PROJECT SUMMARY/ABSTRACT

This project aims to leverage the comprehensive research database established by the Center for International Blood and Marrow Transplant Research® (CIBMTR) for hematopoietic cell transplantation (HCT). The project is divided into three specific aims: a descriptive analysis of enrolled patients to understand demographic and clinical characteristics, a survival analysis focusing on the time from HCT to seven different endpoints, and the development of an R Shiny application for dynamic and interactive visualization of study results. This initiative seeks to enhance understanding of HCT outcomes, identify factors influencing survival post-transplant, and facilitate data accessibility for clinicians and researchers alike, ultimately contributing to improved patient care and outcomes in the field of cellular therapies.

PROJECT NARRATIVE

Sickle cell disease (SCD) is a group of inherited red blood cell disorders affecting millions of people throughout the world. In someone with Sickle Cell Disease (SCD), their red blood cells contain abnormal hemoglobin, causing them to become misshapen, resembling a sickle. These cells die prematurely, constantly decreasing the body's red blood cell count. They can also block blood flow in small vessels, triggering pain, and severe complications like infections, acute chest syndrome, and stroke. Despite a variety of strategies including supportive care, drug therapies, and red blood cell transfusions are able to potentially alleviate symptoms and extend lifespan of SGD patients, allogeneic hematopoietic cell transplantation (HCT) is the only established potential cure for SCD. Nevertheless, post-HCT SCD patients continue to face severe health challenges because HCT is associated with life-threatening complications most of which occur within the first 2 years after transplantation such as graft-versus-host disease (GVHD) and mortality.

SPECFIC AIMS

A. Exploratory Data Analysis

We will perform exploratory analyses on variables included in the datasets to obtain an overview of the baseline characteristics of patients who received HCT. Specifically, the dataset categorizes the variables into 3 categories: disease-related, patient-related, and transplant-related. By summarizing information on variables in the dataset, we expect to learn the distribution of baseline variables that could facilitate further processing of the data, and provide initial insights on identifying important predictors contributing to predicting the clinical outcome e of patients.

B. Survival Analysis

We will perform a survival analysis to assess the time from HCT to 8 endpoints, including last contact or death, graft failure, neutrophil engraftment, platelet recovery, acute and chronic graft-vs-host disease (GVHD), post-transplant lymphoproliferative disorder (PTLD), and second malignancy. These analyses will help identify key predictors of outcomes and potential factors that could enhance the quality of care for patients undergoing HCT treatment.

C. Shiny App Development

To make the findings accessible and actionable, an R Shiny application will be developed. This interactive tool will allow users to explore the data dynamically, visualize survival curves, compare outcomes across different patient groups, and potentially identify new areas for research or intervention. In order to enhance the understanding and management of Sickle Cell Disease (SCD) post-hematopoietic cell transplantation (HCT), we propose the development of an R Shiny application, a crucial tool designed to visualize and interpret complex datasets.

RESEARCH STRATEGY

A. Significance

Recognizing the significance of HCT regarding the curation of SGD and the post-HCT potential health obstacles, our team aims to explore crucial patient-related, disease-related, and transplant-related factors contributing to clinical outcomes after HCT among SGD patients by utilizing the database from Center for International Blood and Marrow Transplant Research (CIBMTR) in 2021. The project initiative is to specify the granularity for both evaluating success and subsequent health threats post-HCT and assist clinical professionals with the decision on establishing a more comprehensive treatment plan for SGD patients by taking key patient-related, disease-related, and transplant-related features into consideration.

B. Innovation

- Analysis Innovation

  For the survival analysis of this project, the targeted time-to-events include not only outcomes with implications of HCT failure such as "time to GVHD" and "time to PTLD(Post-transplant lymphoproliferative diseases)" but also implications of HCT success such as "time to neutrophil engraftment" and "time to platelet recovery". In this way, we are able to have a more comprehensive understanding to assess the HCT performance for SGD patients

- Technology Innovation

  In this project, we plan to create a standardized and interactive visualization platform to illustrate the analysis results via an R shiny web application. Besides the ease of exploring and engaging with research outcomes, we also recognize the advantage of an R shiny web application to serve as a standardized tool for assisting future researchers to expand upon this project by feeding more recent data to this platform and obtaining more updated results.

