Applications of Data Analytic Modeling for Efficient Stem-Cell Transplants and Donor Registry Management

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Adarsh Sivasankaran

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor: Vladimir S. Cherkassky

January, 2018

Acknowledgements

I would like to express my sincere gratitude to my advisor, Professor Vladimir Cherkassky, for his constant support of my PhD work. His approach to research has been thoroughly inspiring and kept me interested in my work all through. I thank Dr. Eric Williams for his advice and patience. I am particularly indebted to Eric for his help with writing and carefully reviewing this thesis. Thanks to Professor Dan Knights for serving on my committee. Thanks to Professor Claudia Neuhauser for the numerous opportunities she has provided me with over the years.

This work was done in collaboration with the Bioinformatics Research Department at NMDP and I am grateful for their co-operation and assistance. Thanks to Dr. Abeer Madbouly for helping me with the initial intern position at NMDP, which led to this thesis work. Thanks to Mark Albrecht, Jane Pollack, Martin Maiers for supporting this work and extending my stay at NMDP. Mark Albrecht's initial help in identifying possible research directions was of immense help. Thanks also to Debra Turner for helping me get the data I needed for this research.

Several stimulating conversations and discussions were had at KHKH 6-110 and I thank my fellow lab-mates – Sauptik Dhar, Han-Tai Shiao, Thomas Vacek, Jieun Lee, Hsiang-Han Chen, Zhong Zhuang – for all that. I will always fondly remember the time I spent at UMN and Minneapolis, and I want to thank Amith, Sumanth, Bhavana, Deepak, Deepti, Gautham, Mahadev, Komal for being a fun gang to hang around with. Special thanks to Srikanta and Kirthi for all the love, support, and advice ever since I first started to think about applying to grad schools.

And finally, my parents, Shylaja and Sivasankaran, have always believed in me and have had to forego personal comforts to help me succeed. Their unwavering support and confidence has always been important to me and has allowed to me to pursue my interests.

Abstract

Studies have shown 70% of the patients who require Stem Cell Transplants have to rely on Unrelated Donors for a successful treatment. Public registries, such as BeTheMatch maintained by the National Marrow Donor Program, are responsible for meeting this demand. Maintaining a large volunteer registry is a complex and expensive process. We propose data-analytic modeling for three specific problems that can aid in unrelated donor search and registry management.

I. *Donor Selection:* Donor selection for Stem Cell Transplant often requires physicians to manually select 3-5 donors from a long list of genetically compatible donors with varying degrees of match probabilities as identified by Human Leukocyte Antigen (HLA) matching algorithms. The decision process is based upon non-HLA donor attributes, and is very time consuming. We provide a binary classification model that is trained on historical past donor selection data to help make future donor selections faster and consistent.

II. *Donor Availability:* Donors can decline a sample donation request for a number of reasons, which adversely affects the time taken to complete a transplant. Past responses show that only half the requests receive a positive response. We propose a binary classification model for predicting donor availability based on donor demographic information and responses to outreach programs.

III. *Donor Utility and Recruitment:* Power law like distribution of HLA types implies that a large number of registered donors are never utilized. We provide a mathematical framework to combine the Donor Selection and Availability models with donors' HLA type to determine donor utilization, that can be used to identify donors for future cost management efforts in the registry, such as advanced typing. Studies indicate that a large number of patients also don't find a match due to lack of diversity in the registry. We provide recommendations for targeted donor recruitment to enhance diversity based on donors' geographic information.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

Stem cells are immature cells produced in the bone marrow that develop into all types of blood cells in the human body. Cancer patients who have had their bone marrows destroyed due to an aggressive chemotherapy or radiation the ability to produce new blood cells is lost. In such cases, a stem cell transplant is required to replace the damaged bone marrow to regain the ability to produce healthy blood cells. Stem cell transplants are also used to treat certain blood disorders, auto-immune diseases, and genetic disorders.

Family members are the best source of donors. However, only 30% of patients who require stem-cell transplants can find a match within their families (Gragert *et al.*, 2014; Besse *et al.*, 2016). The other 70% have to rely on Unrelated Donors (URDs) or Cord-Blood Units (CBUs) for transplants. URD and CBU searches are facilitated by adult donor registries and cord blood banks. A viable CBU and URD is identified by genetic matching algorithms (Bochtler *et al.*, 2016). To achieve the best post-transplant outcome, patient and donor's Human Leukocyte Antigen (HLA) must be matched. HLA is a system of genes found on chromosome 6 and is responsible for the immune system. A match is determined at a 5 locus level (HLA – A,B,C,DQB1,DRB1), i.e., 10 alleles are considered to find matching URDs and CBUs. For a match to be viable, at least 8 of the 10 alleles should match, with 10 of 10 match being the optimal. It has been shown that mismatched transplants have a higher rate of post-transplant complications (Lee *et al.*, 2007; Cunha *et al.*, 2014; Kanda *et al.*, 2015; Petersdorf, 2015).

Advances in HLA matching algorithms and therapeutic protocols have resulted in increased number of patients being treated with stem-cell transplants (Copelan, 2006). Donor registries around the world are maintained for meeting this higher demand. Typically, registries are responsible for recruiting donors, maintaining a database of genetic and secondary donor information of donors, developing algorithms and software to perform HLA matching, acting as the intermediary between transplant centers and donors,

and other transplant related tasks. Each of these tasks involves significant costs and resources. One such registry operating in the United States of America is BeTheMatch registry operated by the National Marrow Donor Program (NMDP). NMDP has over 10 million registered volunteer members and 185 thousand CBUs. It is also known to be one of the most diverse registries in the world.

1.1 Transplant Process at NMDP

A successful transplant involves a series of steps to be completed by the registry. A URD (or CBU) search is initiated when patients do not find a match among family members. An HLA matching algorithm, such as HapLogic^{SM} (Dehn *et al.*, 2016) developed by the NMDP, is used to identify genetically compatible donors in the registry. The NMDP uses a search interface, Traxis^{TM}, for two main purposes:

1. Entering patient information (HLA type) to initiate a donor search. The matching algorithm, HapLogic^{SM}, then identifies a list of registered members who, with varying probabilities, are genetically matched with the patient. The potential for matching depends not only on the actual set of alleles identified, but the ability to identify the alleles based on the resolution of the typing, and the typing technology used.
2. Displaying donor information to the search experts to make an informed donor selection.

Secondary donor characteristics, such as Age, Gender, Cytomegalovirus (CMV) test result, etc., that are known to be associated with favorable post-transplant success outcomes are displayed along with the donor matching information in Traxis^{TM}. A search expert then sifts through the list of matched donors and identifies the donor with the most favorable characteristics. Typically, 3-5 donors are chosen per patient. The chosen donors are then contacted and asked if they would be willing to donate a blood sample. If the identified donors agree, they are requested to visit a local hospital where a sample is collected to perform a Confirmatory Typing (CT), which is used to resolve ambiguous HLA typing and screen for infectious disease markers. If everything is found to be suitable for proceeding

with the transplant, the donor proceeds to what is referred to as the *work-up* stage, where a licensed medical practitioner obtains a detailed medical and travel history. If no red flags are discovered at work-up, the donor is asked for either a Peripheral Blood Stem Cells or a Marrow Stem Cells donation.

Additionally, registries are also responsible for recruiting donors. The main aim of donor recruitment is to maintain an active set of donors who are likely to be used for transplants. Diversity in the registry ensures that patients from different race groups are likely to find a match. A study by the NMDP in 2014 (Gragert *et al.*, 2014) showed that the likelihood of finding a perfectly matched available donor for patients of White European descent was 75% and for Black American patients of all ethnic backgrounds was between 16% and 19%. The disparity between match rates for these two ethnic populations was due to both higher genetic variance in Black Americans and a lower availability rate. Hence, enhancing the diversity of the registry and improving availability both are important matters of concern.

1.2 Thesis Contribution

Time to transplant is an important metric upon which the efficiency of a registry is measured. Any delay in the process not only adds to the momentary suffering of the patient but also adversely affects the transplant outcome. Data-driven predictive modeling can be employed to ease the bottlenecks involved in the process. In specific, we have identified three major areas where predictive modeling can be used.

1. *Donor Selection*: After HapLogic^{SM} identifies HLA compatible donors in the registry for a patient, a physician or a search expert is required to choose a handful of donors (typically 3-5) based on their secondary characteristics. While there are no strict guidelines for selecting donors, a general recommendation is provided in (Stephen R Spellman *et al.*, 2012) that uses medical studies that identify association between donor secondary characteristics and post-transplant success. Depending

on the HLA type, a patient can potentially find tens of thousands of identically matched donors. Selecting the most favorable donor in this case can be a very time-consuming process. Extensive expertise and time is required to make this selection. We propose a Predictive model trained on data from historical donor searches and selections to assist in making future selections. We present this modelling approach in Chapter 2.

2. *Donor Availability*: *Chosen* donors are contacted to request for a sample donation for Confirmatory Typing (CT) and willingness to proceed with the transplant. A consent is required at this stage to proceed with the transplant. Studies have shown that only 50%-55% of the sample donation requests for confirmatory typing are accepted. A declined request causes significant delays since a new search and selection has to be performed. A Machine Learning model is trained to predict donor's response to a CT request based on previous donor responses. We use donor demographic information and other outreach programs used by NMDP to measure donor engagement, as described in Chapter 3.

3. *Donor Utility and Recruitment*: Chances of a donor being utilized for a transplant is dependent on multiple factors. Only a small portion of the registered members on the registry are used in transplants. Having a large number of donors who are not going to be used creates waste in terms of resources used in managing them. We provide a mathematical framework to combine the two above models with the HLA type of the donors to generate a unified utility score to identify donors who are likely to contribute in a search. Another aspect that is important to registry management is diversity, which affects the likelihood of patients of different races finding a match. The most effective way to enhance diversity is by effective recruiting. We provide recommendations based on geo-coded information of donors for targeted donor recruitment.

Chapter 2 Donor Selection

2.1 Introduction

HLA compatibility between a patient and URDs (and CBUs) is established by HLA matching algorithms (Bochtler *et al.*, 2016; Dehn *et al.*, 2016). For URDs, a donor should match at least 8 of the 10 alleles (at HLA-A, B, C, DQB1, DRB1 loci) to be considered viable. A donor who matches at all 10 alleles is the most preferred. After a list of suitable donors have been identified by matching algorithms, a physician or a search expert selects a short list of donors who are likely to provide the optimal post-transplant outcome. This selection process is based on donors' secondary characteristics such as age, gender, Cytomegalovirus (CMV) test, weight, etc. While there are no strict guidelines for physicians to make this selection, there have been studies that make recommendations to guide the donor selection process (Stephen R Spellman *et al.*, 2012). These recommendations are based on medical studies that associate secondary characteristics of donors with positive post-transplant outcomes (Kollman *et al.*, 2016).

Hence, donor selection is aimed at choosing the donor with the optimal clinically relevant factors that will give the best chance of survival for patients. Depending on the HLA type of the patient, the search process might involve selecting 3-5 donors for Confirmatory Typing (CT) from a long list of HLA potentially matched URDs with varying levels of typing resolution and genetic matching. At the NMDP, patients with common HLA type can potentially find tens of thousands of donors who are identical genetic matches. Donor selection can thus involve sifting through a long list of URDs based on non-genetic factors.

Donor search and display systems, such as Traxis^{TM} developed by the National Marrow Donor Program (NMDP), are used to search for donors and convey donor selections to the registry. All clinically relevant donor characteristics are displayed in

Traxis™ for the search experts to make an informed decision. Making a choice between identically matched donors can be an extremely long and difficult process and is done while the patient is under critical care. The selection process is based on evaluating multiple secondary donor characteristics. A computational model can help ease this selection process by quantitatively identifying donors with more preferable secondary characteristics based on past searches (Shouval *et al.*, 2014). In this thesis, we develop a Machine Learning model that can mimic the donor selection process. Using a trained model, we can assign a single numerical score to every HLA matched donor for a patient to indicate combined favorability of secondary characteristics. Such a score will reduce the comparative multivariate decision process to a decision based on a single score that combines all the relevant donor features. It can be of particular assistance to physicians and TCs which lack the expertise and man-power to make such a critical decision (Irene *et al.*, 2017).

The NMDP saves all donor searches performed using their system. Matching information and the donor specific information that was relevant when the decision was made is saved. Figure 2-1(a) describes the decision process for donor selection and the modeling goal. A search expert is presented with list of donors via Traxis™. This list of donors is identified by the HLA matching algorithm, HapLogic℠, based on patient's and donors' HLA type. All the donors on the registry who match the patient's HLA are presented to the search expert. The goal of this effort is to train a model that can imitate this decision process via predictive data-analytic modeling. In Figure 2-1(b) we show the proposed change to the process using a trained predictive model. This model can be integrated into the Traxis™ system to help in donor selections. The model utilizes all secondary donor characteristics and matching information, that is available for donor selection.. We utilize historical donor searches with corresponding selections for modeling.

