Decision Support App Development for Patients Receiving HCT

Han Yu, Lotus Liu, Xiai Mao, Xinyu Zhang, Yujin Han

February 2023

1 PROJECT NARRATIVE

Sickle-cell disease (SCD) is one of the most common severe monogenic disorders in the world [1]. It is a genetic blood disorder that affects the structure of hemoglobin, a protein found in red blood cells that carries oxygen throughout the body. People with SCD have abnormal hemoglobin molecules that cause their red blood cells to become stiff and sickle-shaped instead of flexible and round. These sickle-shaped cells can get stuck in small blood vessels, leading to reduced blood flow and oxygen delivery to the body's organs and tissues. This can cause a variety of health problems, including severe pain, anemia, organ damage, and increased susceptibility to infections. Early diagnosis and appropriate medical care are important for preventing and managing complications of SCD. Hematopoietic cell transplantation (HCT) is a well-established treatment to control many malignant and nonmalignant diseases [2], which has been used as a potential treatment for SCD. However, HCT also carries significant risks and potential adverse events. Some potential adverse events associated with HCT include Graft-versus-host disease (GVHD), Infections, Organ damage and Bleeding and clotting disorders.

2 PROJECT SUMMARY

In this project, we aim to develop a comprehensive decision support tool for SCD in the form of a Shiny app that utilizes machine learning techniques to provide personalized risk assessments for HCT treatment. This tool will provide clinicians with detailed information about patients and a comprehensive risk assessment of HCT treatment. Our overarching hypothesis is that machine learning methods can be used to provide personalized decision support for HCT treatment in SCD, which is both feasible and acceptable for use in clinical practice.

To test this hypothesis, we proposed three-level aims.

1. First, provide clinicians and patients with a detailed understanding of a certain patient group's characteristics by utilizing clustering and visualization techniques, as well as the development of several key disease indices over a two-year period.
2. Second, we aim to perform a comprehensive evaluation of patients who have received HCT treatment by survival analysis approach based on a single-group visualized KM curve. Statistical and descriptive information would also be displayed to provide clinicians with an understanding of the factors associated with patient outcomes following HCT.

3. Third, we aim to develop a comprehensive machine learning-based method to predict HCT treatment in patients and provide a report detailing the causes of the predicted risks to assist clinicians in their decision-making process. This machine learning algorithm would be trained on a dataset of SCD patients who have received HCT treatment, using a variety of demographic, clinical and HCT related features.

The resulting decision support tool would provide personalized risk assessments for HCT treatment in SCD patients, based on their individual characteristics and HCT related features. This tool has the potential to improve patient outcomes and quality of life by assisting clinicians in selecting appropriate candidates for HCT treatment and avoiding potential risks and complications. The development and implementation of this decision support tool could represent a significant advance in the management of SCD and the treatment of HCT.

3 SPECIFIC AIMS

3.1 Specific Aim 1:

We aim to develop an interactive portal that will provide healthcare providers and patients with a comprehensive decision-support tool for sickle cell disease (SCD). The Shiny app will be implemented to create a user-friendly platform that displays visualizations of the patient's demographics and characteristics of SCD-related symptoms.

To achieve this goal, we will leverage various visualization techniques such as bar graphs, line graphs, radar charts, and heat maps. These techniques will be used to display information about desired patient groups, allowing for a detailed understanding of patient characteristics. Additionally, we aim to show the progression of several SCD-related indices over a two-year period. This innovative platform can facilitate more effective communication between healthcare providers and patients, and enable clinicians and patients to make informed decisions about HCT treatment for SCD.

3.2 Specific Aim 2:

We aim to provide a comprehensive assessment of the patient's physical and psychological status after HCT treatment for a given sub-population.