C. Data Pre-Processing

After comparing the dataset from 2020 to 2021, our team has decided to focus on the dataset from 2021 due to its enhanced data completeness and inclusiveness of the dataset from 2020. To ensure the quality of analysis, columns for variables excluding outcomes with over 20% missing data were removed. In addition, variables with unchanging values were disregarded. In the end, we have the following numbers of variables in addition to patient ID (see Appendix):

- 11 patient-related variables
- 1 disease-related variable
- 10 transplant-related variables
- 8 outcome variables/endpoints

D. Specific Aim 1: Descriptive Analysis of Enrolled Patients

Hypothesis: The exploratory data analysis aims to characterize the distribution of baseline variables and to initially assess associations between the distribution of baseline characteristics and the occurrence of clinical endpoint events. Specifically, for each categorical variable, we will calculate the count and proportion of each value level, and for continuous variables, we will calculate the mean and standard deviation. We will then examine the association between occurrence of clinical endpoints (all-purpose death, signs of recovery) and baseline variables using Logistic regression, and examine the associations between explanatory variables by either Pearson's Chi-squared test (categorical VS categorical), linear regression (numerical VS numerical), and t-test or Wilcoxen test (numerical VS categorical).

Rationale: Conducting exploratory analyses allows us to understand the structure and the pattern in the data, and thus facilitates the process of identification of potentially unbalanced variables and the decision whether to keep these variables in downstream analyses. By testing associations between outcome variables and response variables, we could gain initial insights into characteristics that might be clinically relevant in HCT patients.

Experimental Approach: Summary tables will be created by processing the data with R package `dplyr` and presented with the package `gt`. When examining associations between explanatory variables, we will implement Pearson's Chi-squared test by the R function `chisq.test`, linear regression by R function `lm` , t test by `t.test` and Wilcoxen's test by `wilcox.test`. Highly associated explanatory variables will be further examined and the decision on exclusion of one of the highly associated variables will be made. To fit Logistic regression to test the association between explanatory variables and outcome variables, the R function `glm` will be used.

Interpretation of Results: From the model fit, we could extract predictors that are associated with significant p-values with a pre-specified significance level (0.05), and consider them as important predictors for the occurrence of specific clinical outcome.

Potential Problems and Alternative Approaches: The results of the Logistic regression could be biased by unbalanced distribution of predictors, and this could be mitigated by combining rare categories.

E. Specific Aim 2: Survival Analysis

Hypothesis: The time from hematopoietic cell transplantation (HCT) to various clinical endpoints (last contact or death, graft failure, neutrophil engraftment, platelet recovery, acute and chronic graft-vs-host disease (GVHD), post-transplant lymphoproliferative disorder (PTLD), and second malignancy) significantly differ among patient groups defined by specific demographic and clinical characteristics.

Rationale: Understanding the time to key post-HCT outcomes is crucial for several reasons.

- Firstly, it enables the prediction of patient prognosis by providing estimates on when significant post-transplant events might occur, thus informing both patients and clinicians about possible recovery paths and any potential complications that could arise. This knowledge is invaluable in setting realistic expectations and preparing for any necessary interventions.

- Secondly, insights gained from analyzing the timing of events such as graft-versus-host disease (GVHD) are instrumental in guiding clinical decisions. They allow healthcare providers to customize post-transplant care according to the specific needs of each patient, potentially mitigating risks and enhancing recovery.

- Lastly, by identifying the factors that influence the speed of recovery or the occurrence of delayed complications, adjustments can be made to the transplant process itself. This could involve selecting different donors, modifying conditioning regimens, or altering supportive care practices, all aimed at improving the overall success and safety of the transplant procedure.

Experimental Approach:

- Cox Proportional Hazards (CoxPH) Model: To adjust for potential confounders and assess the impact of various predictors on the time to each endpoint, CoxPH models will be fitted. A test of the proportional hazards assumption can be based on testing whether there is a linear relationship between the Schoenfeld residuals and some function of time. Variables will include patient demographics, disease characteristics, transplant-related factors, and pre-transplant comorbidities. We can conduct variable selection by analyzing the log-likelihood and corresponding p-values calculated using the `anova` function in the survival package for the specific model fit.

- Kaplan-Meier (KM) Survival Curves: For each of the eight endpoints, KM survival curves will be generated to visualize the unadjusted probability of reaching each endpoint over time post-HCT.