2.1.1 Relevant Donor Factors

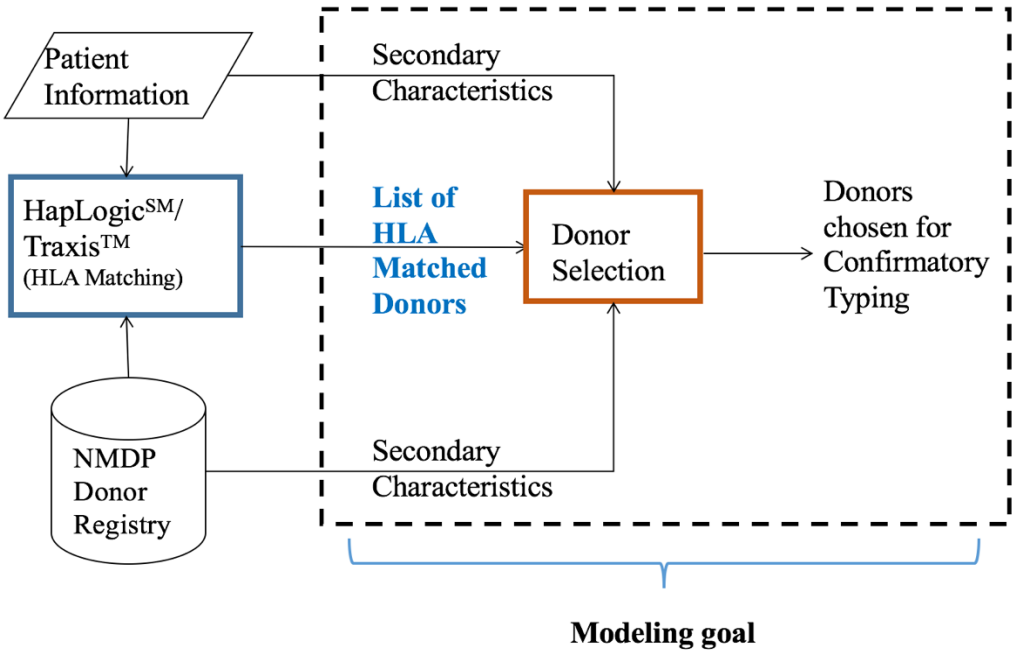
Patient survival after transplantation is the primary concern during donor selection. The main factor that adversely affects patient survival is the number of HLA mismatches

between donors and patient. (Lee *et al.*, 2007) investigates the effect of each locus on patient survival. When a recipient and donor are *Fully Matched* (10/10 allele match), the considerations for donor selection are entirely based on non-genetic donor factors. For *Partially Matched* donors (8/10 and 9/10), the number of mismatches and location of mismatch are considered for selection. (Stephen R Spellman *et al.*, 2012) makes the recommendation that when perfectly matched donors are not available a donor mismatched at HLA-B or -C may be less detrimental than a donor mismatched at HLA-A and -DRB1. These rules translate to different selection criteria for fully matched and partially matched donors. Hence, these two scenarios need to be modeled separately.

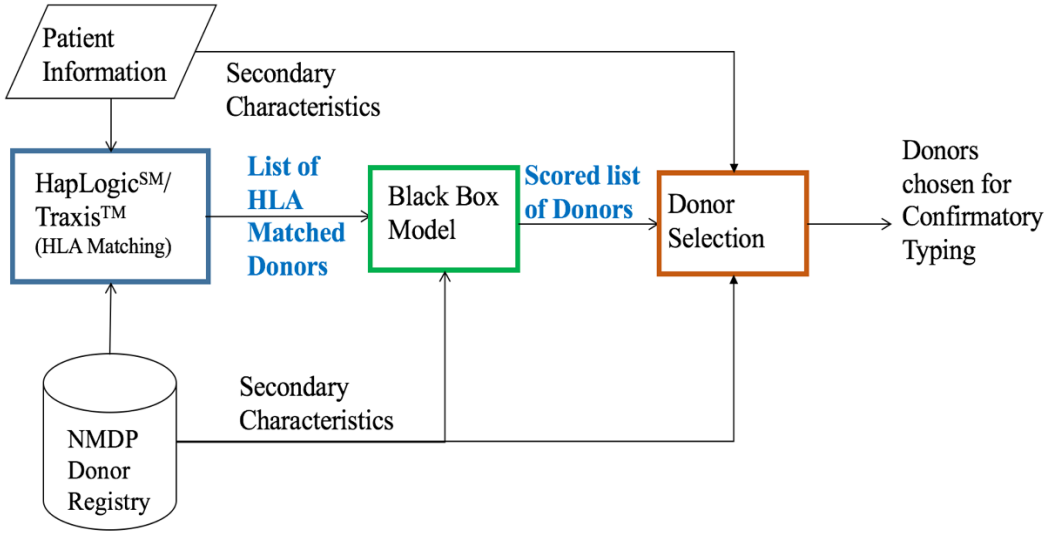
To facilitate this difference in model selection criteria we separate donor searches in to two categories based on selected donors' HLA mismatches:

- **A.** Fully Matched donor searches: All selected donors where perfectly matched (10/10).
- **B.** Partially Matched donor searches: At least one selected donor was partially matched (9/10 or 8/10).

The search categories will be referred to as Category **A** and Category **B**, respectively. HapLogic^{SM} estimates an overall match probability as well as match probabilities at the individual allele level (Dehn *et al.*, 2016), that is, a match probability at 10/10, 9/10, and 8/10 are estimated. Higher match probability corresponds to a higher chance of an exact match established at CT. For partially matched donors at least 1 allele mismatch has been unambiguously identified. Overall match probabilities (at 10/10, 9/10, and 8/10 match grades) and the mismatch locus information are important for donor selection. All this information is referred to as *secondary characteristics* in this thesis. Table 2-1 has more details about secondary characteristics used for modeling.



(a)



(b)

Figure 2-1: (a) Block diagram of the donor selection process. (b) shows the proposed system. The black box model will be used to score the list of HLA matched donors, which can be integrated into the donor search and display system.

2.2 Description of Available Data

Available data used for modeling consists of donor searches facilitated by the NMDP over a 3-month period (May 15, 2016 to Aug 9, 2016) for modeling. In consultation with search experts at the NMDP we have identified donor secondary characteristics that are important for the decision process. These are listed in Table 2-1 along with a description of clinical significance. Notice that several donor characteristics are missing in the database. This is due to either donors not sharing the complete information with NMDP when they were recruited or the information not being entered in the database.

Table 2-1: Secondary donor characteristics that are considered important to the donor selection process

Donor Characteristics	**Significance**
Donor's Age	Younger donors are preferred over older donors (Kollman *et al.*, 2016). Information is available for all donors.
Number of matched alleles	At least 8 of the 10 alleles must be matched. 10/10 match is preferred over 9/10, which is preferred over 8/10 (Petersdorf *et al.*, 2007). Information is available for all recipient-donor pairs.
Match Probability	Assigned by HapLogic^{SM} (Dehn *et al.*, 2016) for every donor, in the range 1% - 99% for each level of match (i.e., 8/10, 9/10, or 10/10 level). A donor with higher match probability is preferred. Match probability at 10/10, 9/10 and 8/10 level are used appropriately. Information is available for all recipient-donor pairs.
Donor's Blood type and Rh Factor	When possible a match between recipient and donor's blood group and Rh factor is preferred (Kollman *et al.*, 2016; Worel, 2016). This information was available for about 24% of the donors in the dataset.

Donor's Gender	Male donors are favored (Stephen R. Spellman *et al.*, 2012). Information is available for all registered members.
Donor's Race	Donors' Race is listed in Traxis™ (Dehn *et al.*, 2016) as belonging to one of the following broad categories: • African American (AFA) • Asian/ Pacific Islander (API) • Caucasian (CAU) • Declined to Answer (DEC) • Hispanics (HIS) • Multi-Group (MLT) • Native American Indian (NAM) • Other Group (OTH) • Unknown (UNK)
Donor's Weight	This information is used to estimate the volume of stem cells that can be harvested from the donor, and if the donor can meet the requirements for the recipient. This information was only available for 3.43% of the donors.
Donor's CMV report	CMV-seropositive or seronegative donors are preferred for CMV-seropositive and seronegative recipients, respectively (Boeckh and Nichols, 2004). 11.97% donors had CMV screening results.
Low/high Resolution Typing	High resolution typing indicates a stronger accuracy in match probabilities.
DPB1 Permissivity	A DPB1 permissive is preferred in addition to a 10-allele match. DPB1 match or a permissive mismatch is considered to be equally preferable (Fleischhauer *et al.*, 2012; Shaw *et al.*, 2014)
Donor Chosen	Information is collected to identify which of the matched donors is chosen to be asked for confirmatory typing.

Any variable that is identified to have missing information is encoded as a binary variable to indicate presence or absence of information at the time search was performed. DPBI information was also transformed as binary indicator to indicate if the donor-recipient match is completely matched or permissively mismatched versus other categories. Match probabilities and Donors' age were numeric variables and are transformed on a [0,1] range. Maximum and minimum values in the collected are recorded for scaling future data. Number of mismatches and Donor Race groups are categorical variables. One-hot encoding was used for categorical variables, where each level of the categorical variable is assigned a separate binary variable. DEC, MLT, OTH levels in the Race variable are merged with UNK level to account for data sparsity. In addition to the variables listed in Table 2-1, we also collected Donor ID, Recipient ID, and Transplant Center ID for measuring model performance. These IDs are unique identifiers assigned by the NMDP for internal identification and communication. There are a few other factors, such as the previous pregnancy indicator for female donors, which did not have enough representation in the registry to be effectively modeled. Such factors have been ignored from consideration for this analysis.

Available data consists of a total of 2,138 donor searches. These 2,138 donor searches resulted in a total of 8486 selections. That is, an average of 4 selections were made per recipient. These searches are identified to belong to either of the two categories as described in Section 2.1.1. Among these 2,138 searches, 1,439 searches belong to Category **A** and 699 belong to Category **B**. For modeling Category **A** searches, we remove all partially (8/10, 9/10) matched donors since these were not considered during selection. For modeling Category **B** searches, we retain complete search results to model. In all, we have 1.5 million donor-recipient pairs in Category **A** searches and 0.9 million donor-recipient pairs for Category **B** searches. Each donor-recipient pair constitutes a sample (data instance) for modeling. Two separate models can be estimated two sets of data. Table 2-2 shows raw counts of chosen donors broken by key donor characteristics, which may indicate selection preferences.

Table 2-2: Raw count of chosen donors by various donor characteristics. This gives some indication of what the preference is. We see younger and male donors are preferred

Donor Broad Race Groups

AFA	499 (5.88%)
API	534 (6.29%)
CAU	4299 (50.66%)
DEC	26 (0.31%)
HIS	767 (9.04%)
MLT	480 (5.66%)
NAM	57 (0.67%)
OTH	21 (0.25%)
UNK	1803 (21.25%)

Donor Gender

Female	3279 (32.64%)
Male	5207 (61.36%)

Donor Age

<= 32	5155 (60.75%)
[33-50]	2717 (32.02%)
> 50	614 (7.23%)

Donor-Recipient HLA Match Count

		Mismatched Location					
	Total	*A*	*B*	*C*	*DQB1*	*DRB1*	
10 on 10	6410	-	-	-	-	-	
9 on 10	2000	1108	458	90	136	208	⇐ *1st Mismatch*
8 on 10	76	37	7	1	5	26	⇐ *1st Mismatch*
		0	11	13	49	3	⇐ *2nd Mismatch*

2.3 Problem Formalization

The physician's donor choice is dependent on the presented list of HLA compatible donors. Hence, the decision process can be viewed as assigning preferences based on donor's secondary characteristics. We assume that when a donor is chosen, the physician is deliberately not choosing the other donors when a choice is made, that is, chosen donors have the most optimal characteristics among the matched set of donors. This implies donors are explicitly labeled as *chosen* or *not chosen*. This assumption leads to a binary classification setting as follows:

$$y_i = \begin{cases} 1, & if\ donor\ \boldsymbol{i}\ was\ chosen \\ -1, & if\ donor\ \boldsymbol{i}\ was\ not\ chosen \end{cases} \tag{2.1}$$

We further assume that all physicians make similar choices when presented with the same set of genetically matched donors. Physicians also do not have an order of preference for the donors chosen for confirmatory typing. This prevents us from formalizing the problem as a ranking problem. Considering all the restrictions in the data and the requirements of the application, a binary classification formalization is the best option. Two different models are trained for *fully matched donor selection* (Category **A**) and *partially matched donor selection* (Category **B**). Donors' secondary characteristics detailed in Table 2-1 are used as input for the model. The majority of the donor characteristics were either binary or categorical. After one-hot encoding of categorical variables, the dimensionality of input features is 14 for Category **A** searches and 33 for Category **B** searches. The practical application of this modeling effort is to enhance the donor search experience. We can achieve this by displaying the more favorable donors at the top of the search list. Hence, we need to modify our predictive model for ranking HLA matched donors in a donor search.

In Category **A**, an average of 257 donors were not chosen for every chosen donor. For Category **B** searches 298 donors were not chosen for every chosen donor. This indicates a severely imbalanced dataset. Usually the learning algorithm accounts for this imbalance by using an appropriate application specific cost function. This information is typically provided by domain experts. For the application, the costs cannot be provided by medical experts, since a patient's life is potentially at risk. Hence, we use the ratio of positive to negative samples as the misclassification cost. Misclassification cost ratio ***r*** is defined as:

$$\frac{C_{fn}}{C_{fp}} = \boldsymbol{r} = \frac{Number\ of\ Not\ Chosen\ donors}{Number\ of\ Chosen\ donors} \tag{2.2}$$

where:

C_{fn} – Cost of False Negative errors

C_{fp} – Cost of False Positive errors

For the analyzed dataset, ***r*** is 257 for Category **A** searches and 298 for Category **B** searches.

2.4 Methods

As noted in the previous section, the learning problem under binary classification setting is heavily imbalanced. Researchers have introduced many techniques to handle data imbalance (Elkan, 2001; Zadrozny, Langford and Abe, 2003; Weiss, McCarthy and Zabar, 2007). These techniques follow two basic approaches:

i. Cost-Sensitive Learning: Incorporate predefined misclassifications costs to the learning model and formulation.

ii. Undersampling/Oversampling: Samples from one class is undersampled or oversampled to maintain class balance so the classifier is not biased towards the majority class (Chawla *et al.*, 2002; Juanjuan *et al.*, 2007).

For our current problem, Undersampling could potentially eliminate information critical to the classifier. Therefore, we use a model that incorporates unequal costs into the learning model.

Cost-sensitive SVM (CS-SVM) is known to be capable of handling data imbalance. It is also known for its robustness in estimating predictive models from noisy and high-dimensional data and has been successfully used in several applications.

2.4.1 Cost-Sensitive SVM

SVM (Support Vector Machine) is a learning procedure based on statistical learning theory (Vapnik, 1995). It is popularly used for predictive learning problems (classification and regression). We are currently dealing with a binary classification problem, i.e., $y = \{+1, -1\}$. The model is to be estimated from finite data $(\boldsymbol{x_i}, y_i), i = 1, 2, \dots, n$, with $\boldsymbol{x} \in \mathbb{R}^d$ and $y = \{+1, -1\}$. The goal of SVM is to find the optimal decision function $f(\boldsymbol{x}, \omega) = sign\ (\boldsymbol{w}.\boldsymbol{x} + b)$ with good generalization performance.

If the training data is linearly separable, there are many separating hyperplanes satisfying the constraints $y_i(\boldsymbol{w}.\boldsymbol{x_i} + b) \geq 1, i - 1, \dots, n$. SVM identifies the hyperplane for which the margin (i.e., the distance between the closes data points to the hyperplane) is maximized, which is called the *optimal hyperplane*. The concept of margin is illustrated in Figure 2-2.

