BIS 687 Group 5 Proposal

Chengxi Wang, Jia Wei, Yuchen Chang, Yun Yang

2024-02-16

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. 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.

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.

Research Strategy

A. Significance

The development of a robust risk-stratification model for predicting ACS post-HCT in SCD patients carries profound significance for both clinical practice and patient outcomes.

This model would revolutionize pre-transplant counseling and decision-making processes. Patients and clinicians could make more informed choices about pursuing HCT by assessing the personalized risk of post-transplant complications like ACS. This understanding could lead to a more balanced discussion of the potential benefits and risks of transplantation, allowing patients to weigh their options and make decisions aligned with their individual health goals and preferences.

A successful risk-stratification model would enable the implementation of targeted interventions. Clinicians could utilize the model to identify high-risk patients who may benefit from preemptive measures aimed at reducing the likelihood or severity of ACS following HCT. These interventions could encompass personalized conditioning regimens, prophylactic antibiotic therapy, enhanced monitoring protocols, and tailored supportive care plans. By intervening proactively based on individualized risk profiles, healthcare providers could potentially mitigate the occurrence and impact of ACS, thereby improving patient outcomes and reducing the burden on healthcare resources.

In practice, clinicians would integrate the risk-stratification model into their pre-transplant assessment process. Before undergoing HCT, patients would undergo comprehensive evaluations to determine their risk profile for post-transplant complications, including ACS. This evaluation would encompass factors such as patient demographics, disease characteristics, and transplant specifics. Based on the identified risk level, clinicians would tailor their approach to transplantation, including the selection of appropriate interventions to mitigate ACS risk.

The risk-stratification model could also inform post-transplant care strategies. Patients identified as high-risk for ACS could receive intensified monitoring and follow-up care, allowing for early detection and intervention in case of complications. Additionally, the model could facilitate ongoing adjustments to treatment plans based on evolving patient needs and risk profiles, ensuring that care remains personalized and responsive to individual circumstances.

Overall, the development and implementation of a predictive risk-stratification model for ACS post-HCT in SCD patients would represent a significant advancement in the field of transplant medicine. By providing actionable insights into patient risk profiles and guiding personalized interventions, such a model has the potential to optimize patient outcomes, enhance the safety and efficacy of HCT, and ultimately improve the quality of life for individuals living with SCD.

B. Innovation

Firstly, we are applying a multifactorial approach to this project. By integrating patient demographics, disease characteristics, and transplant specifics, our project is more comprehensive compared to studies focusing on isolated features like neurological factors Vichinsky et al. (2000) or blood testing results Castro et al. (1994). This approach will provide us a multidimensional understanding of the patient experience, which would also benefit any future post-HCT interventions.

Secondly, compared with traditional statistical methods (like nonparametric hypothesis tests, ANOVA, etc.) Vichinsky et al. (1997), that are currently applied into previous ACS studies, we are employing advanced machine learning and causal inference techniques to find out influential factors of ACS post HCT and develop predictive models for ACS occurrence, which will enhance the accuracy and reliability of the results given by our analysis.

Finally, we focus on post-HCT patients. As ACS is a frequent cause of hospitalization and death for patients with SCD, there's already a lot of studies Paul et al. (2011) that focuses on the population with SCD. Our project contributes to filling the gap of narrowing down the population into patients who have received HCT specifically, we aim to find out which factors are influencing the ACS occurrence, thereby contributing valuable insights to post-HCT interventions.

C. Research Plan

One potential curative treatment method for these SCD patients is HCT, but it is associated with certain risks, including the development of the disease ACS. ACS is a life-threatening complication characterized by pulmonary vaso-occlusion and can lead to respiratory failure and even death. Despite the development of current transplantation techniques, ACS remains an important cause of morbidity and mortality post-HCT in SCD patients. Understanding the factors for ACS occurrence post-HCT is crucial for risk stratification and improving patient outcomes.

Our research will use a dataset coming from the Center for International Blood and Marrow Transplant Research (CIBMTR) database. This dataset contains detailed clinical and demographics data for individuals diagnosed with SCD and have undergone HCT as part of their treatment.

Prior to conducting our research, we will conduct a comprehensive literature review of existing studies related to ACS post-HCT in patients diagnosed with SCD. Based on the findings from this review, we will propose using statistical methods and machine learning techniques to investigate certain specific demographics factors, disease characteristics, and transplant specifics that may act as predictive factors for the occurrence of ACS post-HCT in patients with SCD.

To ensure the reliability and applicability of our predictive models without modifying the dataset, we will focus on robust modeling techniques. This includes rigorous feature selection to prioritize clinically relevant variables and model regularization to minimize overfitting. We will employ cross-validation within the dataset to assess model stability and consistency across different subsets of data, this ensures the internal validation. Additionally, we will explore alternative validation methods such as bootstrapping or resampling to simulate external validation. These approaches will help establish the reliability and generalizability of our findings across diverse patient populations and settings. Analyses within specific patient subgroups (e.g., different age groups, disease severities) can also be employed to identify any performance discrepancies.

D. Specific Aims

Specific Aims 1. To pinpoint predictors of Acute Chest Syndrome post-HCT, including patient demographics, disease severity, and transplant details.

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.