Interpretation of Results: In our analysis using the Cox Proportional Hazards model, we anticipate identifying several key variables that significantly influence the various endpoints of interest post-HCT, such as time to graft failure, neutrophil engraftment, platelet recovery, and the onset of acute and chronic graft-vs-host disease, among others. These variables could range from patient-related factors (e.g., age, gender, underlying disease type, and stage), to treatment-related factors (e.g., type of donor, conditioning regimen intensity), to post-transplant care aspects (e.g., immunosuppression protocol, infection prophylaxis).
Once these influential variables are identified, we plan to conduct a stratification analysis to categorize patients into groups based on their characteristics. For example, patients might be grouped by age range, disease stage at transplant, or by the type of donor (related vs. unrelated) to see how these factors affect outcomes differently.
Following the stratification, we will plot Kaplan-Meier (KM) Survival Curves for each group. These plots will visually represent the survival probabilities over time for each stratified patient group, allowing us to observe any significant differences in survival rates among the groups. For instance, we may see distinct survival curves for younger vs. older patients or for those receiving transplants from matched unrelated donors vs. sibling donors.
The expectation is that these stratified KM Survival Curves will reveal significant disparities in survival times across different patient groups, highlighting the impact of the identified variables on patient outcomes.

Potential Problems and Alternative Approaches: In tackling the survival analysis for our study, we may encounter several challenges that necessitate alternative approaches. One such challenge is the violation of the proportional hazards assumption, a cornerstone of many survival analysis methods. Should this occur, we would pivot to employing time-varying covariate models or consider other survival analysis techniques, such as accelerated failure time models, which do not rely on this assumption. Another potential issue could arise from having sparse data for certain outcomes, which might reduce the statistical power of our analyses. In these instances, we might need to aggregate similar categories or focus our analysis on the most clinically relevant outcomes to ensure meaningful results. Furthermore, the identification of complex interactions among predictors could complicate the model's interpretability and accuracy. To address this, machine learning methods like random forests or survival trees could be employed, offering a more nuanced capture of these intricate relationships. These approaches not only help us navigate potential pitfalls but also enhance the robustness and relevance of our findings, ensuring they contribute effectively to the field.

F. Specific Aim 3: Development of an R Shiny Application for Results Visualization

This application will serve as an interactive platform for clinicians, researchers, and policy-makers, enabling them to dynamically explore and analyze the wealth of patient-related, disease-related, and transplant-related variables and their impact on post-HCT outcomes.

Rationale: The rationale behind this application is twofold. Firstly, given the complexity and multi-dimensionality of the data, traditional static methods of data presentation are insufficient for capturing the nuanced relationships between variables such as age, ethnicity, disease genotype, transplant type, and conditioning regimen. An interactive tool will allow for a more comprehensive and tailored exploration of these variables, fostering a deeper understanding of their interplay and impact on patient outcomes. Secondly, this

application aims to democratize data access, allowing users to generate custom analyses and visualizations that can inform clinical decision-making and policy development.

Experimental Approach: The app will be implemented using R Shiny, leveraging its capabilities for creating interactive, web-based data visualizations. Key features will include the ability to filter and stratify data based on specific variables like age group, disease genotype, or transplant type, and visualize these in the form of survival curves, bar charts, and heatmaps. For instance, users could compare survival outcomes between different age groups or analyze the impact of donor-recipient HLA matching on post-transplant complications. This approach will enable users to interact with the data, providing immediate visual feedback and insights.

Interpretation of Results: The results presented through the Shiny app will offer valuable insights into the factors that influence post-HCT outcomes in SCD patients. By enabling the exploration of complex datasets in an intuitive manner, the app will not only aid in the interpretation of current research findings but also spark new hypotheses and areas for further study. It will facilitate a more nuanced understanding of the SCD patient journey post-HCT, contributing significantly to the field of hematology and transplant medicine.

