Exploratory Data Analysis Report for Bone Marrow Transplant: Children

Rudraharsh Tewary

November 3, 2023

1 Summary of Work Done

The primary goal of my project is to investigate whether the concentration of CD34+ cells contained in bone marrow transplants has a statistically significant impact on the long term survival metrics of pediatric recipients part of my chosen dataset [8]. In this section, we will move to discuss what has been done previously in support of and against this hypothesis. Note that as the dataset is relatively new (2020), a lot of work that has been cited did not use the exact same dataset, but they did study the same phenomena in pediatric and/or adult patients.

Our first reference [3] discusses the case of Autologous Hematopoietic Stem Cell Transplant in pediatric patients, particularly for the treatment of Central Nervous System Tumours(CNSTs), wherein they investigate whether there is an impact of CD34+ dosage/concentration on the Overall Survival (OS), Progression Free Survival (PFS) and Relapse rate. Their research concluded that above an optimal level of CD34+ concentration $> 3.6 \times 10^6/kg$, the patients had significantly better PFS and OS, with also overall lower Relapse rates as compared to patients with concentration $\leq 3.6 \times 10^6/kg$ of CD34+ cells. Similarly, another study [5] showed that even in adult patients, higher CD34+ dosage $\geq 4.2 \times 10^6/kg$ had decreased relapse rates.

Our next reference [6] investigated the impact of CD34+ dosage on Platelet Engraftment and Neutrophil Engraftment on receivers of Allogeneic Hematopoietic Stem Cell Transplant(AHsct), a difference being the optimal cut-off value chosen by them, with the dosage $> 5 \times 10^6/kg$ of CD34+ being classed as high dose and $\leq 5 \times 10^6/kg$ being classed as low dosage. Even here, we find support for our hypothesis as the researchers reported that patients who received high dosage of CD34+ had shorter times to Neutrophil Engraftment and Platelet Engraftment as well as lower risk of developing Graft vs Host Disease. The case for AHsct in case of pediatric patients is covered by [1] where in they find that an increased CD34+ dosage $7.47 \times 10^6/kg$ showed a marked improvement in OS, Transplant Related Mortality (TRM),and Donor Chimerism (DC).

The following work [2] shows that in the case of pediatric patients receiving AHscts for the treatment of hematological disorders, Higher dose of CD34+ $> 2.42 \times 10^6/kg$ resulted in sped up Platelet Engraftment. Similarly, the following case study [7] also supports our hypothesis. Finally, we have one article which argues that CD34+ concentration has no impact on transplant success in the treatment of Neuroblastoma [4]

2 Data Description

Our Dataset [8]has been collected from the UC Irvine Machine Learning Repository, Titled 'Bone Marrow Transplants: Children'. It contains pediatric patients with a variety of Malignant Disorders (acute lymphoblastic leukemia, acute myelogenous leukemia, chronic myelogenous leukemia, myelodysplastic syndrome) and Non-Malignant Disorders (severe aplastic anemia, Fanconi anemia, with X-linked adrenoleukodystrophy). All patients in the dataset were treated with Unrelated Allogeneic Hematopoietic Stem Cell Transplants. We have a total of 187 observations with 37 features for each observation. There are 12 continuous features and 25 discrete features in the dataset. A table containing the name, type and description of each feature has been included in the report and is a part of the Appendix.

3 Data Issues

Upon exploration of the dataset, multiple issues were encountered. The first and foremost being the presence of missing values, especially in categorical columns like CMVstatus,DonorCMV,RecipientCMV. However, in most observations where one of these 3 variables has a missing entry, one or both of the remaining variables have a recorded entry, so, concerning the small size of the dataset, it would be possible to manually impute the missing value based on other observations. Another significant column with missing values is 'extcGvHD', which signals development of extensive chronic graft vs host disease. I will undertake a more optimistic approach for this variable, Wherein if the value for extcGvHD is missing, then I would replace it with 'No', i.e. extcGvHD was not seen.

Apart from missing values, the authors have also made some questionable choices in the treatment of Binary variables. For example, There is a significant amount of binary features which have 'Yes' or 'No' as their levels. However, the notation is switched for certain variables. If the standard is 0 for 'No' and 1 for 'Yes'. Then it is followed for few variables and discarded for others, for example variable 'IIIV' has 1 for 'Yes', 0 for 'No'. On the other hand, variable 'aGvHDIIIV' has 0 for 'Yes', 1 for 'No'. However, this can be fixed simply through a few updates.