Maximizing the margin translates to minimization of the $\| \mathbf{w} \|$. To achieve this, SVM solves the following optimization problem:

$$\min_{w} \frac{1}{2} \left|\left|\boldsymbol{w}\right|\right|^2 \qquad Subject\ to: y_i(\boldsymbol{w}.\boldsymbol{x_i} + b) \geq 1, i - 1, \dots, n \tag{2.3}$$

When the data is not linearly separable, some training samples are allowed to fall inside the margin, referred to as the *soft margin* as shown in Figure 2-2. Non-negative slack

variables are introduced as $\xi_i = \max(1 - y_i f(\boldsymbol{x_i}), 0)$ to account for the deviations from the margin borders. The learning formulation then becomes:

$$\min_w \frac{1}{2} ||\boldsymbol{w}||^2 + C \sum_{i=1}^{n} \xi_i \qquad (2.4)$$
$$Subject\ to: y_i(\boldsymbol{w}.\boldsymbol{x_i} + b) \geq 1 - \xi_i, i - 1, \dots, n$$

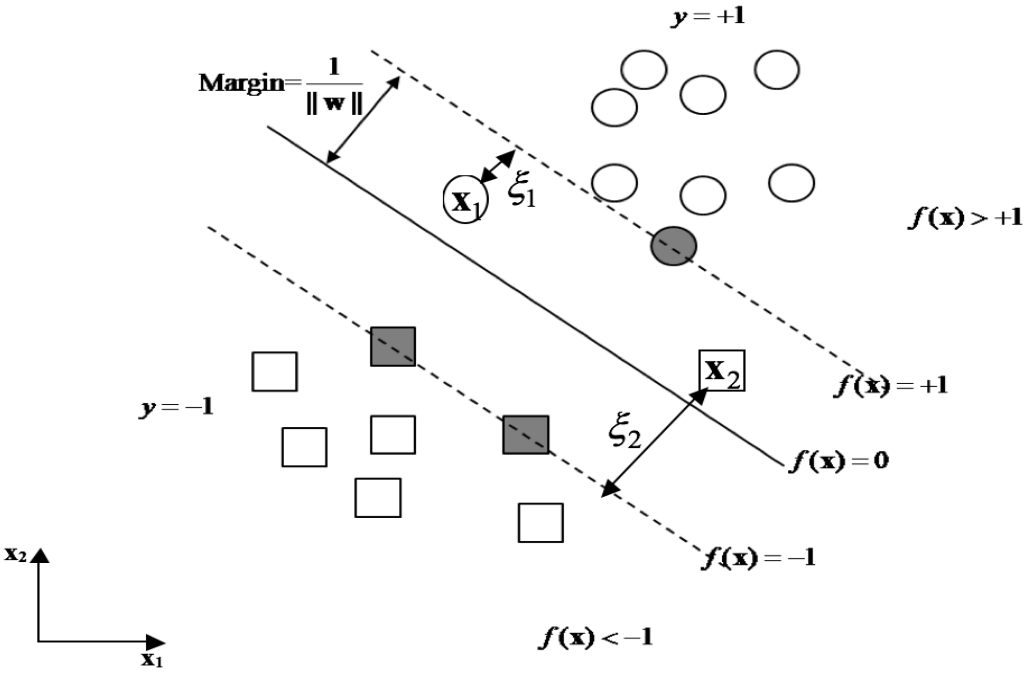


Figure 2-2: Non-separable case of binary classification. Slack variables $\xi_i = max(1 - y_i f(\boldsymbol{x_i}), 0)$ *correspond to deviation from the margin borders.*

In this form, the coefficient *C* controls the trade-off between complexity and proportion of non-separable samples and must be determined by model selection. Problem (2.4) is a quadratic programming (QP) problem, which is typically solved in its dual form:

$$\min_{\alpha} -\sum_{i,=1}^{n} \alpha_i \ + \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\boldsymbol{x_i}. \boldsymbol{x_j})$$
$$Subject\ to: \sum_{i,=1}^{n} y_i \alpha_i = 0, \qquad 0 \le \alpha_i \le C\ , \qquad i - 1, \dots, n \tag{2.5}$$

$(\boldsymbol{x_i}. \boldsymbol{x_j})$ can be extended with a kernel function $\kappa(\boldsymbol{x_i}. \boldsymbol{x_j})$ inner product. This allows for nonlinear decision functions without an explicit transformation of the data to higher levels.

In cost-sensitive setting, we assign importance (or cost) to false positives and false negatives as specified by the ratio of misclassification costs $\boldsymbol{r} = {C_{fp}}/{C_{fn}}$. The learning goal here is to estimate a model that will minimize the weighted error for future test samples.

$$Weighted\ Test\ Error = C_{fp} P_{fp} + \ C_{fn} P_{fn}$$

Here P_{fn} and P_{fp} are probability of false negatives and false positive errors respectively. For SVM, the primal form that incorporates the cost is shown in Equation (2.6).

$$\min_{w,b,\xi} \ \frac{1}{2} \left|\left|\boldsymbol{w}\right|\right|^2 + C_{fn} \sum_{i\ \epsilon + class} \xi_i + C_{fp} \sum_{i\ \epsilon - class} \xi_i$$
$$Subject\ to: y_i(\boldsymbol{w}. \boldsymbol{x_i} + b) \ge 1 - \ \xi_i, \qquad \xi_i \ge 0, \qquad i - 1, \dots, n \tag{2.6}$$

The dual formalization remains unchanged except for the constraints, as shown in (2.7).

$$\min_{\alpha} -\sum_{i,=1}^{n} \alpha_i \ + \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\boldsymbol{x_i}. \boldsymbol{x_j})$$
$$Subject\ to: \sum_{i,=1}^{n} y_i \alpha_i = 0, \tag{2.7}$$
$$0 \le \alpha_i \le C_{fn}\ , \qquad i\ \in\ + class$$
$$0 \le \alpha_i \le\ \ C_{fp}, \qquad i\ \in\ - class$$

2.4.2 Histogram of Projections

Typical model interpretation techniques for SVM classifiers are based on manual selection of low dimensional projections and identification of few important input features. Histogram of projections is a novel method of visualizing SVM models by projecting training (or test) inputs onto the normal direction of the decision boundary. (Cherkassky and Dhar, 2010) show how histogram of projections can be used for improved understanding of optimally trained SVM models, based on the idea that such univariate histograms of projections reflect well-known properties of SVM classifiers. Figure 2-3 illustrates the steps in generating histogram of projections for linear SVM. Similar procedure can be used to generate histograms for nonlinear SVM models. Visual analysis of univariate histogram shows data separability and class overlap. As mentioned in Section 2.3, the modeling dataset is severely class imbalanced. (Cherkassky and Dhar, 2015) demonstrate how these histograms can be used to interpret models with unequal misclassification costs. They also note that this analysis helps in quantifying confidence in SVM predictions, based on the distance from the margin border. Test inputs outside the +/- 1 margin have a higher confidence than the inputs that have projection distances within the margin. In the following Discussion Section (see Section 2.5) we show how these univariate histograms can be used for better donor search presentation.

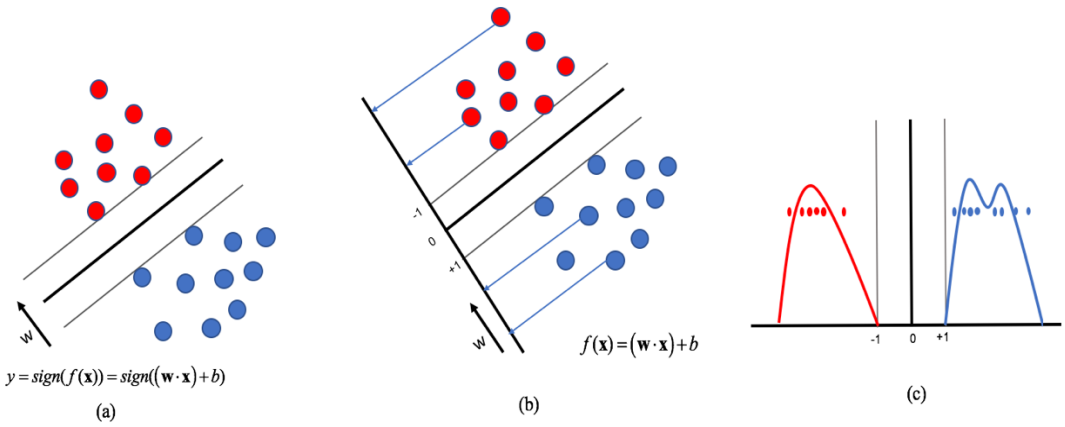


Figure 2-3: Illustration of steps in generating univariate histogram of projections. (a) Optimal SVM model with training data. (b) Projection of training data onto the normal direction of the SVM hyperplane. (c) generated univariate histogram

2.4.3 SVM Model Selection and Experimental Setup

Model selection is necessary to ensure best generalization (performance on out-of-sample data) performance on test data. This is achieved by proper tuning of hyperparameters. SVM model selection involves two main components:

1. Parameter(s) controlling the *margin size*.
2. Model parameterization, that is, the choice of kernel type and its complexity parameter.

There are several approaches to model selection (Chapelle *et al.*, 2002; Cherkassky and Mulier, 2007; Chang and Lin, 2013). We perform exhaustive search of parameter values that achieves the least validation error estimated by cross-validation. For linear kernel, the only relevant factor is the *C* parameter which controls the *margin size*. In cost-sensitive modeling, the ***r*** parameter (miscalculation cost ratio) determines how the points within the margin are penalized. However, this parameter is not to be treated as a tuning parameter.

Since the data is heavily imbalanced, care should be taken during data separation for cross-validation. Class imbalance ratios need to be kept approximately constant when data is split (learning, validation, and test). When using double resampling for model estimation available data is split into training and test sets, and model selection is performed with the training set (Cherkassky, 2013). Depending on the total data available train/ test split ratios were different for the two models. Relevant information is tabulated below in Table 2-3. We use 5 fold cross validation on training data for model selection. A grid search of best parameters based on average validation error across different folds is performed. The only tuning parameter is *C* for linear SVM.

*Table 2-3: Summary of available data of Fully matched (**A**) and Partially matched (**B**) donor selections*

	*Fully Matched (Category **A**)*	*Partially Matched (Category **B**)*
Total # of samples	1347671	939005
Total # of Donor Selections (+ve samples)	5507	2976
Training/Test Ratio	25:75	33:67
Imbalance Ratio (+ve : -ve samples)	1:257	1:298
Dimensionality	14	33

2.5 Results and Discussion

An extension of LIBSVM's (Chang and Lin, 2013) package, LiblineaR (Fan *et al.*, 2008), was used. LiblineaR, a library for larger linear SVM classification, allows for faster model estimation for linear SVMs. We also used RBF SVM, which did not offer any improvement over the linear version on a smaller subset of data. Other techniques, such as kNN, CART, Random Forests, Boosting Decision Trees were initially used for modeling. None of them offered competitive performance in this highly unbalanced data set.

2.5.1 Experimental Results

We use a normalized weighted error to evaluate the model performance for the heavily imbalanced datasets as suggested in (Cherkassky and Mulier, 2007). In this metric False Positives and False Negatives are weighed appropriately using the same cost parameter ***r*** that is used to train the cost-sensitive model. Typically, classifier accuracy is measured by the percentage of samples misclassified by the classifier. However, the class imbalance in the dataset biases the standard accuracy metric towards the majority class. For example, a naïve classifier, which labels all samples as *not chosen*, will have an accuracy of about 99.9%. Hence, using the standard metric will not be effective for either model selection or

measuring model performance. The weighted test error is given below (see Equation 2.8), where ***r*** is the misclassification cost ratio as defined in Equation 2.2.

$$Normalized\ Weighted\ Error = \frac{N_{fp} + \mathbf{r} \times N_{fn}}{N_{-} + \mathbf{r} \times N_{+}} \quad (2.8)$$

where:

N_{fp} – Number of false positives

N_{fn} – Number of false negatives

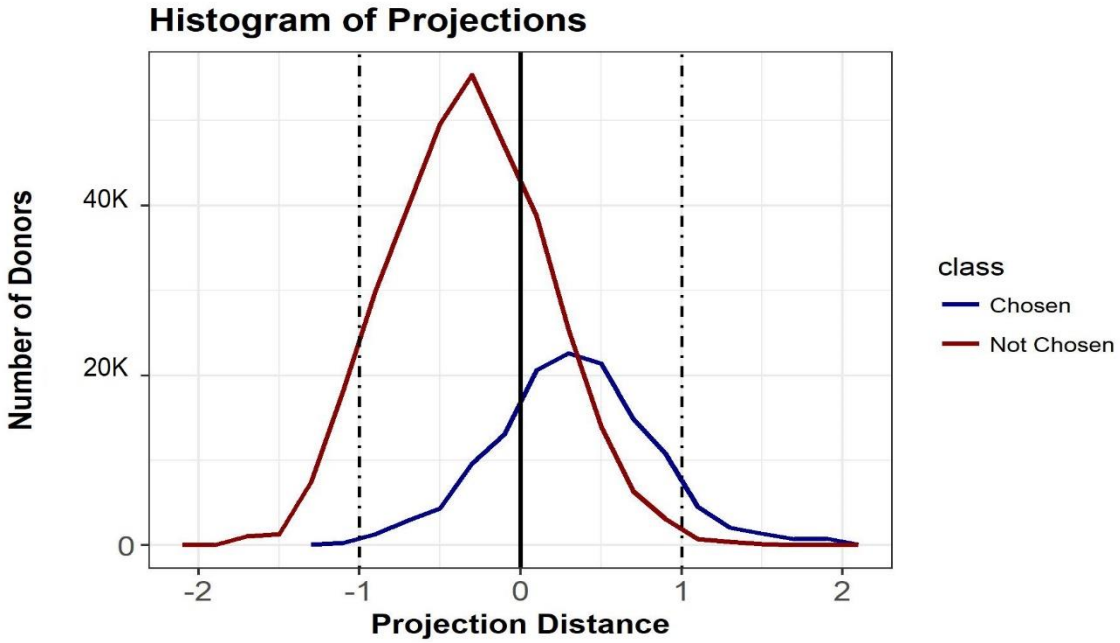
N_{-} – Number of negative samples

N_{+} – Number of positive samples

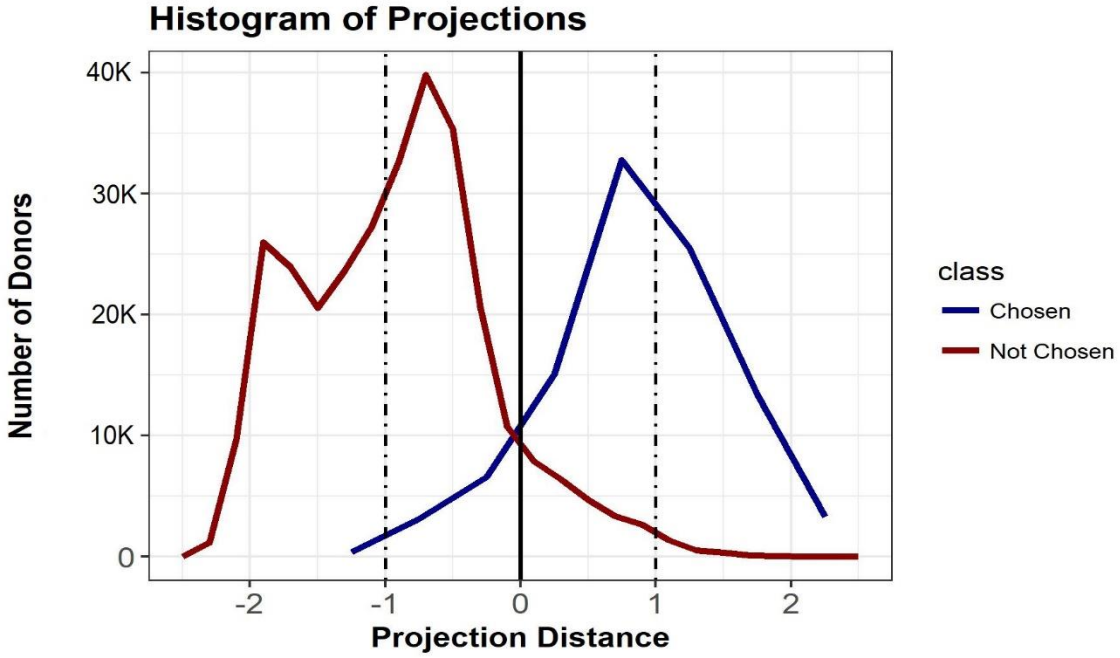
Using this measure, the simple majority classifier will have an error rate of 0.5. Table 2-4 shows error rates for the two models. False Positive Rate (FPR) and False Negative Rate (FNR) are both calculated on the test set. Training and Testing errors are weighted according to Equation 2.8. We can analyze the trained model based *Histogram of Projections* (Cherkassky, 2013). Figure 2-4 shows the Histograms for training data for two models **A** and **B**. The positive class is shown on a different scale for better visual clarity.

Table 2-4: Test errors for the two models.

	*Fully Matched (Category **A**)*	*Partially Matched (Category **B**)*
Training Error	0.2314	0.09260
Testing Error	0.2454	0.09262
FPR	0.2574	0.1061
FNR	0.2342	0.0555



(a)



(b)

Figure 2-4: Histogram of Projections of training data for the two models. (a) is for Fully matched donor selection model (Scenario A), (b) is partially matched donor selection model (Scenario B). The two classes are represented on different vertical scales.

We see a significant class overlap in Figure 2-4 (a) for fully matched donor selection model. This can be explained by understanding the donor search process. When a selection is made for searches with fully matched donors, most HLA matched on the list are identical genetic matches. The not chosen donors have similar secondary characteristics as compared to chosen donors. Consequently, many *not chosen* donors will be labeled as *chosen.* This explains the large False Positive Rate. In contrast, for Partially matched donor selection model, the class overlap is much smaller (see Figure 2.4(b)), suggesting that selection of donors is made largely based on secondary donor information.

2.5.2 Discussion

SVM classifiers are typically used to assign a predicted label to new data instances. Use of hard label assignment will lead to donors being labelled *chosen* or *not chosen*. The end goal of this modeling is to be able to assign donors most likely to be asked to donate with a higher score. Hard label assignment will not help us with this goal. Projection distances, which are used to make class assignments, can be used to assign a real value score to matched donors instead of class labels.

To effectively assist the decision process, donors who are more likely to be chosen should be assigned a higher score than other donors in the list, that is, donors with favorable secondary characteristics should receive higher scores than the donors with less favorable characteristics. Figure 2-5 shows the preferential ordering method for a hypothetical donor search. A list of HLA compatible donors, as identified by the matching algorithm, are assigned a score based on their secondary characteristics. This score is then used to order the donors for the donor display system. This may help Transplant Centers to simplify the decision process for donor selection.

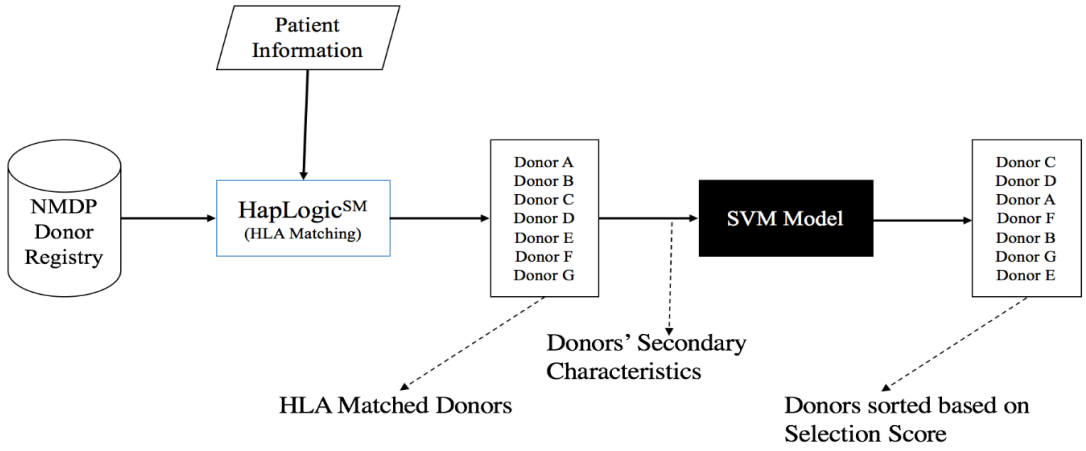
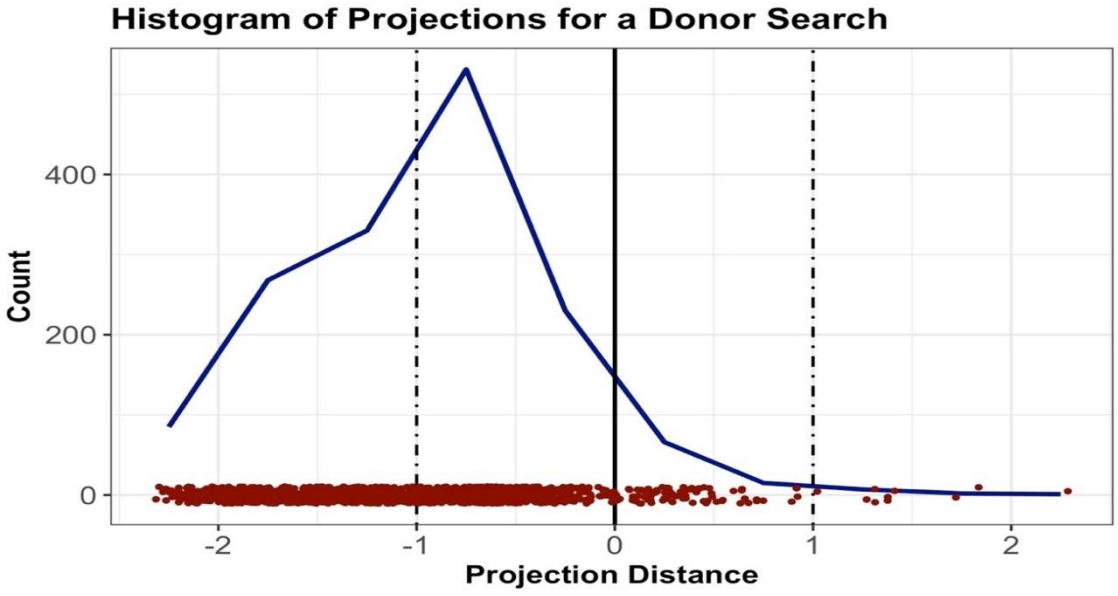


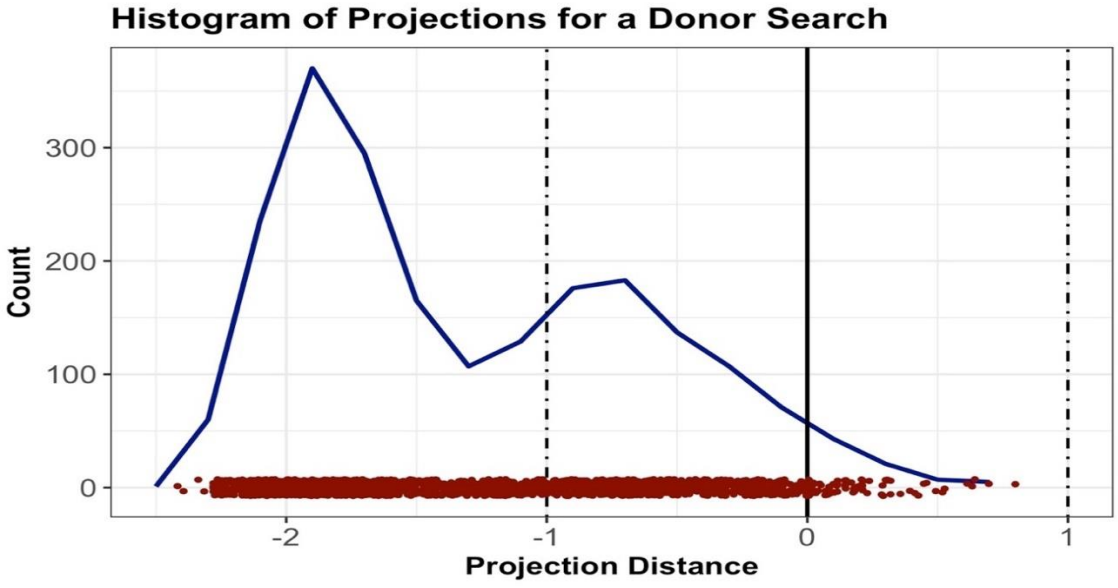
Figure 2-5: Donor sorting based on SVM model in the testing stage. Once the donors are identified to be a match for a patient's HLA type from the registry, the donor's secondary characteristics are fed to the trained SVM model to obtain a score for each donor. The donors are then sorted in ascending order based on the score.

Search Presentation: Histogram of Projections are typically used for analyzing and interpreting optimal models based on training data. Here we show how similar histograms can be generated using test data for graphical representation of test data. Using SVM projection distances, a graphical representation of the high dimensional data can be produced using Histogram of Projections for each patient as shown in Figure 2-6. This representation provides the users with an ability to view the multi-variate high dimensional donor data on a simple histogram. This will help narrow down the number of donors that need to be considered to make the decision. Figure 2-6 (a) and (b) show the histograms of projections for real patients (in test data) with 415 and 456 matched donors respectively, that were not used for training the model. Using the histogram of projections, we can restrict the search size from the entire list to a handful of donors with highest scores. In Figure 2-6 (a), a small portion of the donors have really high scores, indicating donors with really favorable characteristics. In Figure 2-6 (b), donor scores are clustered, indicating donors have similar secondary characteristics. Here too, the search field can be limited to only donors with positive scores. This representation can be easily integrated to the donor

display system. Donors with scores higher than +1 will have extremely favorable characteristics, and similarly, donors with scores less than -1 will have unfavorable characteristics. We notice that in most searches only a small percentage of matched donors were assigned positive scores.



(a)



(b)

Figure 2-6: (a) Histogram of Projections for a search with 415 matched donors. A small portion of donors have a very high score. (b) histogram of Projections for a search with 456 donors, with a more clustered score.

In the available dataset, a donor search resulted in 44,646 matched donors for a patient. Looking for the best donor for this patient would have been extremely time consuming. The donor search experience can be improved using the modeled score. Figure 2-5 shows how the SVM model can be used to sort donors. Matched donors are assigned a rank based on decreasing model score. The donor with the highest score gets rank 1, the donor with the second highest score gets rank 2, and so on. We analyze ranks of chosen donors (per patient) in the test data based on the proposed sorting method. Figure 2-7 has the cumulative distribution of the maximum rank (position of donors in the sorted list) of chosen donors. 75% of all the searches had all their chosen donors ranked within a position of 45. This indicates that the proposed model assigns higher score to favorable donors efficiently.

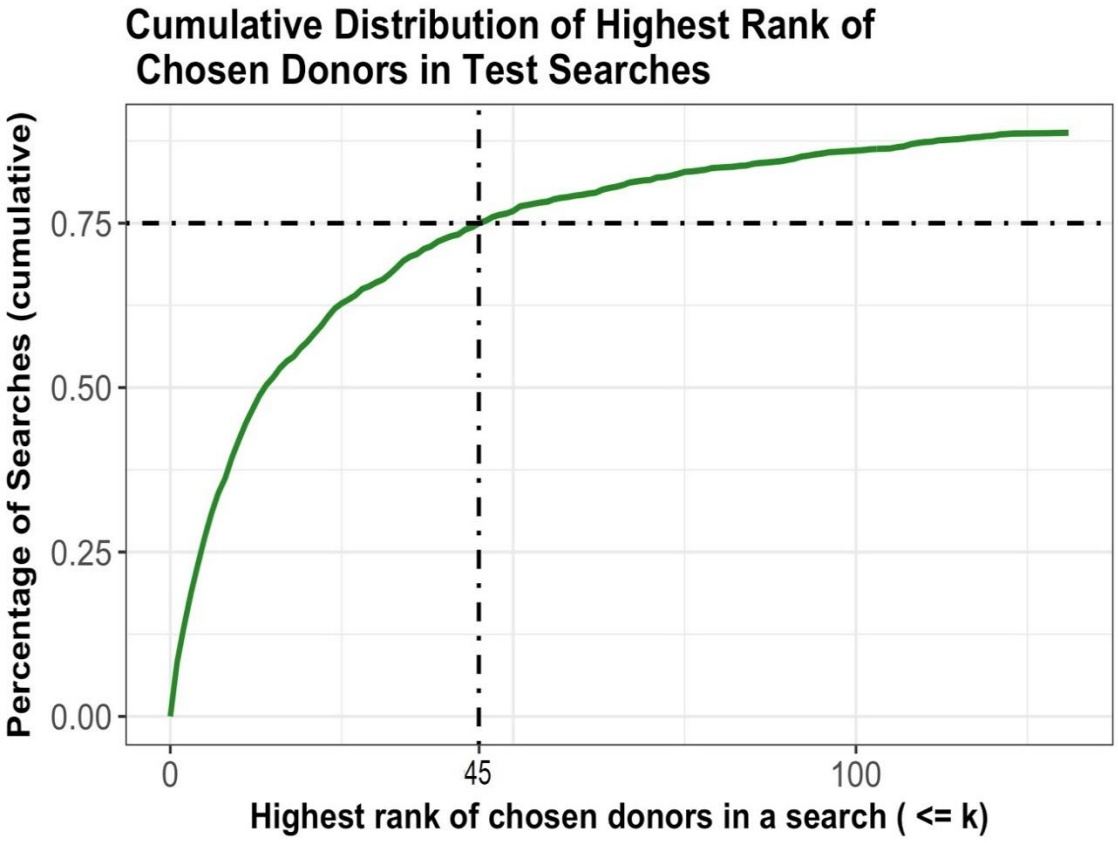


Figure 2-7: Cumulative distribution of highest rank of chosen donors. The graph is truncated on the x-axis for presentation.