Using the Kaplan-Meier curve and descriptive statistical results to demonstrate the rate at which various complications (physical and psychological dimensions) emerge after patients receive HCT. Knowing the rate at which different complications emerge in different groups can be very effective in suggesting appropriate recommendations for the disease in question in the short term and significantly improve prognostic outcomes. By comparing the rates of onset of different complications, we can also investigate in depth whether there is a certain causal relationship between HCT and the disease, thus providing a better understanding of the role of HCT.

Besides, we will try to use the non-proportional hazards methods such as the MaxCombo test or Restricted Median Survival Time difference test, to explore more about the causal effects for those complications.

3.3 Specific Aim 3:

We aim to develop a machine learning-based method to comprehensively predict the health risks of HCT treatment and report important factors for prediction to help clinicians in their decision making.

Specifically, we integrate different variables to creates new dependent variables which shows comprehensive health condition of patients after HCT treatment. We then consider the Conformal prediction (CP) methods [3] for prediction due to the following advantages: (1) CP methods allow clinicians to assess the confidence level of the predictions and avoid over-reliance on potentially incorrect results; (2) tree models-based CP methods have good model interpretability and can show important variables for prediction, providing predictive insight, increasing transparency and accountability; (3) CP methods have good generalization and remain robust under covariate shift and label shift setting, which is crucial complicated medical scenarios.

4 RESEARCH STRATEGY

4.1 SIGNIFICANCE

The objective of this project was to create a decision support tool in the form of a Shiny app for SCD, which provides comprehensive risk assessment and detailed information regarding HCT treatment for clinicians. By utilizing a personalized, predictive machine learning-based model based on registry-level datasets, this tool has the potential to guide clinical practice, enhance patient education, and facilitate shared decision-making for HCT in SCD. The development of this decision support app could significantly impact the care of SCD patients and improve clinical outcomes in HCT.

4.2 INNOVATION

Our project is innovative in utilizing the Shiny app platform to present the decision support tool for HCT treatment in SCD. The significance of this innovation lies in the numerous benefits of the Shiny app for clinicians. Firstly, the Shiny app provides a platform for clinicians to create personalized and customizable tools tailored to specific patient needs. This can lead to improved patient outcomes and reduced risk of adverse events. Secondly, the Shiny app can streamline clinical workflows and improve efficiency in clinical settings. This can save valuable time and resources for clinicians, enabling them to focus on patient care. Thirdly, the Shiny app allows clinicians to interact with real-time healthcare data, including patient information, which can help to identify patterns and trends in SCD datasets, informing clinical decision-making. Implementing our decision support tool on the Shiny app platform can significantly improve the care and outcomes of SCD patients. It represents a valuable contribution to the field of HCT treatment.

4.3 RESEARCH PLAN

4.3.1 Data Pre-processing

The date we are dealing with has 157 variables. After removing the variables that half of their values are missing, we got 47 variables. We selected and classified 45 variables into six categories, as shown in the tables in Appendix A.1.

4.3.2 Specific Aim 1:

- **Rationale**: Our goal is to utilize clustering and visualization techniques to provide clinicians and patients with a comprehensive understanding of SCD patients' characteristics. We will use unsupervised learning algorithms such as k-means and hierarchical clustering to group patients based on shared demographics such as age, gender, and race. In addition, we will develop disease indices related to SCD over a two-year period to track changes in patient' s health status. Disease indices serve as measures of disease severity or progression. The indices can help us identify potential risks and opportunities for improvement.
- **Literature Review**: We will conduct a detailed review of the existing literature on sickle cell disease and SCD complications. The study will involve research, including relevant articles, publications, and national public health websites that provide information on SCD. By conducting the literature review, we can gain insights into the current state of knowledge about SCD and identify areas where this interactive portal can be most helpful to healthcare providers and patients.
- **Visualization Techniques**: The resulting clusters will be displayed through various visualization techniques. We will first examine the desired characteristics to be displayed, including its format, scale, and complexity. We will examine the desired characteristics to be displayed and then determine the most effective visualization techniques to present the data. Based on these factors, we will determine which visualization techniques will most effectively present the data. For example, bar and line graphs are commonly used for displaying continuous data, while radar charts and heat maps may be more suitable for displaying categorical or discrete data.