Proceeding on, some variables which should have a categorical component to them, have that missing, which causes a skew to appear. An example would be the variable 'PLTrecovery', which tells us about the time in days to platelet recovery. However, there are some patients who never recovered to baseline levels. Now, as the variable is recorded as a type 'Integer' the authors decided to put in the value of days as '1000000' for those patients who never recovered. Same issue is also prevalent for another variable 'ANCrecovery'. My approach to fix this would be to create a new factor variable, which encodes whether the patients recovered to baseline levels or not.

Finally, Overall the number of observations are lesser than desired. It is usually said that ideally each feature should have 10 observations, making it so that we should have at-least 370 observations to make stronger claims. However, we only have close to half of the required amount at 187 observations.

4 Exploratory Data Analysis

We begin our Exploratory Data Analysis by taking an overview plot which shows us the number of discrete and continuous columns, as well as the number of complete and incomplete rows along with the percentage of missing observations

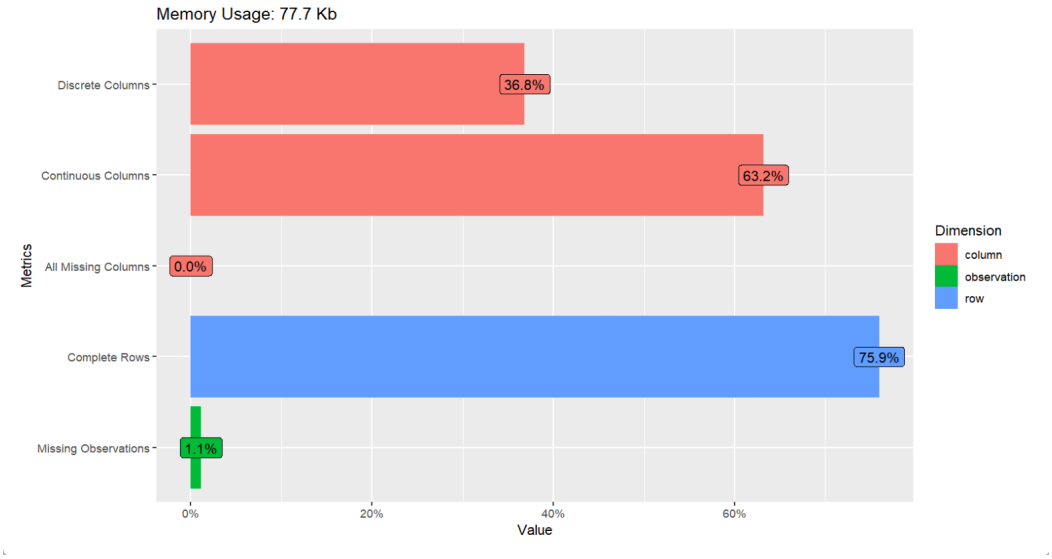

Figure 1: Overview plot of the Dataset

Now, we will proceed to generate a plot which orders columns by their share of missing values