Transplant Center Monitoring: Donor selection requires dedicated staff with expertise and knowledge of selection protocols and is a time-consuming process. The current modeling effort provides a direct method to quantify the search efficiency. Model assigned scores can be used to analyze the effectiveness of donor selection behavior of TCs and provide feedback when it is noticed suboptimal choices are made repeatedly. Figure 2-8 shows a hypothetical behavior for two Transplant Centers. An optimal selection (shown in Blue line) occurs when most of TCs selected donors have a positive selection score. Suboptimal donor selection practice will also lead to donors with unfavorable secondary characteristics being chosen over donors with more favorable characteristics (as shown in Red line in Figure 2-8).

Figure 2-8: Selection Score densities for Optimal and Not Optimal choices.

In conclusion, donor searches are often performed when patients are under critical care. Having to choose between identically matched donors can be a huge burden on physicians and search experts. We have shown that use of Machine Learning can alleviate some of this pain and help make decisions faster. The trained model provides a quantitative way to compare and select donors and the decision can be reduced to a single variable. This will help to make choices faster and complete transplants more quickly. Further analysis of variable weights has shown that they correspond to how decisions are made in practice. Incorporating the model information into donor display systems can help streamline the URD search process and improve efficiencies. Time to transplant is an important concern for both TCs and Donor Registries (Dehn *et al.*, 2016). The proposed model promises to reduce time spent on reviewing search results to make the most suitable choice.

Chapter 3 Donor Availability

3.1 Introduction

A successful stem-cell transplant involves a series of inter-related steps. After donor selections are made (as described in Chapter 2), consent from selected donors is required to proceed with Confirmatory Typing (CT). Chosen donors are contacted to request a sample donation. Historically, only 50-55% of such sample donation requests receive positive response. When donors decline a sample donation request, donor selection has to be performed again which causes significant delays in completing the transplant. Inability to predict which donor is likely to donate remains a major point of concern. For the purposes of this study, we define *availability* as a positive donor response to a CT request after a physician has identified the donor to be a suitable match.

Typically, donor availability is an important consideration during donor selection to avoid delays in the transplant process. However, there is no direct way to predict donor availability. Search experts have historically relied CT request responses as they relate to group-averages based on demographic information. These group averages tend to be highly inaccurate when extrapolated to the individual donor level. Rigorous registry wide studies such as (Gragert *et al.*, 2014) also rely on such group averages. Our aim is to develop a machine learning model to predict donor availability.

Even though the average availability is 50-55%, several donor characteristics are known to be associated with varying levels of donor availability. For example, donor race and ethnicity shows high variation in donor availability. Similar relationships are also known for donor age, gender, and other characteristics. In an effort to keep donors engaged and committed to the cause of stem-cell donation, NMDP has devised several outreach programs. These efforts require donors to respond to communication requests sent via email or phone. Donor responses to these requests are tracked and saved in the database.

Analysis of these outreach data has indicated higher availability among donors who respond to such requests. We propose a comprehensive model that is utilizes all donor related information that is known to affect donor availability. We provide average availability rates by donor characteristics in the next section (Section 3.2).

3.2 Description of Available Data

Analysis of available data over time has helped experts identify subgroups with higher (or lower) than average availability rates. Apart from the demographic data collected at the time a donor joins the registry, the NMDP also captures several specific member responses. These include:

(a) Response to questions on a post recruitment survey that has been specifically developed through research to measure member commitment to the donation process

(b) Member responses collected from email and social media invitations to renew commitment

(c) Answering a health history questionnaire when the member has been identified as a potential match on daily generated search reports and contacted by NMDP personnel

(d) Member initiated contact with the NMDP to request updates to their contact information; joining the registry through online registration (versus live drive recruitment where outside influences can more easily sway a person's decision to join the registry).

It has also been observed that race and ethnicity are strong indicators of availability. The standard procedure for analyzing the effect of these characteristics on availability is to calculate the historical average among the donors belonging to a particular population (for example, members of self-identified Caucasian race, or members who have renewed commitment) and extending this average to every member in the population. These averages are also used in studies that assess match rates for different patient populations.

Average availability rates for African American donors have historically been much lower than Caucasian donors. (Gragert *et al.*, 2014) reported average availability rate for Caucasian registry members to be 51% and for African American registry it was only 23% at the time the study was performed.

We analyzed CT request data from the period August 1, 2013 to November 30, 2015. A total of 178,249 CT requests were made during this period. Associated donor data (demographics and response to outreach programs) were collected for modeling. A description of the data collected is listed below.

1. *Donor Race and Ethnicity:* every donor is assigned to one of the following categories based on self-identified race and ethnicity at the time of registration. The broad classification follows the convention used in HapLogic (Dehn *et al.*, 2016). For other studies a finer classification may be used. The following race groups are retained as levels in a categorical variable for modeling.
    a. AFA – African American
    b. API – Asian and Pacific Islander
    c. CAU – Caucasian
    d. DEC – Declined to Answer
    e. HIS – Hispanic
    f. MLT – Multi-Race groups
    g. NAM – Native American
    h. OTH – Others
    i. UNK – Unknown

2. *Donor Age at request:* Calculated from donor birth date and the date the CT request was placed. This information is used as a numeric variable for modeling.
3. *Years on Registry:* Calculated from donor registration date and the date the CT request was placed. This information is also used as a numeric variable.

4. *Donor Gender:* Collected at the time members join the registry. A binary variable is formed to indicate Female or Male donors.
5. *Recommitted:* Registry members are sent communications asking them to renew their commitment to the cause of donation. Whether a donor responds to outreach, or not, is recorded in the NMDP database. A response is associated with higher availability. A binary response indicator is generated.
6. *Address Change:* Identifies member initiated address change requests. A binary indicator variable is used to identify donors who have initiated a change of address.
7. *Do-It-Yourself:* Primary method of donor recruitment is via live drives where a representative (or a volunteer) from BeTheMatch collects information and adds people to the registry. Another way to register is via an online registration which allows members to register themselves. The second method of registration indicates higher availability. A binary variable indicates if a donor chose to join the registry via the online registration form.
8. *Health History Questionnaire (HHQ) response:* Indicates that the member appeared as a match on a URD search report, was contacted prior to a CT request and answered the health history questionnaire. A HHQ response is also indicative of higher availability. A Binary indicator was used for modeling.
9. *Post Recruitment Survey (PRS):* Once a member is added to the registry, a survey is sent to evaluate their commitment to the registry. Donors answer 4 questions about the stem cell donation process that have been found, through research, to be most relevant to availability. Donors are scored according to their responses. We also have an associated variable to indicate if a registered member was asked to respond to this survey or not. An indicator variable was used to identify donors who responded to the survey. Responses from each of the 4 individual questions were changed to a positive or negative response and a total score was generated. Each of these variables (2 indicators and response components) were used for modeling.
10. *Donor Center (DC):* In order to facilitate searches, several different network sources of donors are inventoried within the NMDP search database. Donor availability varies for these networks and the set of recruiting donor centers within

each network. In all there were 13 DC codes used for modeling, as different levels in a categorical variable.

Table 3-1 shows variations in responses to outreach programs by race groups and gender. We notice donor responses are not uniformly distributed in either gender or race groups. Interactions between above variables also exhibit different availability rates. Traditional determination of availability is done by sub-setting data, for example, into groups such as *CAU Males* from *DC Code A* who have a *Recommit Response* and subsequently calculating the group averages. The number of such possible sub-groups grows exponentially with the number of considered variables. For example, there are 468 ways of interaction between just the levels of Race, DC Code, and 2 binary indicators. Sub-setting data into such finer detail might show spurious results. We have thus avoided presenting marginal availability values at this level. Accounting for different subgroups at this level will also not have a large enough sample size to make conclusive inferences. We show in the next section that different variables have different levels of associated average availability. There is also a problem of missing information for donors who are sourced from external donor centers. For example, most international donors have Race listed as Unknown. When experts have to rely on group averages for availability averages, missing or unreliable information distorts group estimates. Furthermore, use of nonlinear machine learning models have mechanisms to account for these interactions without users having to specifically hand-craft features.

Table 3-1: Number of registry members responding to outreach programs and action items broken down by self-identified Race groups and Gender.

	Recommit Response		**Health History Questionnaire**		**Do-It-Yourself**		**Change of Address**	
					Live			
	No	*Yes*	*No*	*Yes*	*Drive*	*Online*	*No*	*Yes*
Race								
AFA	14587	681	14708	560	12270	2998	13588	1680
API	11366	526	11164	728	10453	1439	9763	2129
CAU	91426	7958	91709	7675	43509	55875	82245	17139
DEC	802	6	804	4	13	795	795	13
HIS	18249	1128	18296	1081	16675	2702	16751	2626
MLT	10512	771	10454	829	7339	3944	9384	1899
NAM	1341	73	1346	68	1019	395	1248	166
OTH	328	8	327	9	74	262	322	14
UNK	18480	7	18480	7	34	18453	18471	16
Gender								
Female	65862	6205	65899	6168	42965	29102	58875	13192
Male	101229	4953	101389	4793	48421	57761	93692	12490

3.2.1 Average Availability Rates by Donor Characteristics

Each of these identified donor demographics and responses are associated with varying levels of availability. Figure 3-1 shows observed availability rates by donor race in the available dataset. Average availability rates for some of the above-mentioned factors are shown below. Race is often considered a difficult factor to account for in matching, due to errors in self-identified information and complex ancestry information (Hollenbach *et al.*, 2015). We do not consider the ambiguity of Race groups in self-identified information for modeling. Figure 3-2 and Figure 3-3 show availability rates separated by donor gender

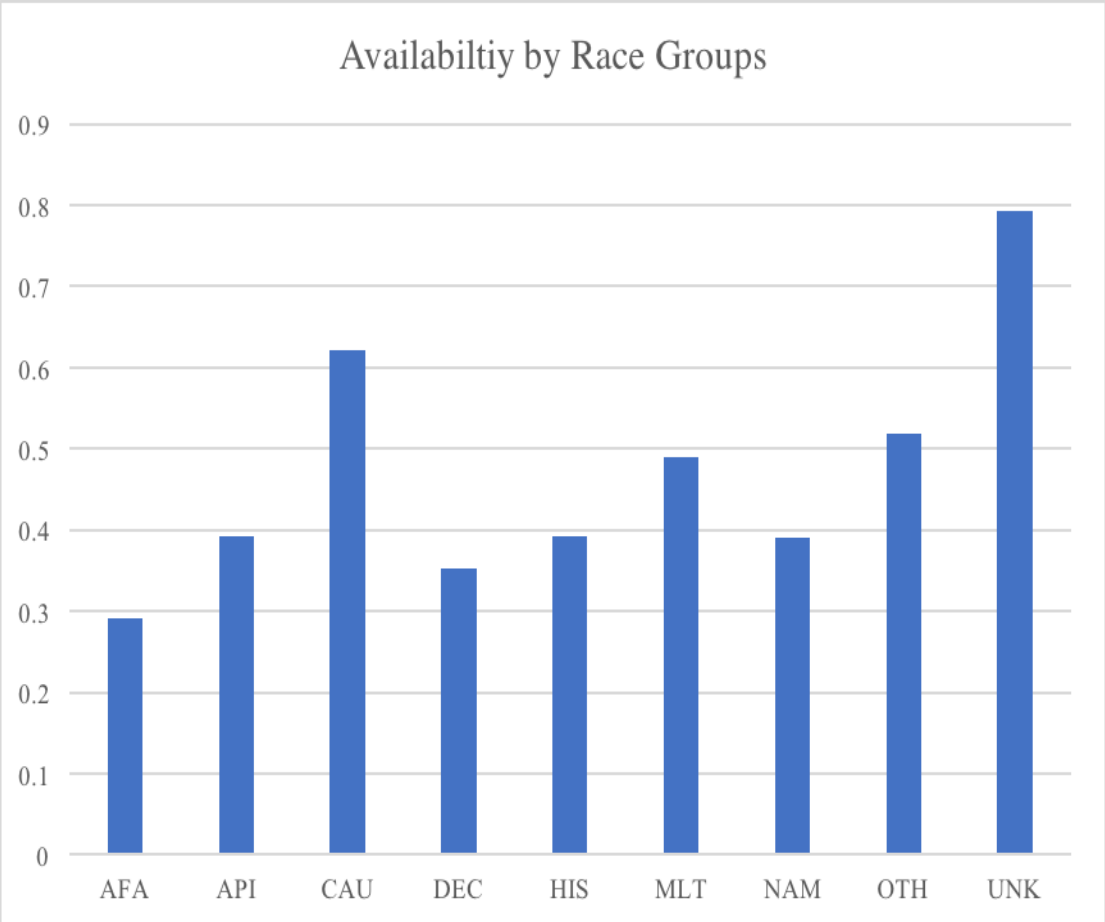


Figure 3-1: Availability by self-identified broad race groups. Significant variation is apparent in the average availability rates between groups. Caucasians have an availability rate of 62% while African American donors have an availability rate of 29%.

and source donor center. We do not notice any difference in Availability between Male and Female donors. But, DCs are associated with varying levels of availability.

