Predictive Factors for ACS Post-HCT in SCD: Towards Personalized Risk Stratification Model

Chengxi Wang, Jia Wei, Yuchen Chang, Yun Yang

1 Introduction

Acute chest syndrome (ACS) is defined by the emergence of a new radiodensity on chest imaging, accompanied by fever and/or respiratory symptoms. As a potentially fatal acute complication of sickle cell disease (SCD), ACS demands immediate intervention across all ages [3]. Hematopoietic cell transplantation (HCT) presents a curative approach for SCD, yet it is accompanied by significant post-transplant risks, including ACS, which poses a substantial threat to patient outcomes [1].

The recurring incidence of ACS following HCT underscores the critical need for the identification of predictive factors that could inform both pre- and post-transplant management strategies. These factors encompass a range of patient demographics, disease characteristics, and specific elements of the transplant procedure itself. Achieving a detailed understanding of these predictors is essential for risk stratification, the refinement of transplant protocols, and the customization of post-transplant care, all aimed at minimizing the risk of ACS and enhancing survival rates.

The core research question explores whether the specifics of patient demographics, disease characteristics, and transplant details can act as predictive factors for the occurrence of ACS post-HCT in patients with sickle cell disease. This inquiry seeks to map out how variables such as age, sex, genetic background, disease severity, baseline hemoglobin levels, and a history of ACS, along with the type of transplant, the source of stem cells, and the conditioning regimen, play a role in forecasting ACS post-transplantation.

Specifically, this research focuses on two key aims to improve outcomes for SCD patients undergoing HCT. The first objective is to identify significant predictors of ACS post-HCT, including patient demographics (age, sex, genetic background), disease characteristics (severity, baseline hemoglobin, history of ACS), and transplant specifics (type, stem cell source, conditioning regimen). The second objective is to develop a risk stratification model to categorize patients by their likelihood of developing ACS post-HCT, enabling targeted monitoring and interventions for those at high risk.

Through these endeavors, the research aspires to enhance patient care and outcomes by enabling personalized treatment strategies and reducing the incidence of ACS.

2 Significance and Motivation

The development of a robust risk-stratification model to predict ACS after HCT in SCD patients holds critical importance for improving clinical outcomes and decision-making [2]. This model would fundamentally change pre-transplant counseling, enabling personalized assessments of the risks associated with post-transplant complications. As a result, patients and clinicians can make better-informed decisions about undergoing HCT, leading to more balanced discussions on the potential benefits and risks of the procedure.

Moreover, a successful risk-stratification model would allow for the implementation of precise interventions aimed at high-risk patients, potentially including customized conditioning regimens, prophylactic treatments, and enhanced monitoring strategies. These proactive measures could significantly reduce the incidence or severity of ACS post-transplant, thereby improving patient outcomes and optimizing resource use in healthcare settings.

By incorporating this model into the standard pre-transplant assessment and ongoing post-transplant care, healthcare providers can offer more targeted and effective management strategies. This approach ensures that treatment plans are continuously adjusted to meet the evolving needs and risk profiles of patients, further enhancing the safety and efficacy of transplantation. Ultimately, the implementation of this predictive risk-stratification model would mark a significant advancement in transplant medicine, providing actionable insights that improve the quality of life for individuals with SCD.

3 Data Description

The data for our research is sourced from the HCT and Cellular Therapy Data within the CIBMTR database. This extensive dataset captures outcomes from HCT in individuals with sickle cell disease, gathered over several years from multiple clinical centers. It includes anonymized patient identifiers, detailed accounts of the transplantation procedures, and comprehensive follow-up data concerning post-transplant complications and outcomes. The dataset contains key variables such as patient demographics, the number and types of transplants (including donor and graft types), and clinical outcomes, notably the presence of cytomegalovirus pre-transplant and various post-transplant health conditions. For our analysis, we have selected ACSPSHI as the target variable, which denotes the occurrence of Acute Chest Syndrome post-HCT.

3.1 Data Pre-processing

In the data preprocessing phase, we focused on refining the dataset to ensure the integrity and accuracy of our analysis on HCT outcomes. Our initial step, data cleaning, involved selectively including only pre-HCT measures while excluding duplicate entries that provided redundant information. We also undertook a comprehensive missing value imputation process to address gaps in the dataset.

The imputation process employed the missForest algorithm, a robust non-parametric method that uses random forests for imputation. This technique begins with initial rough estimates of missing values and iteratively refines these estimates by assessing variable importance, thus ensuring a high level of accuracy in the imputed data.

For inclusion and exclusion criteria, we removed samples lacking records of ACS post-HCT to maintain a focus on our research objectives. The preparation for imputation identified critical numerical variables such as SCREATPR, SCREAULN, HB1PR, INTSCREPR, and AGEGPFF, which are vital for our analysis. Non-numerical variables were converted into factors to suit the requirements of the missForest algorithm, which is adept at handling mixed data types.