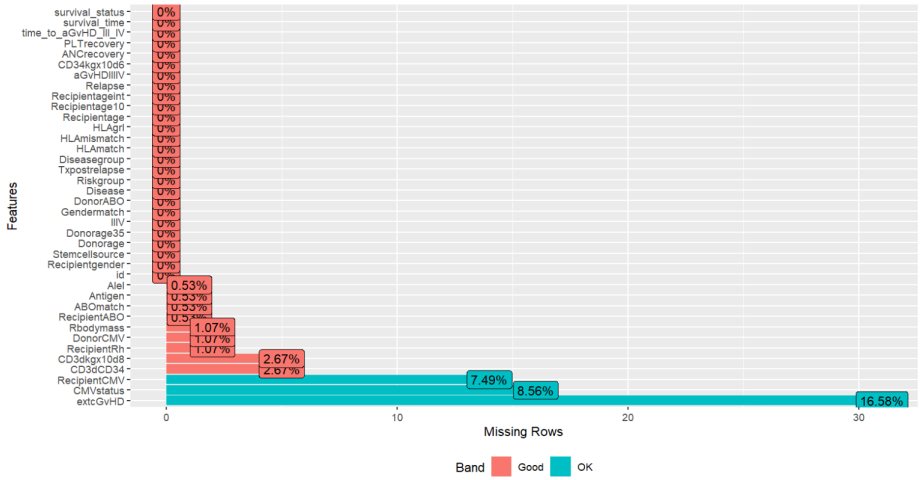

Figure 2: Plot of columns by Percentage of missing values

We see that most columns have very low amounts of missing values, the only culprits being the columns 'RecipientCMV','CMVstatus' and 'exctGvHd' each having 7.5%,8.5%,and 16.5% of missing values respectively. Still overall, we have a decently workable set of observations and as discussed earlier the missing value number may be further reduced through manual imputation.

We now move on to look at the plots of the discrete/categorical variables:

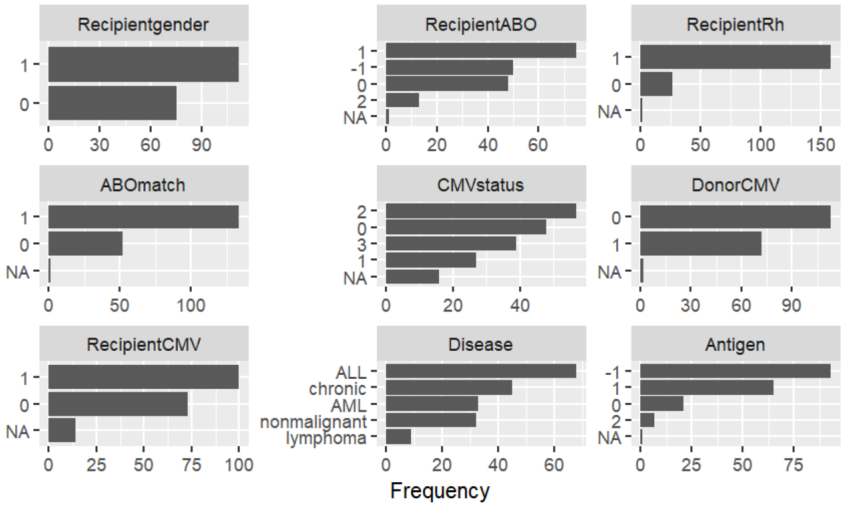

Figure 3: value frequency plot of categorical variables

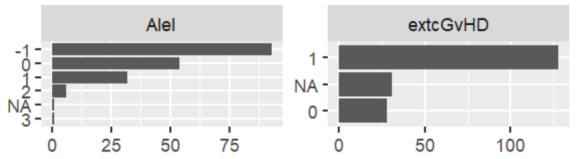

Figure 4: Continuation of value frequency plot

These are the bar plots of a subset of our categorical variables in the dataset. One thing that we can notice from this plot is that a big majority of our dataset, i.e. patients are suffering from malignant disorders. This will impact our ability to make claims about our selected feature 'CD34+ concentration"s impact on transplant outcomes for non-malignant cases. Another noticeable factor is that the Gender of our patients is majority male instead of an even split.

We will now move on to observe the distribution of our continuous variables Right off the bat,

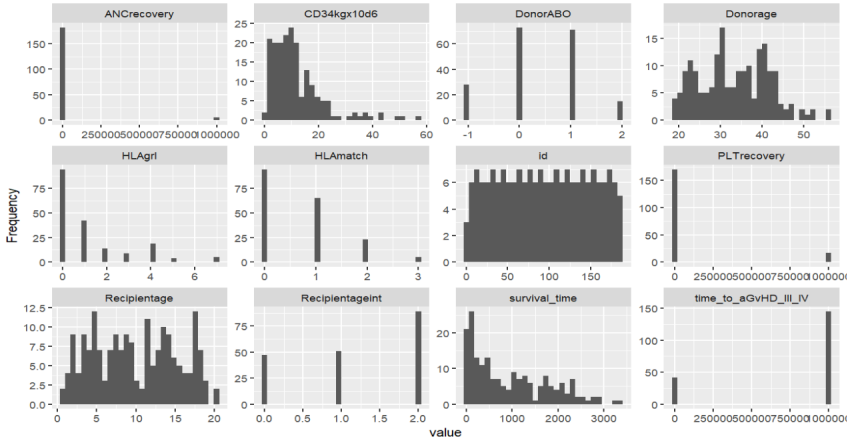

Figure 5: Distribution of continuous variables

we notice a major issue. Which is the presence of '1000000' as an observation in the variables 'ANCrecovery','PLTrecovery', and 'time-to-aGvHD-III-IV' we move to adjust their graphs by removing the max observation as discussed in the issues section Now, we can move to make observations

