C. Bock, T. Mahecic, A. Hofmann and J. Meier, "Machine learning-based prediction of massive perioperative allogeneic blood transfusion in cardiac surgery," *European Journal of Anaesthesiology,* vol. 39, no. 9, pp. 766-773, doi: 10.1097/EJA.0000000000001721, Sep 2022. |
| [7] | J. Young, P. Graham and R. Penny, "Using Bayesian networks to create synthetic data," *Journal of Official Statistics,* vol. 25, no. 4, pp. 549-567, 2009. |
| [8] | M. Di Zio, M. Scanu, L. Coppola, O. Luzi and A. Ponti, "Bayesian networks for imputation," *Journal of the Royal Statistical Society,* vol. 167, pp. 309-322, 2004. |
| [9] | M. Baowaly, C. Lin, C. Liu and K. Chen, "Synthesizing electronic health records using improved generative adversarial networks," *Journal of the American Medical Informatics Association,* vol. 26, no. 3, pp. 228-241, March 2019. |
| [10] | H. Ping, J. Stoyanovich and B. Howe, "DataSynthesizer: Privacy-preserving synthetic datasets," in *International Conference on Scientific and Statistical Database Management*, 2017. |
| [11] | J. Zhang, G. Cormode, C. Procopiuc, D. Strivastava and X. Xiao, "PrivBayes: private data release via bayesian networks," in *SIGMOD International Conference on Management of Data*, 2014. |