After implementing the preprocessing methods described above, we successfully obtained a cleaned and imputed dataset consisting of 804 rows and 32 columns. This refined dataset is now optimally prepared for detailed analysis and further research investigations.

3.2 Exploratory Data Analysis

After cleaning the dataset and imputing missing values, we embarked on an exploratory data analysis to examine the dataset's characteristics and address comments received. We initiated this analysis by generating pie charts to display the numerical proportions of each level for all categorical variables (see Figure 1). Additionally, we created bar plots segmented by the outcome value. Upon reviewing these plots, we observed a significant imbalance in the number of patients with and without ACS post-HCT, with 780 records indicating ACSPSHI=1 and only 24 showing ACSPSHI=0 (see Figure 2). To address this imbalance, we plan to explore various weighting methods in our modeling approach, aiming to achieve more accurate predictions and robust results.

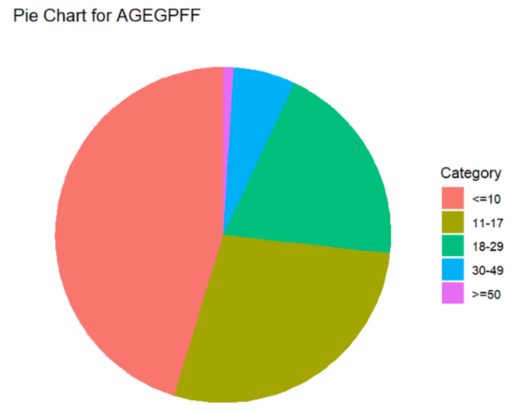


Figure 1: Neural Network Structure

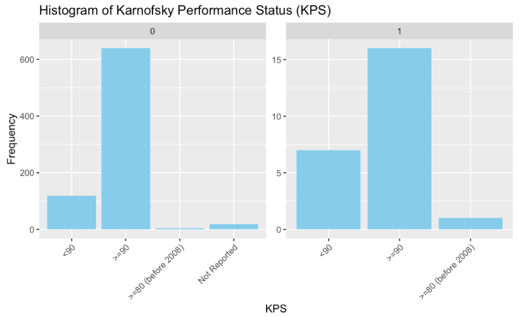


Figure 2: Neural Network Structure

4 Specific Aim 1

To pinpoint predictors of Acute Chest Syndrome post-HCT, including patient demographics, disease severity, and transplant details.

4.1 Hypothesis

We hypothesize that certain patient demographics (age, sex, genetic background), disease characteristics (severity, baseline hemoglobin levels, history of ACS), and transplant specifics (type of transplant, source of stem cells, conditioning regimen) are significant predictors of the occurrence of ACS following HCT.

4.2 Rationale

The variability in ACS occurrence post-HCT among sickle cell disease patients highlights a complex interaction among patient demographics, disease characteristics, and transplant specifics. This complexity is further compounded by the rarity of ACS events relative to non-events, which poses a significant challenge to traditional analytical methods in accurately identifying risk predictors. Therefore, our first aim is to address this analytical gap by employing advanced statistical models capable of handling sparse outcome data, with the goal of enhancing the predictive accuracy and identifying nuanced risk factors that could be obscured in conventional analyses. By unraveling these intricate associations, we aim to generate insights that could lead to improved patient management and outcome prediction in the post-HCT setting for individuals with SCD.

4.3 Experimental Approach

4.3.1 Propensity Score Matching (PSM)

- **PSM-Model 1** utilized forward selection to meticulously build upon a minimal model, guided by improvements in AIC values. Our custom loop in R, while methodical, resulted in a model with suboptimal p-values, prompting us to explore further.
- **PSM-Model 2** employed the "MASS" package's stepAIC function to perform a more automated forward selection. However, this too did not yield the discriminative power necessary to establish causal inferences confidently.
- **PSM-Model 3** took a straightforward approach by evaluating the first five variables for their propensity score contribution but similarly fell short in producing statistically compelling results.

Given the suboptimal p-value outcomes of 1 from the PSM approach, suggesting potential model misspecification, we expanded our experimental framework to incorporate supplemental methods more adept at handling the intrinsic complexities of feature selection, without the direct inference of causality.

4.3.2 Embedded Methods

We incorporated embedded methods as intrinsic parts of the model construction process, which allow for simultaneous feature selection and model training, thereby mitigating the risk of model misspecification.

- **Lasso Regression:** By applying a penalization parameter to the regression model, Lasso regression effectively zeroes out less significant coefficients. This penalty term adds robustness against overfitting and helps in the selection of features that contribute most significantly to the model, independent of p-values from traditional hypothesis testing.
- **Random Forest:** This ensemble method provided an alternative perspective through feature importance scores derived from numerous decision trees. By aggregating the predictive power of various features across different models, we obtained a consensus on feature relevance, independent of the limitations identified in the PSM method.

