BIOS 611 Project Report

Analysis of Children Bone Marrow Transplant Data set

Wanting Jin

Github Repository

October 27, 2022

Abstract

1 Introduction

1.1 Background

Allogeneic hematopoietic stem cell transplantation (HSCT) is a standard therapy for children with a variety of both malignant (i.a. acute lymphoblastic leukemia, acute myelogenous leukemia) and non-malignant diseases (i.a. severe aplastic anemia, Fanconi anemia). It is of significant importance to find matched donor such that ensure success outcome and reduce the risk of complications like graft-versus-host disease after transplantation procedure. In gereral, Human Leukocyte Antigens (HLA) Matching is conducted to find out if hematopoietic stem cells match between the donor and the patient receiving the transplant. However, due to the difficulty and time consuming to find a fully matched donor, it is acceptable to use unrelated donors mismatched at 1-2 HLA alleles. Previous research also proposed that increased dosage of CD34+ cells / kg extends overall survival time without simultaneous occurrence of undesirable events affecting patients' quality of life (Kawak et al., 2010).

1.2 Aims

1.3 Data Description

This data set describes pediatric patients with several hematologic diseases grouped into malignant disorders and nonmalignant. All patients were subject to the unmanipulated allogeneic unrelated donor hematopoietic stem cell transplantation. A total of 187 patients are included in this data set. And 39 attributes about donor and recipient's matching properties and their survival outcomes after transplantation are recorded.

1.4 Exploratory Visualizations

Prior to carrying out any model building or prediction on the heart disease data, we are interested in investigating some exploratory figures to gain an understanding of some patterns that exist in the data and if there are any particularly interesting factors to consider. In Figure 1, Pearson's Correlation among all variables in the original dataset are shown in the heatmap.

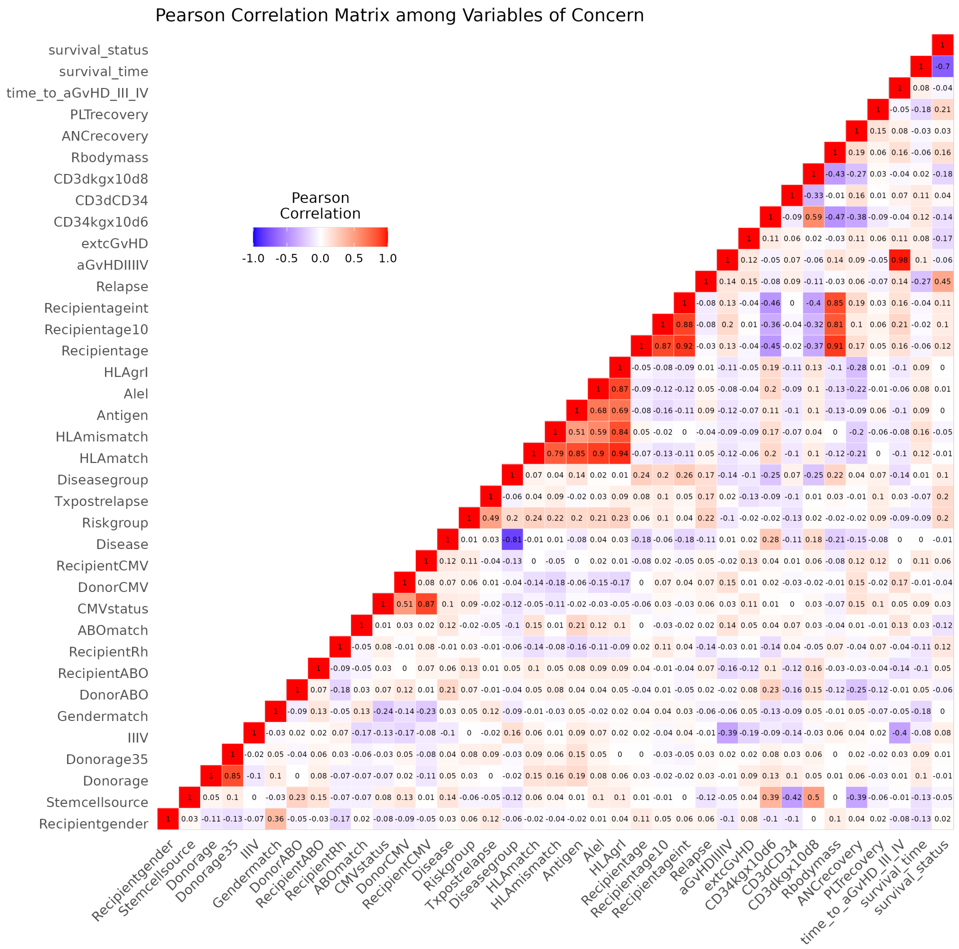

Figure 1: Heatmap Plot of Pearson Correlations among covariates of interests.

2 Methods

2.1 Data Preprocessing

2.2 Logistic Regression

To address the classification problem, one simple way is to use logistic regression. In the logistic regression model, we assume that,

$$y_i|\boldsymbol{x}_i \sim Bernoulli(\pi_i),$$

and that,

$$\text{logit}(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}) = \boldsymbol{x}_i^T\boldsymbol{\beta},$$

for $i = 1, \cdots, n$. Then the log-likelihood function for solving β can be expressed as

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - \log(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) \right\}$$

We can solve for $\boldsymbol{\beta}$ in this expression by maximizing the log-likelihood function.

3 Results

4 Conclusion

5 Discussion

5.1 Limitations & Concerns

References