4.3.3 Specific Aim 2:

- **Rationale**: The main objectives of survival analysis are to estimate the probability of the event occurring over time, identify factors associated with the event, and compare survival between different groups. Survival analysis methods can also be used to estimate the expected duration of survival, to model the hazard rate (i.e., the instantaneous rate of the event occurring at a specific point in time), and to evaluate the effectiveness of interventions or treatments in preventing or delaying the event of interest.
- **Experimental Approach and Visualization**: There are many time-to-event types in the data, where event refers to mental and physical complications. We plot a single Kaplan-Meier image (shown as Figure 2) for each comorbidity and analyze which comorbidities are more likely to occur in a given group by comparison. At the same time, we also monitor the survival rate of patients with various

complications at various time periods after HCT in real time, and summarize the information in the form of graphs to provide doctors with easier information organization and query.

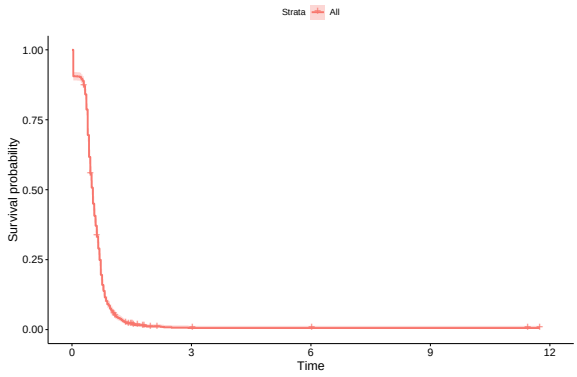

Figure 1: Survival probability of ANC after patient receiving HCT

Since all time-to-event outcomes use calendar time, it's easy to put the KM curves into one plot for simpler comparison.

Since we have stratified the model according to the covariates beforehand, the model may not be influenced much by the covariates. However, to get more accurate causal effect results, we can try some models designed for non-proportional risk in survival analysis, such as the MaxCombo test and the Restricted Median Survival Time difference model. Causal effect can be discussed as the secondary outcome in this section.

- **Interpretation of Results**: To interpret a single group Kaplan-Meier curve, we need to examine the shape of the curve and the location of any vertical drops (which indicate events such as death or disease progression). If the curve for one complication is relatively flat, it indicates a high probability of survival over time, while a steep drop indicates a high mortality rate or risk of disease progression. Through comparison of the Kaplan-Meier curves of all complications, we can identify several complications that are most likely to occur rapidly in the given interested group and intervene in a timely manner, which is very useful for visualizing time-to-event data and can provide valuable insights into the prognosis of patients with a particular disease or condition.

4.3.4 Specific Aim 3:

- **Variable integration**: To solve the problem of missing data and an excessive number of (outcome) variables, we integrate the post-HCT and outcome variables in Table.4 and Table.6 to construct new variables. For instance, variables AGVHD (Acute graft-vs-host disease) and CGVHD (Chronic graft-vs-host disease) can be integrated as a binary variable reflecting graft-vs-host disease, GVH, which means GVH=AGVHD *OR* CGVHD. We also predict other health indicators after HCT, such as the stroke, acute chest syndrome and secondary malignancy. All of these responses are binary or multi-class variables

- **Conformal prediction (CP)**:CP combines the ideas of prediction and hypothesis testing. Assuming only data exchangeability and n random observations, conformal prediction predicts a set of labels $\hat{C}(x_{n+1})$ for a new observation x_{n+1}, and the probability that the set $\hat{C}(x_{n+1})$ contains the true

label Y_{n+1} is $(1-\alpha)$, where α is an artificially set confidence level. It means given the confidence level α, CP achieves marginal coverage that $P(Y_{n+1} \in \hat{C}_{n,\alpha}(x_{n+1})) \geq 1-\alpha$. We consider BCOPS [4] model, which shows promising results on the multi-class classification problem, especially when the training data and the out-of-sample test data may have different distributions. In addition, BCOPS uses random forest [5] as base learners, which facilitates us to calculate feature importances. Specifically, we consider using the Gini index [6] to measure the contribution of each feature as a variable importance measure. Feature importance helps clinicians find variables that are important for prediction, thus providing more supports for decision making.