4.3.3 Filter Methods

- **Information Gain:** We applied this measure to evaluate the entropy reduction achieved by each feature, offering a direct and unassuming assessment of feature relevance. Unlike PSM, information gain does not rely on a specified model structure and therefore is not affected by model misspecification. By ranking features according to their information gain, we were able to identify those that provided the most significant reduction in uncertainty regarding the outcome variable.

4.3.4 Wrapper Methods

- **Recursive Feature Elimination (RFE):** We addressed feature selection as an optimization problem, utilizing RFE to systematically evaluate different combinations of features. This method iteratively constructed models and removed the least impactful attributes, refining the feature set based on model performance. This process continued until the most predictive features were identified, offering a data-driven solution to feature selection that circumvents the pitfalls of p-value reliance.

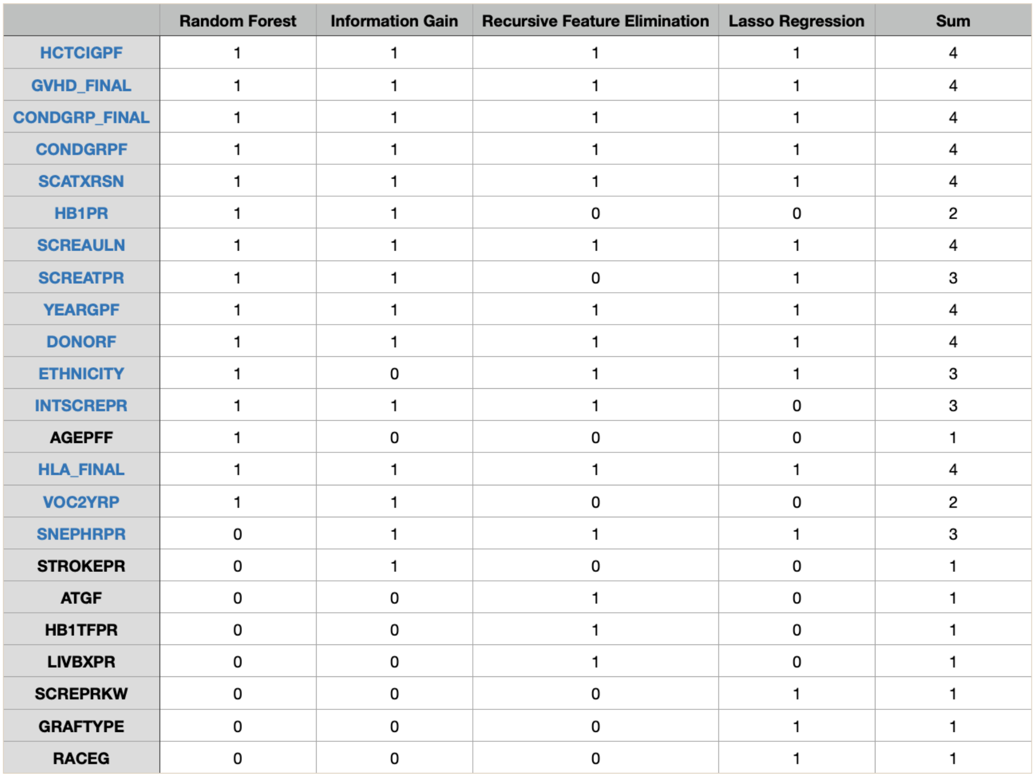
4.4 Results

After implementing the four models—Lasso Regression, Random Forest, Information Gain, and Recursive Feature Elimination—we synthesized the insights drawn from each method to create a multifaceted perspective on the data:

- **Select the Top 15 Features:** Identified the 15 most important features from each model, based on the strength and consistency of their predictive power.

- **Count Feature Occurrences:** Calculated how often each feature appeared in the top 15 across the different models.
- **Aggregate Top Features:** Chose the 15 features that recurred most frequently across the models. This number was selected because the inclusion of more than 15 features led to significantly varied results and less model stability.

By doing so, we sought to identify a robust set of predictors that were not only statistically significant according to one methodology but were also consistently recognized across multiple analytical approaches. The integration of results (See Figure 3) from Lasso Regression, Random Forest, Information Gain, and Recursive Feature Elimination allowed us to compensate for the potential biases and limitations inherent in any single modeling technique. This composite approach enhances the validity of our findings and underscores the complexity of ACS post-HCT, reflecting its multifactorial nature. The key advantage of this integrative strategy is the ability to capture different dimensions of feature relevance. Lasso Regression highlighted predictors by imposing sparsity, thus ensuring only the most impactful features were retained. Random Forest offered a non-linear perspective on feature importance, which is critical for capturing complex interactions. Information Gain served as a filter to rank features based on their individual contribution to entropy reduction, thus prioritizing features that offer the most information about the outcome. RFE provided a systematic reduction approach, reinforcing the selection of features that consistently contribute to model performance.