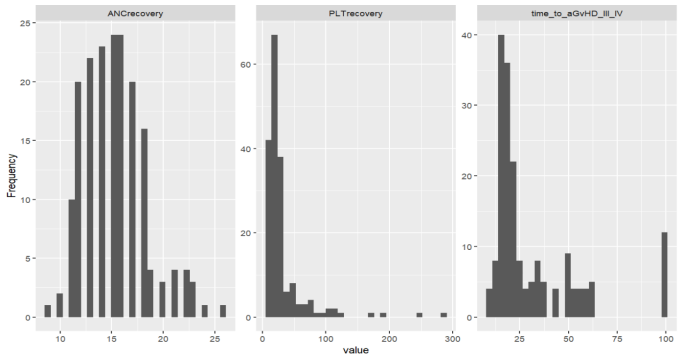

Figure 6: Distribution of continuous variables adjusted

on our obtained data. We see that on average it takes about 2 weeks for our patients to make a recovery in neutrophil count (ANCrecovery), while platelet engraftment takes around a month (PLTrecovery). We also see that even for those who survive, the maximum survival time of patients is around 8 years post transplant.

Now, we move to analyze the relationships between our features, one way to do that is by generating a correlation plot of our features Looking at the heatmap, we can first identify that there is a clus-

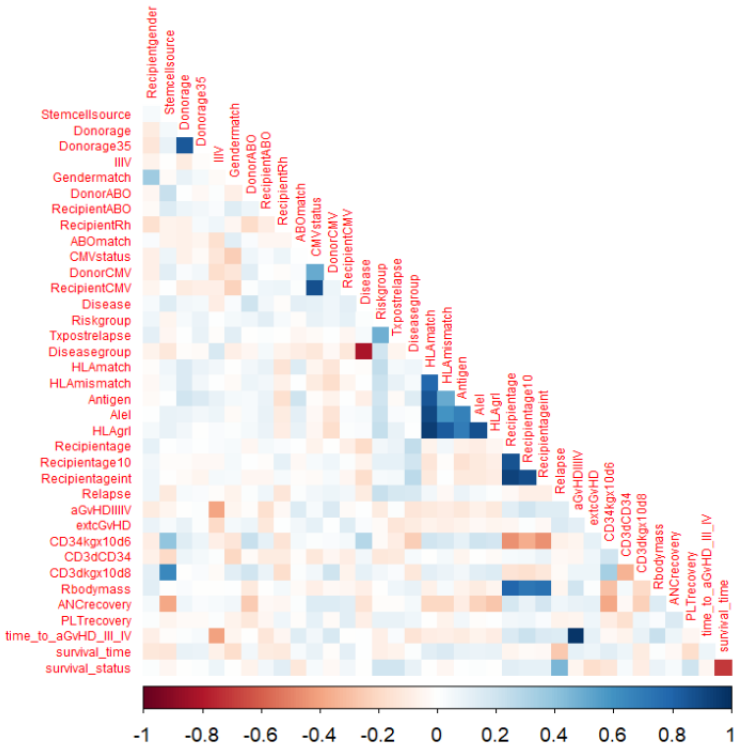

Figure 7: Correlaiton heatmap of our features

ter of strong positive correlation around the variables 'HLAmatch','HLAmismatch','Alel','Antigen' and 'HLAgrl'. However, this is to be expected as the 'HLA'series features are directly dependent on the antigens and alleles(alel) as the HLAgrl/HLAmatch/HLAmismatch variables are used to check differences between antigens and allels to measure donor compatibility. However, in our case, it can be said that few of these variables can be eliminated during model design as the information carried by one is nearly preserved in the other feature.
Though, overall we see that there is weak correlation between most variables be it negative or positive. Which gives us a hint that any model created would have most of the features as a dependency.