References

[1] "Sickle-cell disease". In: *The Lancet* 376.9757 (2010), pp. 2018–2031.

[2] "Current Use of and Trends in Hematopoietic Cell Transplantation in the United States". In: *Biology of Blood and Marrow Transplantation* 26.8 (2020), e177–e182.

[3] Glenn Shafer and Vladimir Vovk. "A Tutorial on Conformal Prediction." In: *Journal of Machine Learning Research* 9.3 (2008).

[4] Leying Guan and Robert Tibshirani. "Prediction and outlier detection in classification problems". In: *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 84.2 (2022), p. 524.

[5] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[6] Bin Li, J Friedman, R Olshen, and C Stone. "Classification and regression trees (CART)". In: *Biometrics* 40.3 (1984), pp. 358–361.

A Appendix

A.1 Variable Description

Table 1: Patient-demography variables

Variable	Explaination
DUMMYID	Dummy CIBMTR Recipient ID
SEX	Sex
ETHNICIT	Ethnicity
COUNTRY	Country of HCT institution
AGE	Patient age at transplant, years
AGEGPFF	Patient age at transplant, years (grouped)
RACEG	Race(regrouped)

Table 2: Patient-health condition variables

Variable	Explaination
SUBDIS1F	Disease genotype
RCMVPR	Recipient CMV serostatus
KPS	Karnofsky/Lansky score at HCT

Table 3: HCT-related variables

Variable	Explaination
TXNUM	Transplant number
TXTYPE	Transplant type
DONORF	Donor type
GRAFTYPE	Graft type
YEARTX	Year of transplant
YEARGPF	Year of transplant (grouped)
HLA_FINAL	Donor-recipient HLA matching
CONDGRPF	Conditioning intensity
CONDGRP_FINAL	Conditioning regimen