	Random Forest	Information Gain	Recursive Feature Elimination	Lasso Regression	Sum
HCTCIGPF	1	1	1	1	4
GVHD_FINAL	1	1	1	1	4
CONDGRP_FINAL	1	1	1	1	4
CONDGRPF	1	1	1	1	4
SCATXRSN	1	1	1	1	4
HB1PR	1	1	0	0	2
SCREAULN	1	1	1	1	4
SCREATPR	1	1	0	1	3
YEARGPF	1	1	1	1	4
DONORF	1	1	1	1	4
ETHNICITY	1	0	1	1	3
INTSCREPR	1	1	1	0	3
AGEPFF	1	0	0	0	1
HLA_FINAL	1	1	1	1	4
VOC2YRP	1	1	0	0	2
SNEPHRPR	0	1	1	1	3
STROKEPR	0	1	0	0	1
ATGF	0	0	1	0	1
HB1TFPR	0	0	1	0	1
LIVBXPR	0	0	1	0	1
SCREPRKW	0	0	0	1	1
GRAFTYPE	0	0	0	1	1
RACEG	0	0	0	1	1

Figure 3: Selected Top 15 Important Features

4.5 Interpretation of Results

After analyzing on the importance values, we conclude that Factors such as pre-existing health conditions, GVHD prophylaxis (pro-pho-laxis) strategies, the intensity of the conditioning regimen, and baseline he-mo-glo-bin and cre-ati-nine levels significantly influence the risk of ACS post-HCT. These insights suggest that personalized, pre-transplant risk assessments and tailored management strategies could potentially reduce ACS incidence and improve patient outcomes.

5 Specific Aim 2

To develop a risk stratification model that categorizes patients based on their likelihood of developing ACS after HCT

5.1 Hypothesis

We hypothesize that a combination of patient demographics, disease characteristics, and transplant specifics can significantly predict the risk of ACS post-HCT in sickle cell disease patients.

5.2 Rationale

The development of ACS post-HCT significantly impacts patient outcomes in sickle cell disease, necessitating early identification and intervention for those at highest risk. By leveraging advanced machine learning techniques, we aim to create a nuanced risk stratification model that can predict ACS occurrence with high accuracy, thus enabling targeted monitoring and preventative care.

5.3 Experimental Approach

- **Random Forest:** We plan to use Random Forest as one of our predictive modeling techniques. This ensemble learning method operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. Random Forest is particularly suited for our dataset due to its robustness against overfitting and its efficacy in handling categorical and continuous variables.
- **XGBoost:** XGBoost is renowned for its performance and speed in classification tasks. This algorithm will allow us to handle the imbalanced nature of our dataset efficiently, offering a sophisticated mechanism to boost the predictive accuracy. By adjusting its parameters, we aim to optimize the model's ability to predict the nuanced risk scores of ACS post-HCT, providing a precise categorization of patients based on their risk levels.
- **Feed Forward Neural Network:** Through our EDA process, we observed significant imbalance in our dataset, especially for the number of subjects that have/do not have ACS post-HCT. As Neural networks have the capacity to learn complex relationships in the data through their layered architecture, we thought it can be beneficial for capturing patterns in imbalanced datasets.

The rationale behind choosing these three models is based on the fact that all of them are well-suited for handling imbalanced classes, which is a problem that existed in our dataset. These algorithms employ ensemble learning techniques that aggregate predictions from multiple decision trees, making them robust against class imbalance. Besides, both Random Forest and XGBoost offer mechanisms for feature importance assessment, while Feed Forward Neural Network allows for summarizing the weights across all neurons. This process provides valuable insights into the contribution of each feature to the predictive performance, and could be really helpful for us to interpret the final results.

Compared to other machine learning methods such as logistic regression, decision tree, and support vector machine, Random Forest, XGBoost and Feed Forward Neural Network are much better to use for our modeling objectives. Logistic regression may struggle with capturing non-linear relationships and interactions between predictors, which are essential for predicting ACS occurrence accurately. A single decision tree might be more prone to overfitting than the technique Random Forests. SVMs may require careful tuning of hyperparameters and kernel selection, which can be computationally expensive and less interpretable compared to ensemble tree methods.

We approach model fitting following the classic way of training a machine learning model. We repeated the model fitting process 100 times. For each step, we randomly split the dataset, take 70

For the model setting of Random Forest and XGBoost, according to the result of 5 fold cross validation and grid search for selecting best combinations of hyper-parameters, we set the the number of features randomly selected from the full set of predictors at each split to be 2, and the number of trees growth in a forest to be 500 for random forest model and the learning rate as 0.1, the max depth of tree as 3 for XGBoost. To deal with the imbalance dataset, we also set the class weight parameter in Random Forest Model and the imbalance ratio in XGBoost according to each training set.