Responses to outreach programs are also associated with varying levels of availability. The recommit option is one such effort. We notice the donors who have recommitted have a significantly higher availability – 93% vs 53% (see Figure 3-4). A similar trend is seen in donors who have other recorded responses (see Figure 3-5, Figure 3-6). While these Figures show a very strong indication of higher availability, we should note from Table 3-1 that donors of different race groups and gender respond differently to these programs. Hence, just using these raw univariate estimates isn't informative enough. Current analyses are restricted to these univariate calculations.

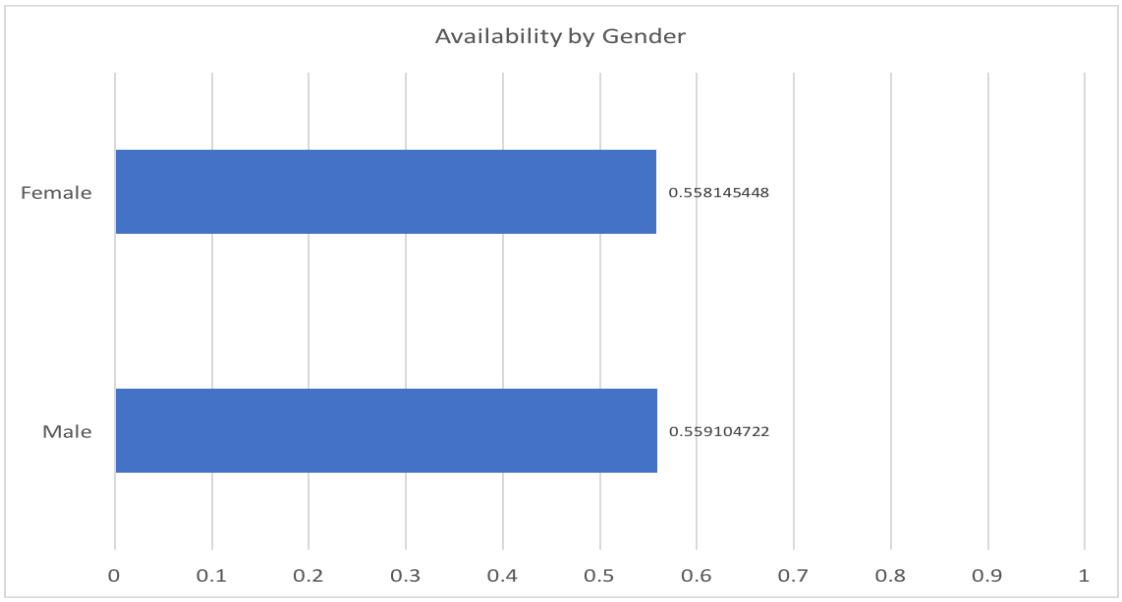


Figure 3-2: Availability rates by Donor Gender. We do not see any difference

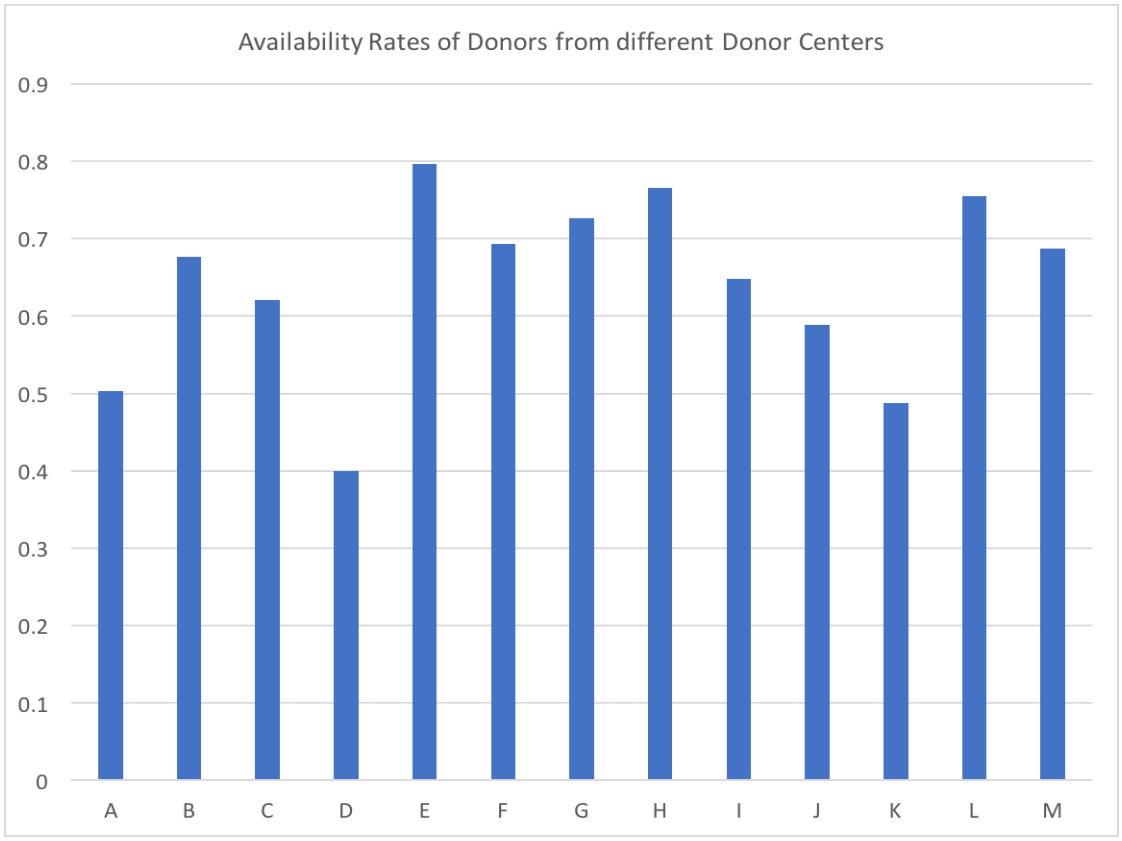


Figure 3-3: Different Donor Centers have different Availability rates. Shown here are the rates for 13 different DCs used in modeling

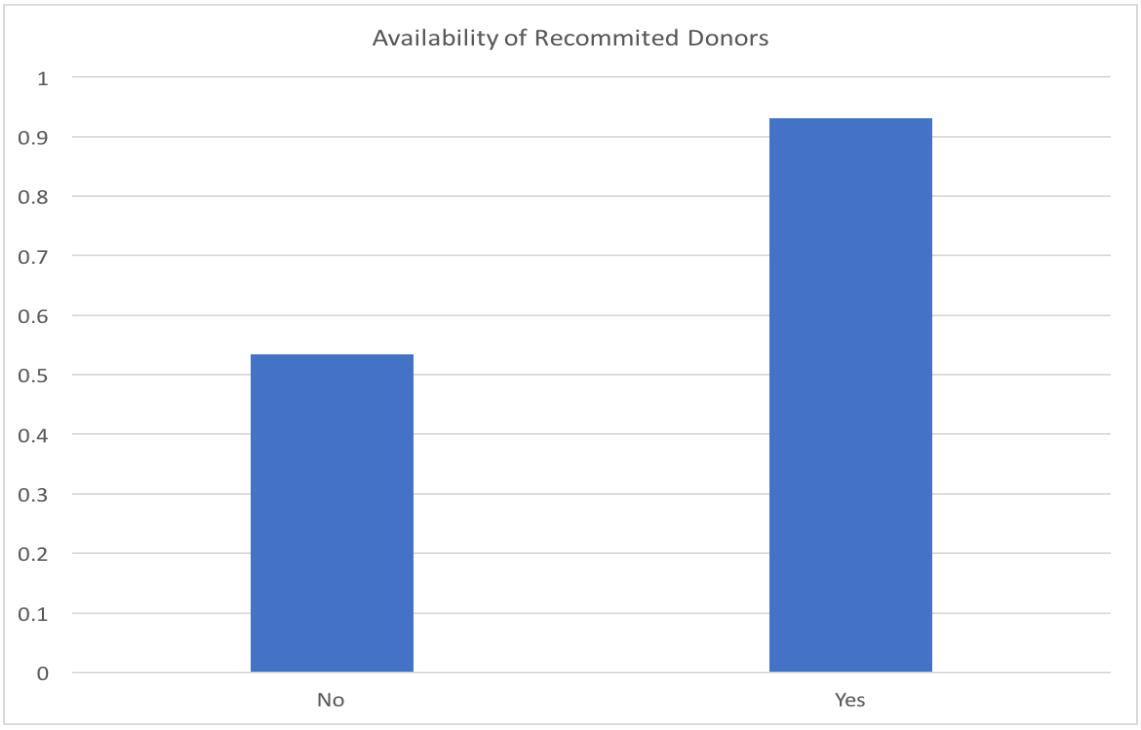


Figure 3-4: Availability Rates by Recommit response

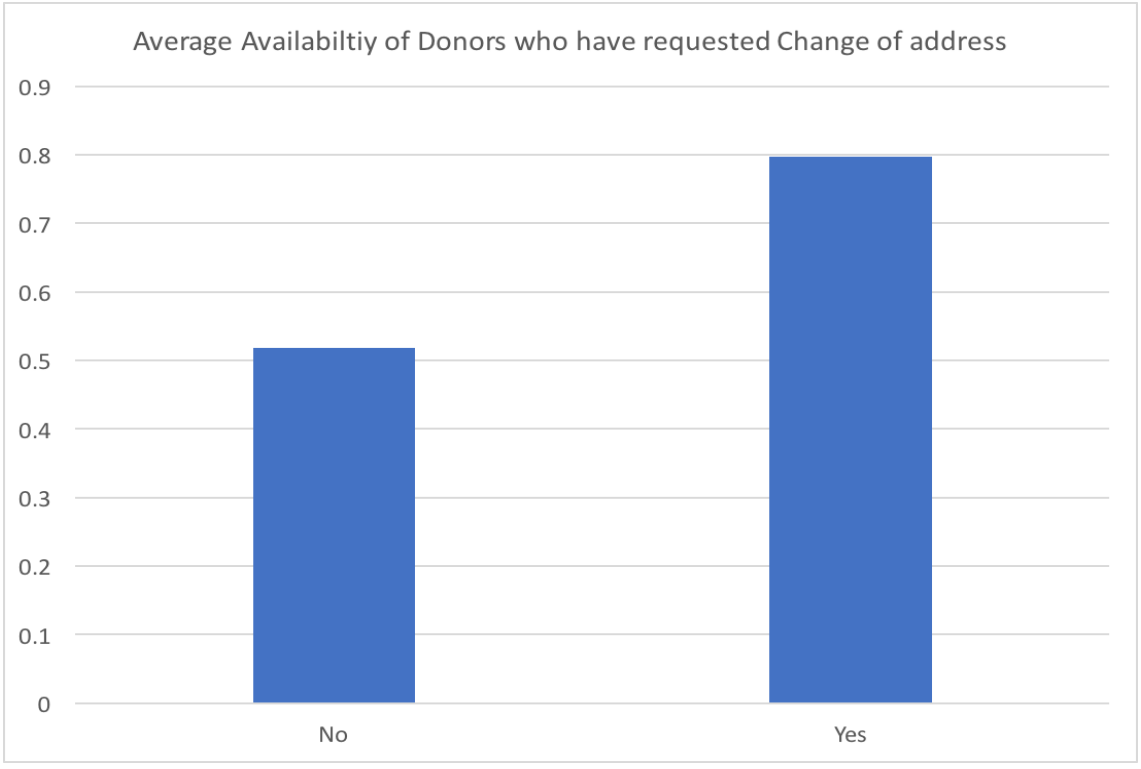


Figure 3-5: Availability Rates by Change of Address Request

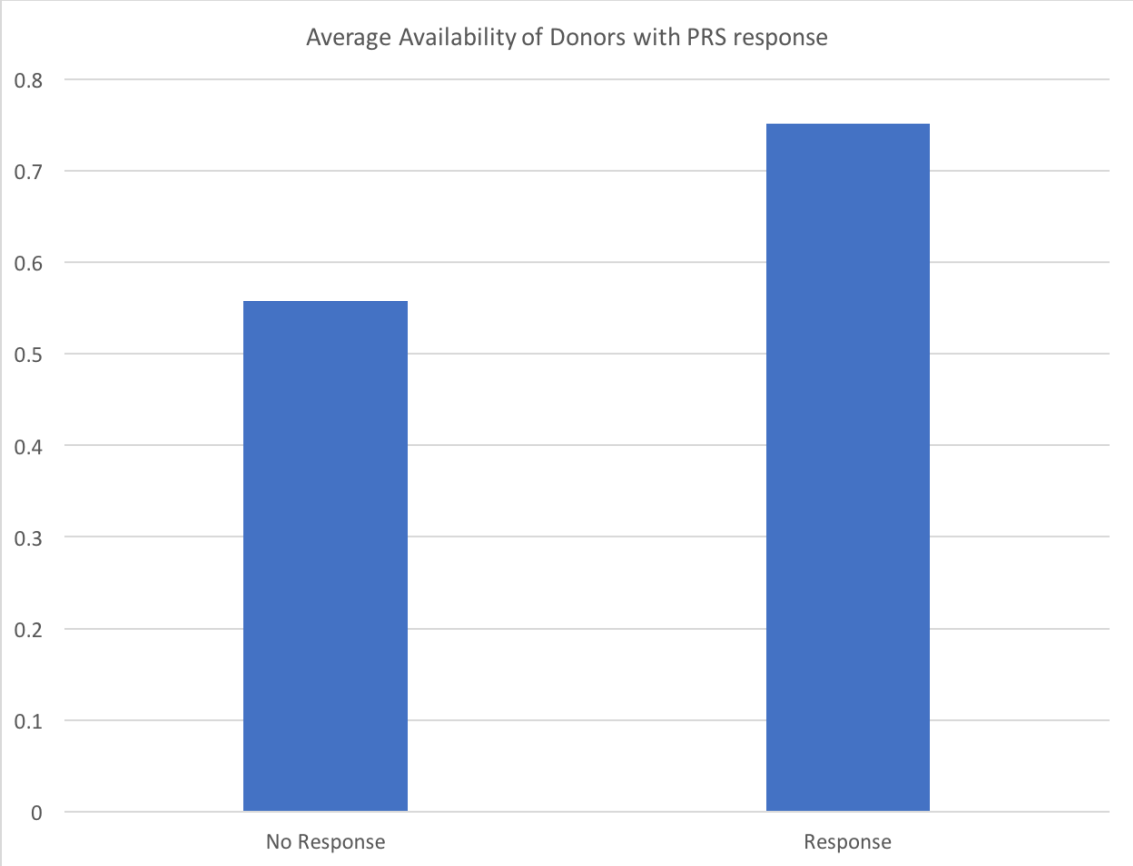


Figure 3-6: Availability Rates by Response to Post-Recruitment Survey

3.3 Problem Formalization

The goal of this modeling effort is to develop a Machine Learning approach for predicting donor availability. The target variable is the donor response (Yes/No) to a CT sample donation request. However, a donors' response can be negative for a variety of reasons – temporarily unavailable, unable to contact, medically unavailable to donate, not interested.