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. Given the rarity of ACS events relative to non-events, traditional analysis struggles to accurately identify risk predictors. Propensity score matching (PSM) offers a solution by matching patients on key characteristics, reducing confounding and isolating the effects of specific factors on ACS risk. This method mirrors the nuanced approach needed to understand biological systems where different factors dictate outcomes, such as the transition from fetal-type to adult-type hematopoietic systems affecting immune responses. By applying PSM, we aim to clarify how individual variations contribute to ACS risk post-HCT, enhancing our ability to personalize patient care and improve outcomes in sickle cell disease treatment.

Experimental Approach. To systematically identify significant predictors of ACS following HCT among patients with SCD, we will conduct a comprehensive analysis utilizing PSM to address the imbalance in our dataset. This approach will allow us to compare patients who developed ACS post-HCT with those who did not, across various patient-related factors (country, age group, sex, ethnicity, race group), disease-related factors (subtype of disease), and transplant-related factors (number of transplants, type of transplant, donor type, graft type, conditioning intensity, prophylaxis for graft-versus-host disease), as well as early post-transplant outcomes (veno-occlusive disease, graft failure, hemorrhagic cystitis).

- Data Preparation: Data will be standardized and cleaned to ensure consistency and accuracy.
- Propensity Score Calculation: For each patient, propensity scores will be estimated using logistic regression based on the aforementioned predictors. This score represents the probability of developing ACS given the patient's characteristics and transplant specifics.
- Matching Process: Patients who developed ACS post-HCT will be matched with those who did not, using nearest-neighbor matching within a specified caliper width to ensure balance across predictors. This method aims to create comparable groups, minimizing bias due to confounding factors.

- Statistical Analysis After Matching: After obtaining a matched cohort, we will analyze the data using logistic regression to identify which predictors significantly influence the occurrence of ACS post-HCT. This analysis will be adjusted for any residual imbalances to ensure robustness of findings.
- Sensitivity Analysis: Given the potential for unmeasured confounding in observational studies, we will perform sensitivity analyses to estimate the impact of such factors on our results, ensuring the reliability of our conclusions.

Interpretation of Results. The analysis results, interpreted through causal inference, will highlight the direct impact of specific predictors on the risk of ACS post-HCT. This interpretation will not only identify which factors significantly influence ACS risk but also guide the development of targeted interventions and monitoring strategies, informing clinical decisions and resource allocation for high-risk patients.

Potential Problems and Alternative Approaches. In utilizing PSM to identify predictors of ACS post-HCT within our highly imbalanced dataset, we anticipate challenges such as achieving balanced matches due to the scarcity of ACS cases and ensuring accurate model specification to avoid biased estimates. The inherent limitations of PSM, including its inability to account for unmeasured confounders, further complicate this approach. To mitigate these issues, we may explore alternative strategies like stratification or full matching for better balance, perform covariate adjustments post-matching to correct for residual imbalances, and conduct sensitivity analyses to gauge the impact of unmeasured confounding.

Specific Aims 2. To develop a risk stratification model that categorizes patients based on their likelihood of developing ACS after HCT

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.

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.

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.

The rationale behind choosing these two models is based on the fact that both Random Forest and XGBoost 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. 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 and XGBoost 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.

Interpretation of Results. We will use 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. This can inform clinicians about the key factors to monitor. Partial Dependence Plots (PDPs) can show the effect of a particular feature on the predicted outcome, holding all other features constant. This is helpful for understanding non-linear relationships and interactions between features and the target variable in a clear, visual format. Clinical interpretation also involves understanding how well the model performs, using metrics like AUC-ROC for overall discrimination ability, precision-recall curves for balance between sensitivity and precision, and calibration plots for reliability of predicted probabilities.

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. 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.

Potential Problems and Alternative Approaches. A primary challenge is the dataset's imbalance, with only 3% of samples being positive for ACS post-HCT, which could bias the model towards predicting the majority class. Additionally, there's a risk of overfitting due to the dataset's complexity. To address these issues, we will explore strategies 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.

References

Castro, Oswaldo, Donald J Brambilla, Bruce Thorington, Carl A Reindorf, Roland B Scott, Peter Gillette, Juan C Vera, and Paul S Levy. 1994. "The Acute Chest Syndrome in Sickle Cell Disease: Incidence and Risk Factors. The Cooperative Study of Sickle Cell Disease."

Paul, Rabindra N, Oswaldo L Castro, Anita Aggarwal, and Patricia A Oneal. 2011. "Acute Chest Syndrome: Sickle Cell Disease." *European Journal of Haematology* 87 (3): 191–207.

Vichinsky, Elliott P, Lynne D Neumayr, Ann N Earles, Roger Williams, Evelyne T Lennette, Deborah Dean, Bruce Nickerson, et al. 2000. "Causes and Outcomes of the Acute Chest Syndrome in Sickle Cell Disease." *New England Journal of Medicine* 342 (25): 1855–65.

Vichinsky, Elliott P, Lori A Styles, Linda H Colangelo, Elizabeth C Wright, Oswaldo Castro, Bruce Nickerson, and Cooperative Study of Sickle Cell Disease. 1997. "Acute Chest Syndrome in Sickle Cell Disease: Clinical Presentation and Course." *Blood, The Journal of the American Society of Hematology* 89 (5): 1787–92.