We then fit a simple feed forward neural network with only one hidden layer. We specify the number of neurons in the hidden layer to be 5, the maximum number of iterations (epochs) for training is 200, which is the number we started to observe the convergence. For initialization, we set the weights to be randomly within the range of -0.1 to 0.1. To prevent overfitting, we also added L2 regularization coefficient in our model.

5.4 Results

After 100 epochs of training, we got the average accuracy for the three models, 0.97 for Random Forest, 0.939 for XGBoost, and 0.932 for Feed Forward Neural Network. The average specificity is 1 for Random Forest, 0.966 for XGBoost, and 0.957 for Feed Forward Neural Network. The average recall is 0 for Random Forest, 0.0788 for XGBoost, and 0.106 for Feed Forward Neural Network.

By observing the model fitting result of these two models, we can see that the random forest is predicting all samples to have class 0, and thus result in a higher accuracy than XGBoost. However, this is not what we expected for reality, instead, we will focus on the higher recall value achieved by XGBoost, as a higher recall means that the model is better at identifying individuals who actually have the ACS post HCT. This is crucial in healthcare, that accurately identifying positive cases is essential for providing timely treatment and preventing the spread of the disease.

In general, the Neural Network model has better performance than the XGBoost model in terms of a higher recall value. By checking the variance of the recall value (both are around 0.15), we noticed that they have similar variances, thus we confirmed that this improved performance should not be caused by randomness.

So the neural network model is our final choice for predicting the risk scores of each patient. We visualized the network structure to understand its inner relationship. (See Figure 4)

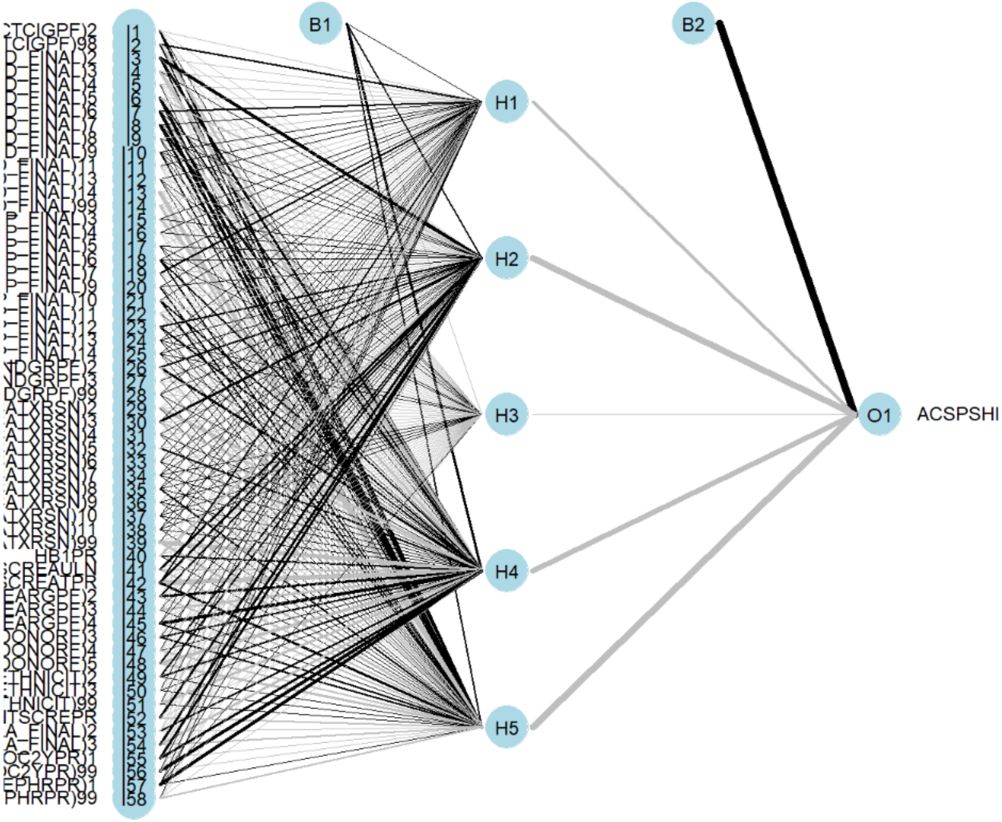


Figure 4: Neural Network Structure

5.5 Interpretation of Results

We used the built-in feature importance metrics provided by both Random Forest and XGBoost to identify which variables (e.g., patient demographics, disease characteristics, transplant specifics) most strongly influence the risk prediction, and selected the top 15 variables for building our advanced predictive models(See Figure 3). This can also inform clinicians about the key factors to monitor.