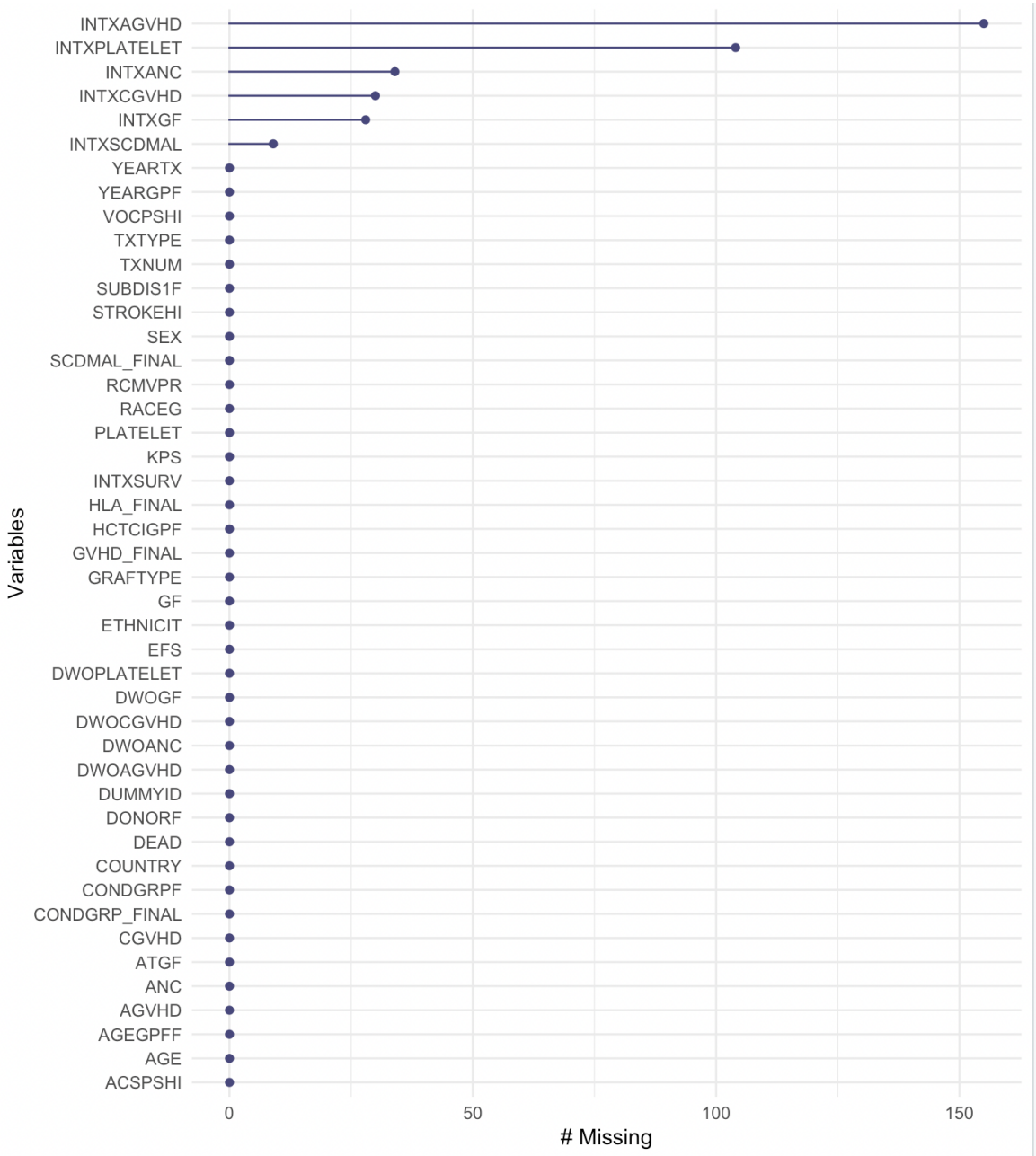

Figure 2: Missing values of 45 selected variables

Table 4: Post-HCT complication variables

Variable	Explaination
STROKEHI	Stroke post HCT
ACSPSHI	Acute chest syndrome post HCT
VOCPSHI	Vaso-occlusive pain post HCT
AGVHD	Acute graft versus host disease, grades II-IV
INTXAGVHD	Time from HCT to acute graft-vs-host disease, months
CGVHD	Chronic graft-vs-host disease
INTXCGVHD	Time from HCT to chronic graft-vs-host disease, months
SCDMAL_FINAL	Secondary malignancy
INTXSCDMAL	Time from HCT to second malignancy, months

Table 5: Prophylaxis-related variables

Variable	Explaination
ATGF	A TG/Alemtuzumab given as conditioning regimen/GVHD prophylaxis
GVHD_FINAL	GVHD prophylaxis

Table 6: HCT efficacy (HCT outcome) variables

Variable	Explaination
ANC	Neutrophil engraftment
INTXANC	Time from HCT to neutrophil engraftment, months
PLA TELET	Platelet recovery
INTXPLA TELET	Time from HCT to platelet recovery, months
GF	Graft failure
INTXGF	Time from HCT to graft failure, months
EFS	Event-free survival (Graft failure or death are the events)
DWOAGVHD	Death without acute graft versus host disease, grades II-IV
DWOCGVHD	Death without chronic graft-vs-host disease
DWOANC	Death without neutrophil engraftment
DWOPLA TELET	Death without platelet recovery
DWOGF	Death without graft failure
INTXSURV	Time from HCT to date of last contact or death, months
DEAD	Survival status at last contact
HCTCIGPF	HCT-comorbidity index