F.1. Appendix

Table 1: List of Predictors

Category	Variable.name	Description
Disease-related	subdis1f	Disease genotype
Patient-related	Dummyid	Unique patient identifier
Patient-related	flag_lancet	Cases from 2019 Lancet Heamatology publication
Patient-related	flag_blood	Cases from 2016 Blood publication
Patient-related	flag_0601	Cases from BMT CTN 0601
Patient-related	age	Patient age at transplant, years
Patient-related	agegpff	Patient age at transplant, years
Patient-related	sex	Sex
Patient-related	ethnicit	Ethnicity
Patient-related	kps	Karnofsky/Lansky score at HCT
Patient-related	hctcigpf	HCT-comorbidity index
Tranplant-related	donorf	Donor type
Tranplant-related	graftype	Graft type
Tranplant-related	condgrpf	Conditioning intensity
Tranplant-related	condgrp_final	Conditioning regimen
Tranplant-related	atgf	ATG/Alemtuzumab given as conditioning regimen/GVHD prophylaxis
Tranplant-related	gvhd_final	GVHD prophylaxis
Tranplant-related	hla_final	Donor-recipient HLA matching
Tranplant-related	rcmvpr	Recipient CMV serostatus
Tranplant-related	yeargpf	Year of transplant
Tranplant-related	yeartx	Year of transplant

Table 2: List of Time-to-Event Variables

Category	Variable.name	Description
Outcomes	intxsurv	Time from HCT to date of last contact or death, months
Outcomes	intxgf	Time from HCT to graft failure, months
Outcomes	intxanc	Time from HCT to neutrophil engraftment, months
Outcomes	intxplatelet	Time from HCT to platelet recovery, months
Outcomes	intxagvhd	Time from HCT to acute graft-vs-host disease, months
Outcomes	intxcgvhd	Time from HCT to chronic graft-vs-host disease, months
Outcomes	intxptld	Time from HCT to PTLD, months
Outcomes	intxscdmal	Time from HCT to second malignancy, months

Table 3: List of Outcome Varaibles (Exclude Time-to-Event)

Category	Variable.name	Description
Outcomes	dead	Survival status at last contact
Outcomes	efs	Event-free survival (Graft failure or death are the events)
Outcomes	gf	Graft failure
Outcomes	dwogf	Death without graft failure
Outcomes	anc	Neutrophil engraftment
Outcomes	dwoanc	Death without neutrophil engraftment
Outcomes	platelet	Platelet recovery
Outcomes	dwoplatelet	Death without platelet recovery
Outcomes	agvhd	Acute graft versus host disease, grades II-IV
Outcomes	dwoagvhd	Death without acute graft versus host disease, grades II-IV
Outcomes	cgvhd	Chronic graft-vs-host disease
Outcomes	dwocgvhd	Death without chronic graft-vs-host disease
Outcomes	ptld	Post-transplant lymphoproliferative disorder (PTLD)
Outcomes	scdmal_final	Secondary malignancy

References

Sickle cell disease – outcomes & advances. https://bethematchclinical.org/transplant-indications-and-outcomes/disease-specific-indications-and-outcomes/sickle-cell-disease/, n.d.

Centers for Disease Control and Prevention. Data & statistics on sickle cell disease. https://www.cdc.gov/ncbddd/sicklecell/data.html, 2023a. Accessed: July 6, 2023.

Centers for Disease Control and Prevention. What is sickle cell disease? https://www.cdc.gov/ncbddd/sicklecell/facts.html, 2023b. Accessed: July 6, 2023.

D.-P. Chen, Y.-H. Wen, P.-N. Wang, A.-L. Hour, W.-T. Lin, F.-P. Hsu, and W.-T. Wang. The adverse events of haematopoietic stem cell transplantation are associated with gene polymorphism within human leukocyte antigen region. *Scientific Reports*, 11(1), 2021. doi: 10.1038/s41598-020-79369-w. URL https://doi.org/10.1038/s41598-020-79369-w.

D. Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC, 4 edition, 2023. doi: 10.1201/9781003282525. URL https://doi.org/10.1201/9781003282525.

A. St. Martin, K. M. Hebert, A. Serret-Larmande, V. Jouhet, E. Hughes, J. Stedman, T. DeSain, D. Pillion, J. C. Lyons, P. Steinert, P. Avillach, and M. Eapen. Long-term survival after hematopoietic cell transplant for sickle cell disease compared to the united states population. *Transplantation and Cellular Therapy*, 28(6), 2022. doi: 10.1016/j.jtct.2022.03.014. URL https://doi.org/10.1016/j.jtct.2022.03.014.