Clinical interpretation also involves understanding how well the model performs. We used metrics like AUC-ROC for overall discrimination ability, precision-recall curves for balance between sensitivity and precision. (See Figure 5, 6, 7)

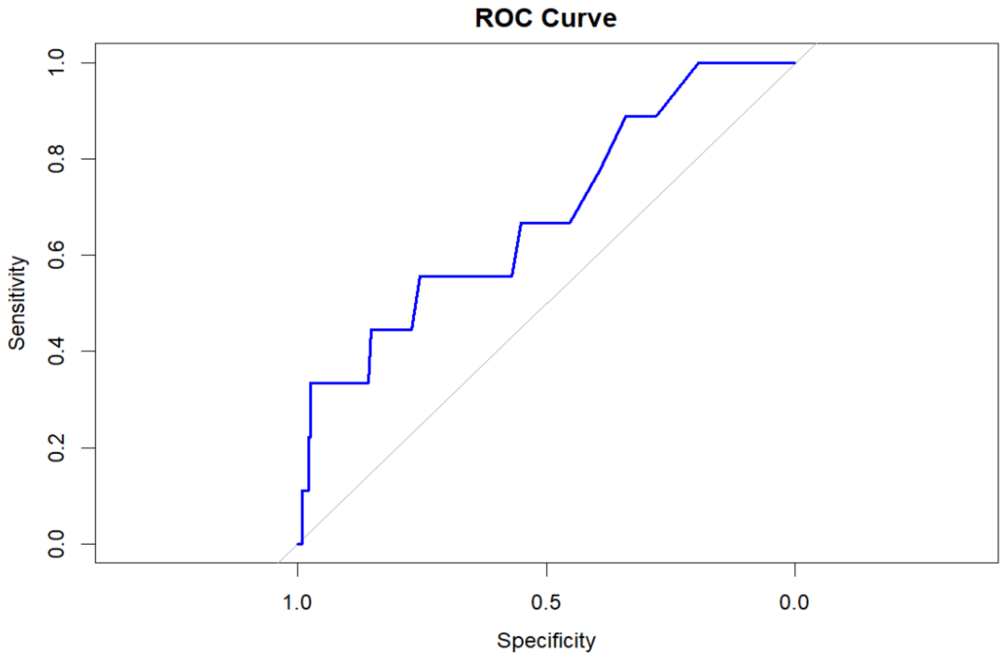


Figure 5: ROC Curve for Random Forest

For a more intuitive interpretation of the result, we calculated the importance by the Olden's algorithm, which calculates the product of the input to hidden and hidden to output layer weights between for each neuron and sums the product across all hidden neurons. A larger absolute importance value indicates higher influence of a predictor on the outcome. (See Figure 8)

Upon evaluating the model's performance, the focus will be on its clinical utility, particularly its capability to assign a numeric risk score to each patient. We take the estimated probability from the neural network model as a risk score, this score indicates the likelihood of developing ACS post-HCT, where a higher score suggests a greater risk. In practice, patients with higher risk scores might warrant closer monitoring, more aggressive pre-emptive treatment, and perhaps modifications to their HCT regimen. Conversely, those with lower scores could be managed with standard post-HCT care protocols.

By analyzing the values of variable importance, we found that when subjects having type CD 34 selection, CNI + MMF, CNI + MTX and CNI alone for GVHD prophylaxis (GVHD_FINAL = 2,5,6,7) will decrease the risk of having ACS post-HCT. Subjects with Sickle nephropathy pre-conditioning (SNEPHRPR1=1), vaso-occlusive crisis requiring hospitalization within 2 years pre-HCT (VOC2YPR = 1) and who received surgery during 2018-2020 (YEARGPF = 4) will result in lower risk of having ACS post-HCT. We also observed that subjects with type Post-CY + siro +/- MMF and MTX + siro (GVHD_FINAL = 3, 13) will have a higher risk of having ACS post-HCT. For subjects with mismatched unrelated donor and cord blood (DONORF = 5) will also result in higher risk of having ACS post-HCT.