Glossary

Allogeneic Hematopoietic Stem Cell Transplant A type of stem cell transplant in which the patient receives stem cells from a donor.. 1

Autologous Hematopoietic Stem Cell Transplant A type of stem cell transplant in which the patient's own stem cells are used.. 1

Donor Chimerism The presence of donor cells in the patient's body after a stem cell transplant.. 1

Neuroblastoma A type of cancer that develops in the nervous system.. 1

Neutrophil Engraftment The process by which the patient's neutrophil count returns to normal after a stem cell transplant.. 1

Overall Survival The percentage of patients who are alive at a certain time after diagnosis or treatment.. 1

Platelet Engraftment The process by which the patient's platelet count returns to normal after a stem cell transplant.. 1

Progression Free Survival The percentage of patients whose disease does not worsen for a certain time after treatment.. 1

Transplant Related Mortality The percentage of patients who die from complications of a stem cell transplant.. 1

A Appendix A

Data Dictionary of Dataset Bone Marrow Transplant : Children

Index	Variable Name	Role	Type	Description
1	Recipientgender	Feature	Binary	1 for Male, 0 for Female
2	Stemcellsource	Feature	Binary	Source of Stem cells, 1 for Peripheral Blood, 0 for Bone Marrow)
3	DonorAge	Feature	Integer	Age of Donor at time of Stem Cell Apheresis
4	DonorAge35	Feature	Binary	Donor age $<$ 35-0, Donor age $\geq$ 35-1
5	IIIV	Feature	Binary	Development of Acute Graft vs Host Disease, Stage II,III or IV yes-1,no-0
6	GenderMatch	Feature	Binary	Compatibility of Donor and Recipient Based on Gender yes-1,no-0
7	DonorABO	Feature	Categorical	ABO blood group of the Donor of Hematopoietic Stem Cells (B = -1,O=0,A=1,AB=2)
8	RecipientABO	Feature	Categorical	ABO blood group of the Recipient of Hematopoietic Stem Cells (B = -1,O=0,A=1,AB=2)
9	RecipientRH	Feature	Binary	Presence of the RH factor on Recipient's Red blood cells ('+'- 1,'-'-0)

10	ABOmatch	Feature	Binary	Compatibility of the Donor and Recipient of the Hematopoietic Stem Cells according to ABO blood group (matched - 1, mismatched - 0)
11	CMVstatus	Feature	Categorical	Serological Compatibility of the Donor and the Recipient of the hematopoietic stem cell according to Cytomegalovirus
12	DonorCMV	Feature	Binary	Presence of Cytomegalovirus infection in the donor of the hematopoietic stem cell transplant (present-1,absent-0)
13	RecipientCMV	Feature	Binary	Presence of Cytomegalovirus infection in the donor of the hematopoietic stem cell transplant (present-1,absent-0)
14	Disease	Feature	Categorical	Type of Disease (ALL,AML,chronic,nonmalignant,lymphoma)
15	Riskgroup	Feature	Binary	High Risk-1,Low Risk-0
16	Txpostrelapse	Feature	Binary	Second Bone Marrow transplant post relapse (yes-1,no-0)
17	DiseaseGroup	Feature	Binary	Type of Disease (malignant-1,non-malignant-0)
18	HLAmatch	Feature	Categorical	Compatibility of antigens of the main Histocompatiblity complex of the donor and the recipient of hematopoietic stem cells according to all international BFM SCT 2008 Criteria (10/10 -0, 9/10-1,8/10-2,7/10-3(allele/antigens))
19	HLAmismatch	Feature	Binary	HLA matched-0, HLA mismatched -1
20	Antigen	Feature	Categorical	In how many antigens there is a difference between the donor and the recipient (-1 no difference, 0- one difference,1 - two differences, 2- three differences, 3- four differences)
21	Allele	Feature	Categorical	In how many alleles there is a difference between the donor and the recipient (-1 no difference, 0- one difference,1 - two differences, 2- three differences, 3- four differences)
22	HLAgrl	Feature	Categorical	The difference type between the donor and the recipient (HLA matched -0,Difference in one antigen -1, Difference in only one allele -2, Difference in only one DRB1 cell -3, two differences(two alleles or two antigens) -4, mismatched-5)
23	RecipientAge	Feature	Integer	Age of Recipient of Hematopoietic Stem cells at the time of transplantation
24	RecipientAge10	Feature	Binary	Recipient Age ¡10 = 0, Recipient Age ¿= 10 -1
25	RecipientAgeint	Feature	Categorical	Recipient Age (0,5],(5-10]-1,[10,20)-2
26	Relapse	Feature	Binary	Reoccurence of the disease (yes-1,no-0)
27	aGvHDIIIV	Feature	Binary	Development of Acute Graft vs Host Disease Stage III or IV (Yes - 0, no-1)
28	extcGvHD	Feature	Binary	Development of Extensive Chronic Graft vs Host Disease (Yes - 0, no-1)
29	CD34kgx10d6	Feature	Integer	CD34+ cell dose per kg of recipient Body weight
30	CD3dCD34	Feature	Integer	CD3+ to CD34 cell ratio
31	CD3dkgx10d8	Feature	Integer	CD3+ cell dose per kg of recipient weight
32	Rbodymass	Feature	Integer	Body mass of Recipient at time of transplant
33	ANCrecovery	Feature	Integer	Time to neutrophil engraftment defined as $0.5 * 10^9/L$
34	PLTrecovery	Feature	Integer	Time to platelet engraftment defined as $> 50000/mm^3$
35	time-to-aGvHD-III-IV	Feature	Integer	Time to development of acute Graft vs Host Disease Stage III or IV
36	survival-time	Feature	Integer	Time of Observation(If alive) Time till event (If dead)
37	survival-status	Target	Integer	(0-Alive,1-Dead)