We remove any donors who were temporarily unavailable from analysis, and the other reasons for a negative response are clubbed into a single category and collectively referred to as *Not Available* donors, to develop a more patient-focused model that avoids

modelling sub-categories of unavailable donors. This assumption leads to a binary classification formalization as below. The donors in the dataset are divided into two groups based on their response to a CT request: 1. *Available* (if the requested member agreed to donate), 2. *Not Available* (if the requested member declined to donate). These categories are treated as the target variable.

$$y_i = \begin{cases} 1, & if\ donor\ available \\ 0, & if\ donor\ not\ available \end{cases} \quad 3.1$$

In all, we have 15 features to represent each donor and 178,249 donors in the dataset. The overall availability observed in the data is 56%. This implies we have a reasonably balanced dataset. In all we have 99,558 positive samples and 78,691 negative samples.

Misclassifying a donor likely to donate would involve looking for new donors while misclassifying a donor who may not donate will also result in having to look for a new donor. This means a false positive is equally as expensive as a false negative. Under this assumption we can train a binary classifier with equal mis-classification costs.

3.4 Methods

Input features like Donor Race, DC code, PRS score are all categorical variables with multiple levels. One way to handle categorical predictors is to encode them as binary variables using one-hot encoding strategy. But this strategy increases the dimensionality of the data. Another solution is to use methods that are suitable to handle categorical predictors. Tree based methods can handle categorical data since they do not require additional encoding as a part of preprocessing. Tree based methods also allow for interaction between variables and for non-linear function estimation (Friedman, 2001). Boosted Trees is a popular non-parametric modeling method known to have competitive

performance (Caruana and Niculescu-Mizil, 2006). In particular, we use a variation of Boosted Decision Trees called the Gradient Boosting method (Breiman, 1997).

3.4.1 Gradient Boosting

Empirical and theoretical evidence suggests that there is no single 'best' method for all classification problems. It has been suggested that combining multiple methods provides better generalization. Boosting is one such learning strategy that combines several 'weak' classifiers. A classifier is called a *weak* classifier if it has an accuracy slightly better than random guessing.

Boosting was introduced by (Breiman, 1997; Freund and Schapire, 1997; Friedman, 2001) for classification and was later extended to regression problems. Classification trees are popularly used as base learners in boosting. This has practical benefits since trees can handle mixed input data types, missing values, are insensitive to monotone transformations, and deal with irrelevant inputs (Cherkassky and Mulier, 2007).

Weak classifiers are sequentially applied to different realizations of the data to produce a sequence of m classifiers $g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), g_3(\boldsymbol{x}), \ldots g_m(\boldsymbol{x})$. The final classifier is constructed using the weighted sum of the sequence of the individual classifiers as in Equation 3.2.

$$f(\boldsymbol{x}) = sign\left(\sum_{j=1}^{m} w_j \, g_j(\boldsymbol{x})\right) \tag{3.2}$$

The weights w_j are dependent on the corresponding individual component classifier's training error. Classifiers with lower training error receive greater weights and therefore have more influence on the final combined model.

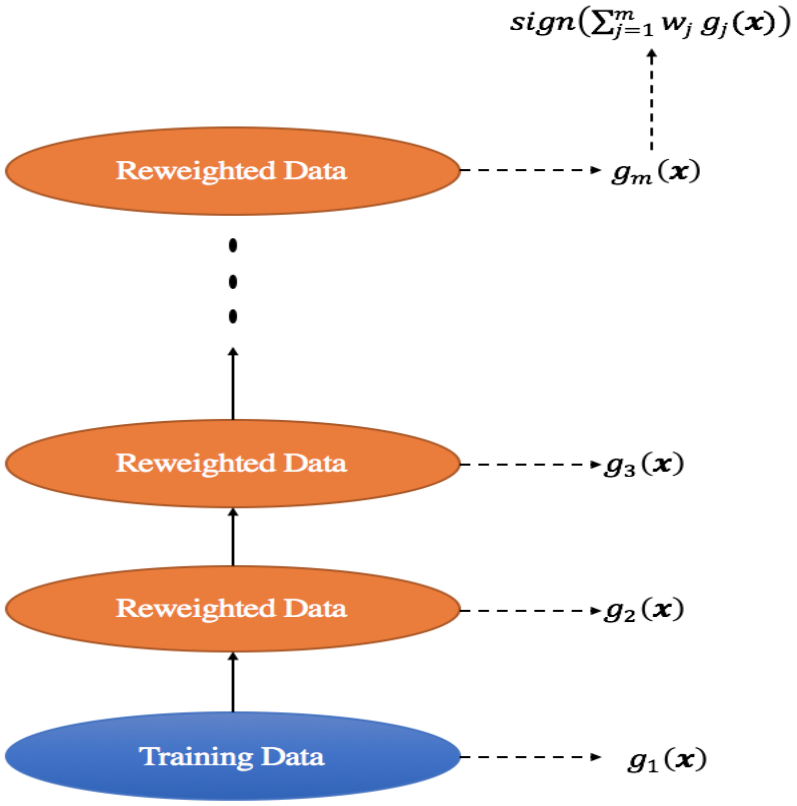


Figure 3-7: Schematic diagram of Boosting methods. Classifiers are sequentially learned on reweighted training data

Boosting can be numerically solved using the Gradient Descent algorithm. This method is called *Gradient Boosting Method* (GBM). The optimization here is performed in the function space instead of the usual parameter space. The update for sequential model training is chosen to be the negative of the direction of steepest descent. Updates at each step are given by:

$$g_j = g_{j-1} - \rho_j h_j \tag{3.3}$$

$$h_{ij} = \left[\frac{\partial L(y_i, g(\boldsymbol{x_i}))}{\partial g(\boldsymbol{x_i})}\right]_{g(\boldsymbol{x_i}) = g_{j-1}(\boldsymbol{x_i})} \tag{3.4}$$

h_m is the stage-wise update considered by applying the above gradient for all *N* data points in the training set. The loss function $L\big(y, f(\boldsymbol{x})\big)$ must be differentiable for the above condition to be valid. ρ_m is a scalar. For smooth loss functions it can be shown that successive models are fit on residuals from the previous stage (Hastie, Tibshirani and Friedman, 2001). A simplified version of the algorithm is described next. The recommended loss function for a binary classification problem in the *gbm* package (Ridgeway, 2006) is shown below (Equation 3.5). The gradient for the loss function in 3.6.

$$L(y, g(\boldsymbol{x})) = -2\sum_{i=1}^{n} y_i\, g(\boldsymbol{x_i}) - \log(1 + \exp\big(g(\boldsymbol{x_i})\big)) \qquad 3.5$$

$$h_i = \left[y_i - \frac{1}{1 + \exp(g(\boldsymbol{x_i}))}\right] \qquad 3.6$$

	Gradient Boosting Algorithm
1	Fit $g_0(\boldsymbol{x})$ on Training data $\{(\boldsymbol{x}_1, y_1),\ (\boldsymbol{x}_2, y_2), (\boldsymbol{x}_3, y_3), \ldots, (\boldsymbol{x}_n, y_n)\}$
2	For j = 1 to *m*
	(a) For *i* = 1, 2, 3,…, *N* compute residuals $r_{ij} = -\left[\frac{\partial L(y_i, g(\boldsymbol{x_i}))}{\partial g(\boldsymbol{x_i})}\right]_{g(\boldsymbol{x_i}) = g_{j-1}(\boldsymbol{x_i})}$
	(b) Fit $g_j(\boldsymbol{x})$ on data $\{(\boldsymbol{x}_1, r_{1j}),\ (\boldsymbol{x}_2, r_{2j}), (\boldsymbol{x}_3, r_{3j}), \ldots, (\boldsymbol{x}_n, r_{nj})\}$
	(c) Update $f_j(\boldsymbol{x}) = f_{j-1}(\boldsymbol{x}) + \eta g_j(\boldsymbol{x})$
3	Output $\hat{f}(\boldsymbol{x}) = f_m(\boldsymbol{x})$

Classifiers are commonly trained to assign hard labels on data instances. For this application, it is more beneficial to assign a real-value that can be used to identify availability. Boosted tree methods provide a normalized score for each donor in the range of [0,1].

3.4.2 Tuning Parameters

The update function at each stage for GBM is $f_j(\boldsymbol{x}) = f_{j-1}(\boldsymbol{x}) + \eta g_j(\boldsymbol{x})$, where η is called the learning rate (or shrinkage parameter). η is a scalar and it controls for the rate at which the boosting algorithm scales the contribution of new functions (trees) when it is added to the current approximation. (Friedman, 2001) states that smaller values of η favor better test error and is shown to provide improvements over no shrinkage (η = 1). However, there are no analytical rules to calculate the parameter. Smaller update values create a computational constraint as more terms are needed for function approximation.

Another important parameter that determines the optimal generalization of the model is the size (number of nodes) of individual trees in the model. Historically, at each stage, large trees were inducted and then pruned with a bottom-up strategy. This strategy is computationally expensive and was shown to degrade performance (Hastie, Tibshirani and Friedman, 2001). A simpler strategy is to restrict all trees to be the same size, J. Thus, J becomes a boosting model parameter that needs to be tuned for optimal performance. For $J > 1$ individual trees are shown to be functions of J predictor variables. J is also called the *Interaction Depth* of trees.

3.5 Results and Discussion

3.5.1 Model Selection and Experimental Setup

For reliable error estimates *test* error is always measured on out-of-sample data, that is, data that is not used to develop the model. Random sampling is used to split the data into *training* and *test* sets. 25% (44,544 samples) of the data is set aside for testing, and the 5-fold cross validation is performed on the remaining 75% (133,705 samples) of data (double resampling strategy) for model selection. A 5-fold cross validation is used in the

training set to tune parameters. The two most important parameters – Learning Rate and Interaction Depth – are tuned. Validation errors are based on the standard classification error metric (percentage of misclassified samples). Table 3-2 shows average validation errors for GBM from 5-fold cross validation. The parameters that correspond to the lowest errors is used for the final model, and is highlighted in Table 3-2.

Table 3-2: Average Validation Error rates for Gradient Boosting Parameter Selection

		Learning Rate				
Interaction Depth		***0.001***	***0.01***	***0.10***	***0.25***	***0.50***
	1	0.4396	0.2986	0.2951	0.2951	0.2959
	2	0.4390	0.2958	0.2933	0.2930	0.2934
	3	0.2987	0.2957	0.2929	0.2927	0.2939
	4	0.2987	0.2948	0.2923	0.2926	0.2927
	5	0.2987	0.2936	0.2922	0.2929	0.2935
	6	0.2987	0.2926	0.2917	0.2923	0.2928
	7	0.2987	0.2923	0.2910	0.2920	0.2957
	8	0.2987	0.2926	0.2918	0.2929	0.2939
	9	0.2987	0.2920	0.2929	0.2935	0.2961
	10	0.2986	0.2925	0.2920	0.2939	0.2956

3.5.2 Experimental Results

For comparison purposes, we repeat the experimental setup described in the previous section with Logistic Regression and Linear SVM. Run-times for non-linear SVM (rbf kernel) were too long (several weeks) and were hence not considered. The classification accuracies for the 3 methods are shown in Table 3-3. We notice Boosted Trees had significantly lower error rates and further analysis is based on this model.

Table 3-3: Training and Testing Accuracy of models measured on test dataset

	Training Accuracy	***Testing Accuracy***
Logistic Regression	0.64	0.62
Linear SVM	0.65	0.63
Boosted Trees	0.73	0.70

Modeling is done in R using the *gbm* package (Ridgeway, 2006) for Gradient Boosting and *LiblineaR* (Fan *et al.*, 2008) for SVM and Logistic Regression.

gbm provides a variable importance plot from the model that is estimated by calculating the classification improvement provided by each variable split in the trees in the model. See (Hastie, Tibshirani and Friedman, 2001) for details. Figure 3-8 shows the variable importance for the trained model. We notice the variables with highest influence are all categorical variables. Raw data in the Data Section (Section 3.2.1) support relative influences derived from the model. Variables with significant marginal differences in availability have a stronger impact in the model. Interactions were also allowed in the model, which contributes to the relative influence values.