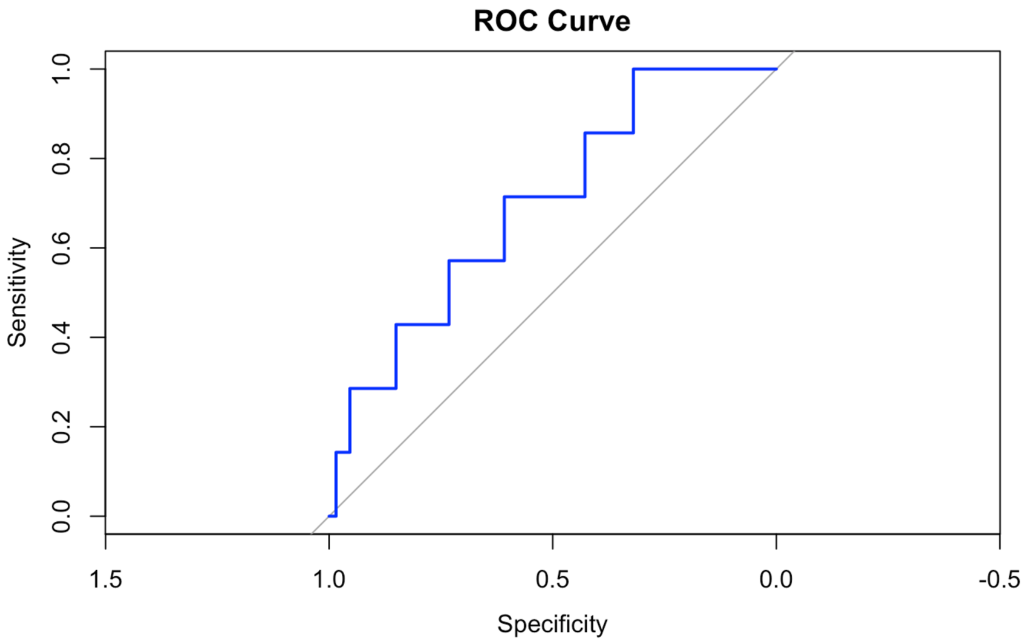


Figure 6: ROC Curve for XGBoost

6 Conclusion

In conclusion, our comprehensive analysis utilizing different modeling techniques has provided valuable insights into the factors influencing the risk of ACS post-HCT. Through the integration of Lasso Regression, Random Forest, Information Gain, and Recursive Feature Elimination methods, we have identified 15 predictors that consistently contribute to predictive performance across multiple analytical approaches. Our findings emphasize the multifactorial nature of ACS post-HCT, highlighting the significance of some pre-existing health conditions specifically.

Besides, the evaluation of model performance has led us to prioritize the use of the Neural Network model, due to its superior recall value compared to other models such as XGBoost. This enhanced recall indicates the model's effectiveness in accurately identifying individuals at risk of ACS post-HCT, crucial for timely intervention and patient care.

From the Neural Network model, we can get variables' importance that can shed light on specific factors that either increase or decrease the risk of ACS post-HCT, informing clinical decision-making and highlighting areas for targeted interventions for patients. These insights are quite important to design personalized risk assessment and management strategies, ultimately contributing to the improvement of patient care and outcomes in the context of Hematopoietic Cell Transplantation.

7 Clinical Interpretation

7.1 Personalized Risk Assessment

This clinical application will involve providing personalized risk scores and implementing tailored strategies for high risk individuals. Our neural network model will provide risk probabilities for each patient, and stratify them based on their likelihood of developing ACS, post-HCT. For those patients identified as high-risk, we will engage in shared decision regarding their treatment options, offering comprehensive counseling sessions to discuss the implications of their risk profiles. If the high-risk patients choose to proceed with HCT, we prioritize the implementation of personalized pre-strategies aimed at mitigating the risk of ACS, post-HCT.