Table 2: Table containing information of all features

References

[1] Sara Bowman et al. "CD34 Stem Cell Boost in Pediatric Allogeneic Stem Cell Transplant Recipients: A Case Series and Review of Literature". In: *Clinical Hematology International* 5 (2023), pp. 155–164. URL: https://api.semanticscholar.org/CorpusID:258008865.

[2] Yingjun Chang et al. "The impact of CD34+ cell dose on platelet engraftment in pediatric patients following unmanipulated haploidentical blood and marrow transplantation". In: *Pediatric Blood & Cancer* 53 (2009). URL: https://api.semanticscholar.org/CorpusID:11689146.

[3] Tristan E. Knight et al. "Effect of Autograft CD34 + Dose on Outcome in Pediatric Patients Undergoing Autologous Hematopoietic Stem Cell Transplant for Central Nervous System Tumors". In: *Transplantation and Cellular Therapy* (2022). URL: https://api.semanticscholar.org/CorpusID:263423178.

[4] Tristan E. Knight et al. "No impact of CD34+ cell dose on outcome among children undergoing autologous hematopoietic stem cell transplant for high-risk neuroblastoma." In: *Bone marrow transplantation* (2023). URL: https://api.semanticscholar.org/CorpusID:261526787.

[5] Ryotaro Nakamura et al. "Impact of graft cell dose on transplant outcomes following unrelated donor allogeneic peripheral blood stem cell transplantation: higher CD34+ cell doses are associated with decreased relapse rates." In: *Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation* 14 4 (2008), pp. 449–57. URL: https://api.semanticscholar.org/CorpusID:21228681.

[6] Alexandra Pedraza et al. "Effect of CD34+ cell dose on the outcomes of allogeneic stem cell transplantation with post-transplant cyclophosphamide." In: *Transplantation and cellular therapy* (2022). URL: https://api.semanticscholar.org/CorpusID:254717130.

[7] Mats Remberger et al. "The CD34+ Cell Dose Matters in Hematopoietic Stem Cell Transplantation with Peripheral Blood Stem Cells from Sibling Donors". In: *Clinical Hematology International* 2 (2020), pp. 74–81. URL: https://api.semanticscholar.org/CorpusID:215957183.

[8] Sikora,Marek, Wróbel,Łukasz, and Gudyś,Adam. *Bone marrow transplant: children*. https://archive.ics.uci.edu/dataset/565/bone+marrow+transplant+children. DOI: https://doi.org/10.24432/C5NP6Z. 2020.