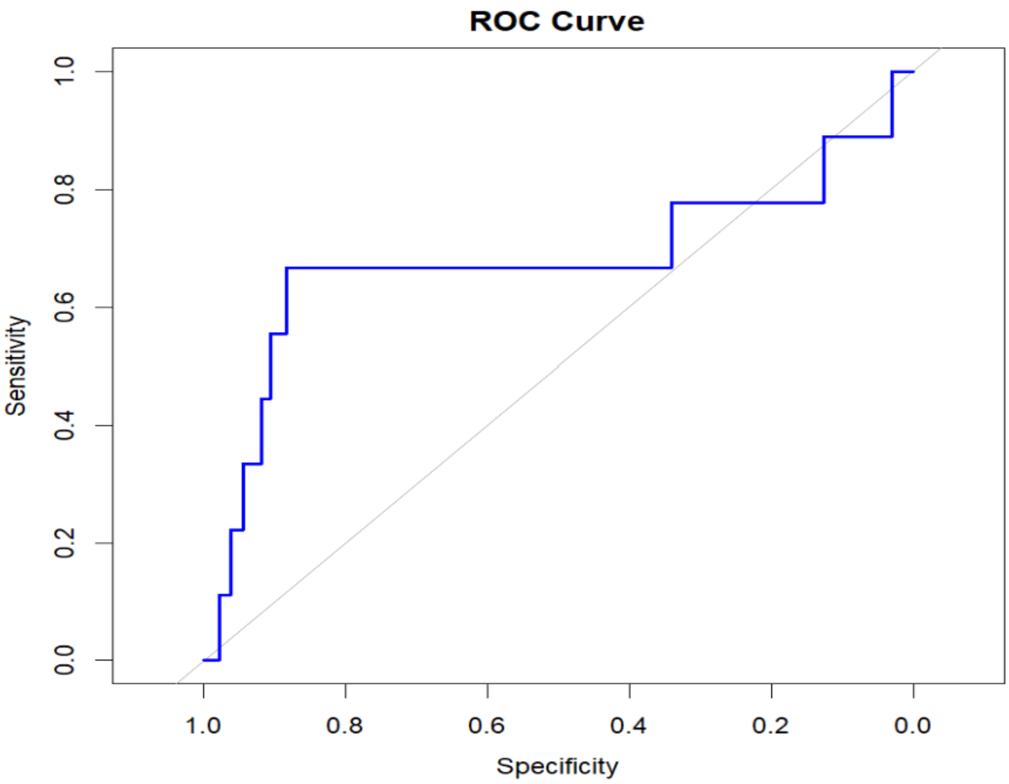


Figure 7: ROC Curve for Neural Network

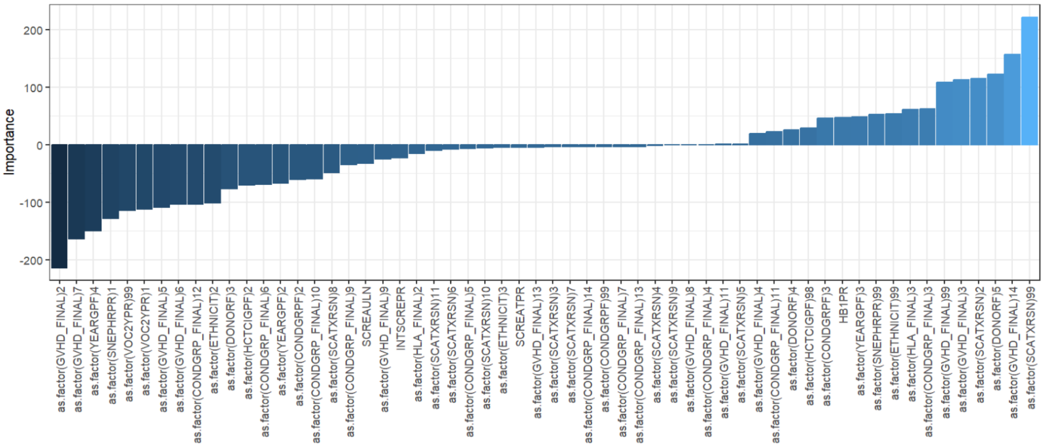


Figure 8: Variable Importance for Neural Network

7.2 Resource Allocation

Our neural network model provides valuable insights related to how patient demographics, disease characteristics, and details of the transplant procedure influences the risk of ACS post-HCT, which can be an instruction for clinicians in optimizing resource use, reducing waste and lowering costs significantly. Besides, efficient resource allocation ensures that patients receive timely and appropriate care, improving overall outcomes. With our model, we can prioritize high-risk patients, directing health resources where they are most needed.

7.3 Limitations and Future Work

There are several limitations that warrant attention in our research study. Firstly, our current dataset has imbalanced problem with only 3% of samples being positive for ACS post-HCT. This imbalance could potentially impact the generalizability and reliability of our results. While we addressed this issue by adjusting the imbalanced ratio as a parameter in our models, we acknowledge the necessity of exploring alternative methods, such as SMOTE for the imbalance and cross-validation to prevent overfitting. If these methods are insufficient, we will consider alternative approaches like cost-sensitive learning and ensemble techniques to enhance the model's performance and robustness.

Another limitation is regarding the inclusion of only pre-HCT patient features, which may restrict the predictive capbilities of our models due to the limited feature set. To enhance our model in the future, we may plan to incorporate features measured after HCT in our model and assess their impact on the model's prediction accuracy and robustness.

Finally, our model's validation currently relies on internal validation methods. However, to further increase its generalizability and reliability to diverse healthcare settings, we are considering validating our model using external datasets for our future work. This approach will provide insights into the model's performance in real-world scenarios beyond the confines of our original dataset, thereby enhancing its practical utility and reliability.

References

[1] O. Castro, D. J. Brambilla, B. Thorington, C. A. Reindorf, R. B. Scott, P. Gillette, J. C. Vera, and P. S. Levy. The acute chest syndrome in sickle cell disease: incidence and risk factors. the cooperative study of sickle cell disease. 1994.

[2] R. N. Paul, O. L. Castro, A. Aggarwal, and P. A. Oneal. Acute chest syndrome: sickle cell disease. *European journal of haematology*, 87(3):191–207, 2011.

[3] E. P. Vichinsky, L. A. Styles, L. H. Colangelo, E. C. Wright, O. Castro, B. Nickerson, and C. S. of Sickle Cell Disease. Acute chest syndrome in sickle cell disease: clinical presentation and course. *Blood, The Journal of the American Society of Hematology*, 89(5):1787–1792, 1997.