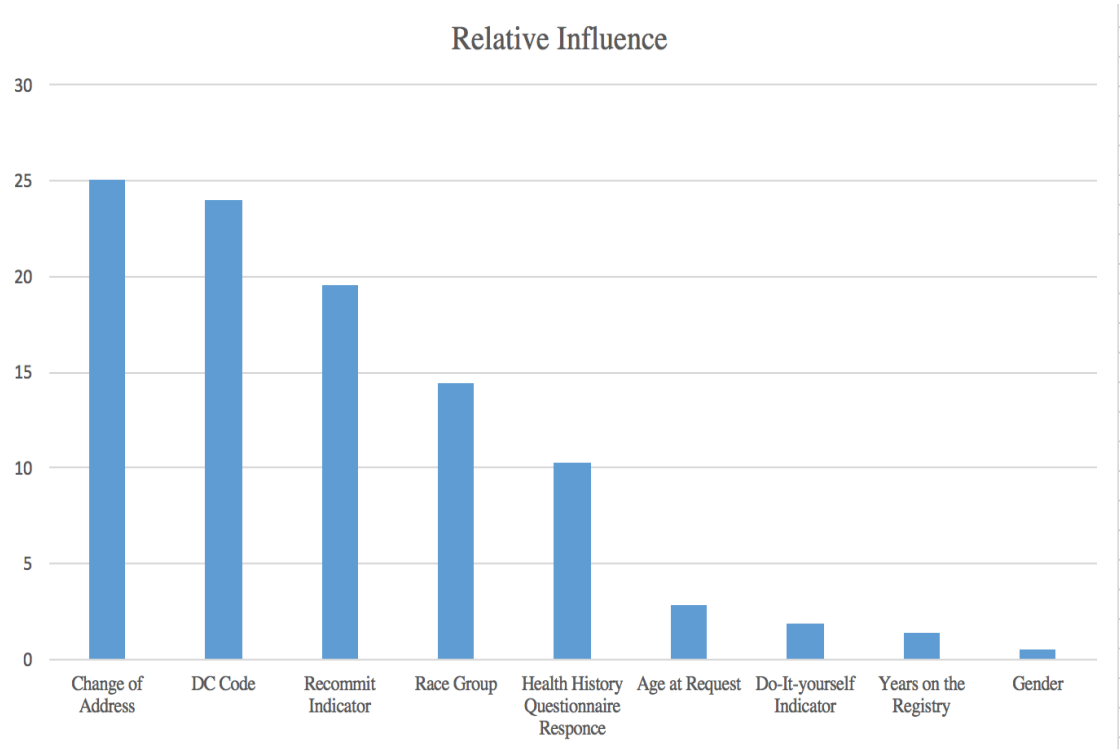


Figure 3-8: Variable Importance plot for the trained model. Other variables listed in Section 2 had lower weights and are not plotted here.

3.5.3 Discussion

The GBM model assigns every donor a score in the 0-1 range. A threshold (typically 0.5) is predict labels to the donors as *available* or *not available*. However, we can use the scores directly to estimate donor availability. To demonstrate effectiveness of the trained model, we show average availability rates broken down by model assigned score ranges in the test set in Figure 3-9. We have binned donors based on these assigned scores and measured the average availability rate within each bin. We notice that among the donors that had a score of between 0.9 and 1.0 observed availability was 93%, marked by the rhombus shaped points on the graph. This (almost) linear relation between model assigned score and average availability holds true for all the brackets as shown in Figure 3-9. Consequently, model assigned scores can hence be used as a direct indicator of a donor's

availability. The bar graph represents the number of donors who are assigned scores in the range shown on the x-axis. Average availability and number of donors in each score brackets are represented on different scales on the y axis as shown respectively on the left and right side of the Figure.

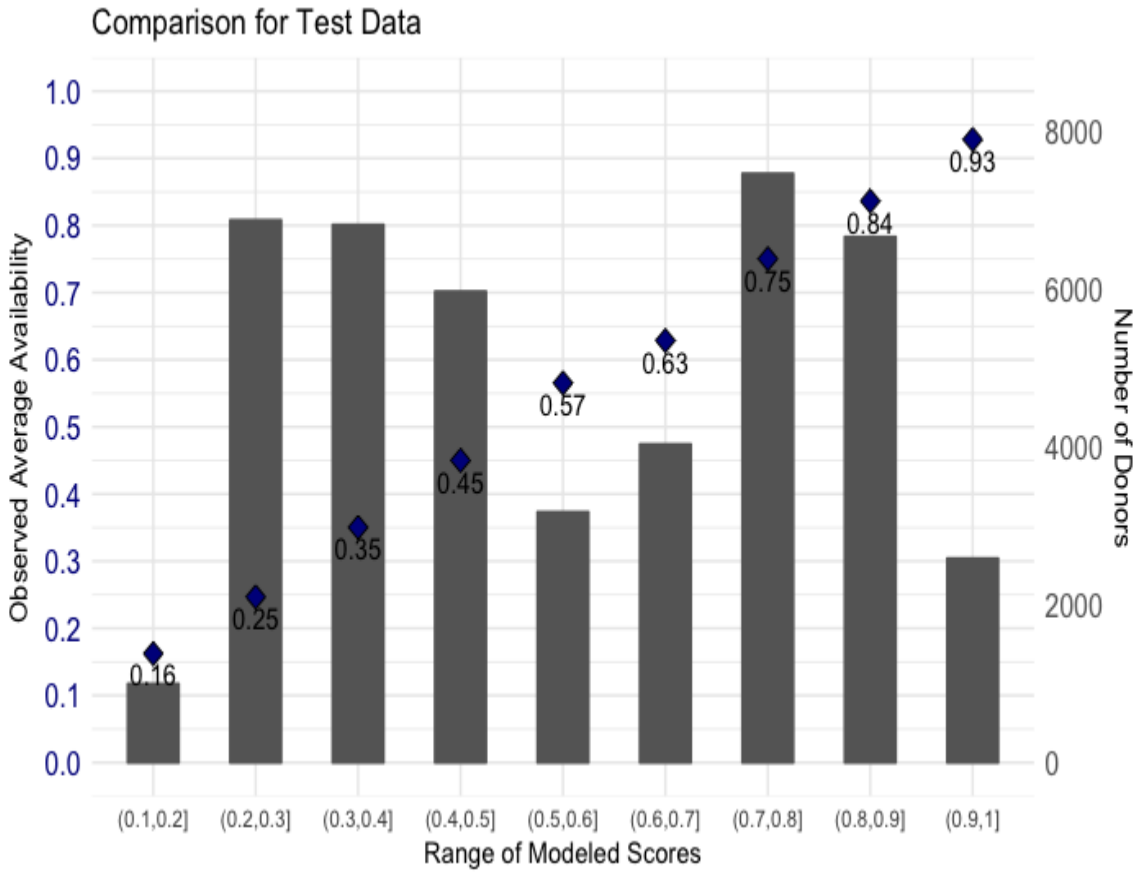


Figure 3-9: Observed availability rates compared to model assigned scores in the test set. The overlaid number in each column represents the average observed availability of donors with modeled scores in the corresponding range. The bar graphs are the number of donors who are in the brackets noted on the x-axis

The NMDP procures donors from several sources. Responses to outreach programs are only recorded for donors within the NDMP controlled network. This results in donors with several different levels of information available for scoring modeled availability. We overcome this problem with careful data encoding to indicate presence or absence instead of using missing values in the modeling set. The model is hence capable of adjusting for varied levels of information and assigns a score to every member in the registry.

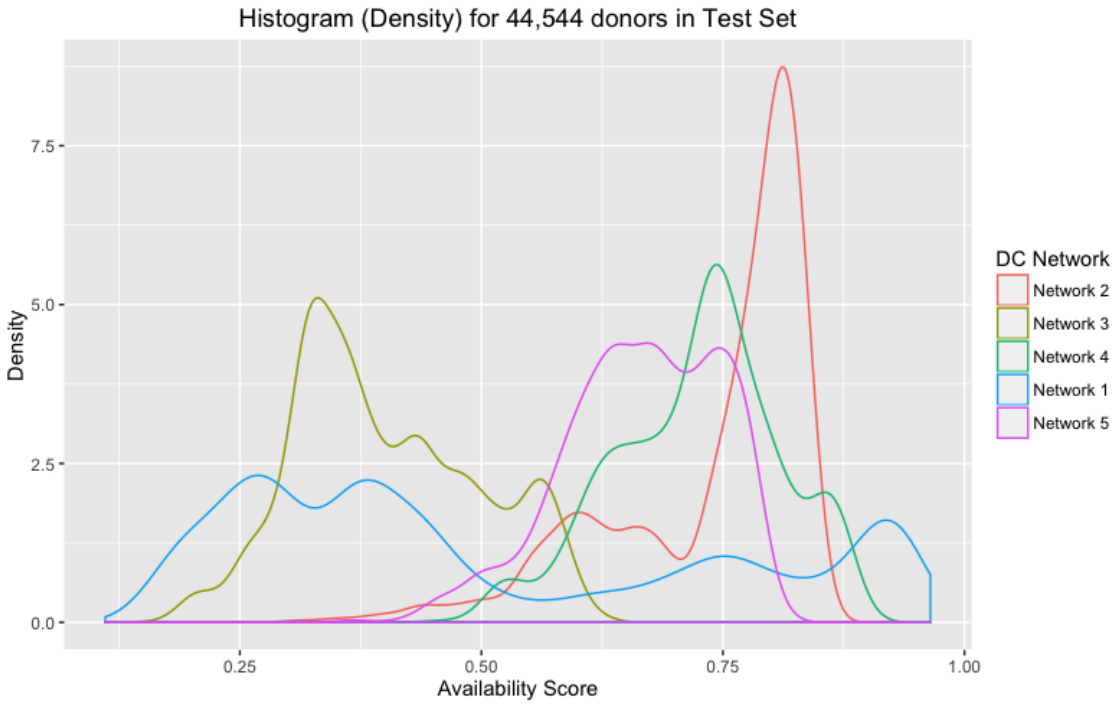


Figure 3-10: Density of donor availability for different donor networks.

Figure 3-10 shows the density of donors across each of the different network sources. The peaks on these densities are affected by the spread of assigned values. We notice *Network 2* donors have a consistent score, which is reflected by a *taller peak* in the graph, whereas *Network 1* donors show more variation. Donors within the NMDP network tend to have more associated information, making it possible for the model to differentiate them. This results in a wider spread in the assigned scores and a better ability to differentiate donors.

Information from this modeling can be integrated directly into the Donor Display Tool. In Figure 3-11 we show how we can use the two modeling efforts to make donor selection easier. The horizontal axis is the Selection Score (as described in the previous Chapter) and the vertical axis is the Availability Score. The most desirable donors are the

top right corner. As we can notice, using the two models will reduce the search field from thousands of donors to a handful instantly.

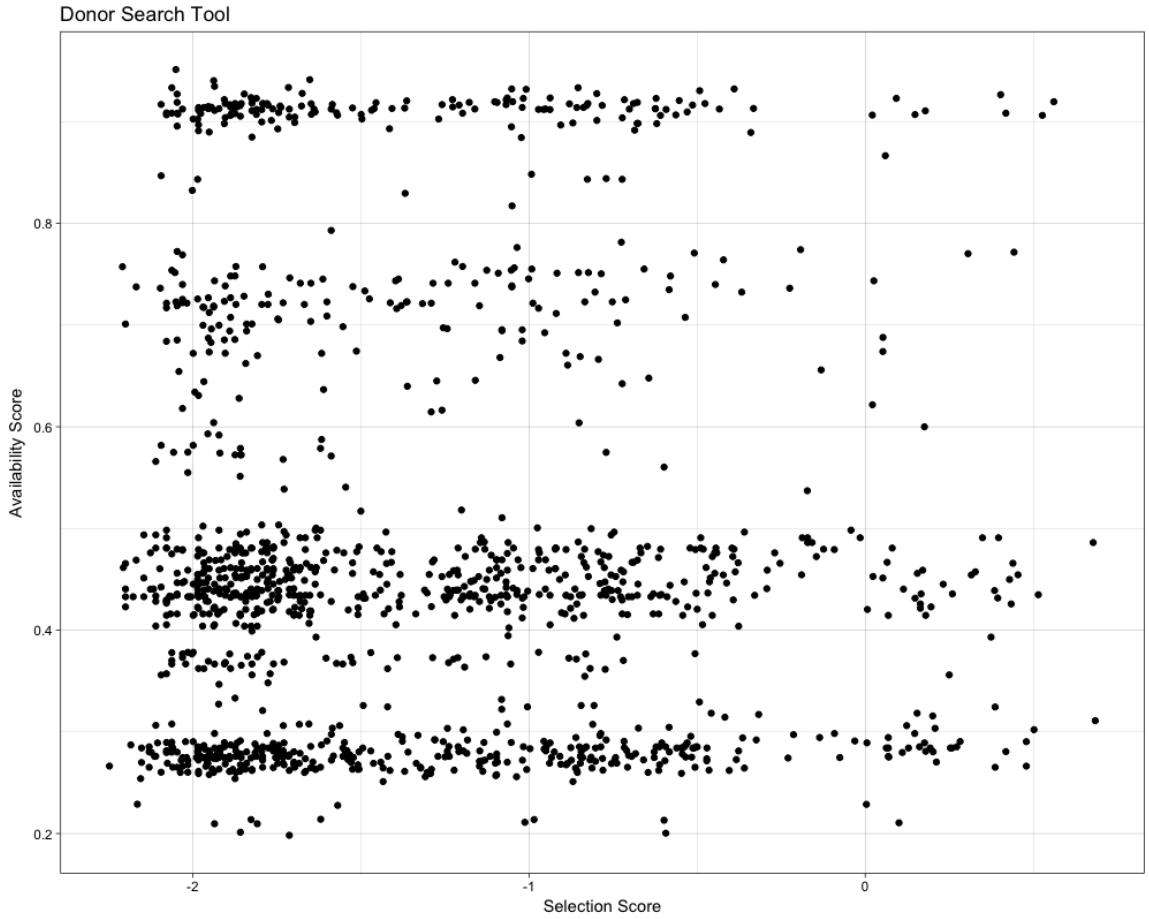


Figure 3-11: Integrating Availability Score and Selection Score into the Donor Selection Tool. Shown here is an actual search with 1341 matched donors.

In conclusion, experts have had to rely on historical averages for availability estimates. This results in either overestimating or underestimating an individual donor's availability. This also limits the ability of clinicians to accurately estimate how many donors should be contacted in order to ensure the patient will be supplied with a bone marrow transplant product when needed. The proposed system helps in providing a point-wise estimate for each donor based on the set of factors available for estimating their availability. Creating a single score for each potential donor can significantly improve selection efficiency.

Chapter 4 Donor Utility and Recruitment

4.1 Introduction

Registry diversity is an important property of volunteer registries. Diversity is measured by the number of copies of unique HLA types present in the registry. This distribution of HLA types is called the *Genotype Frequency* distribution. A diverse registry is beneficial in two ways: (1) it affects the ability to serve patients of different ethnicities; (2) avoids waste in terms of large number of over-represented (HLA types) donors. It is often noted that a large number of patients who require stem-cell transplants do not find a match (related or unrelated donors) and a large number of registered donors are never used. The primary reason for this is the genotype composition of a registry. A large body of studies highlight the need for increasing the diversity of registries to meet the growing demand of stem-cell transplants (Kollman *et al.*, 2004; Hurley *et al.*, 2007; Gragert *et al.*, 2014; Buck *et al.*, 2016; Dehn *et al.*, 2016).

The genotype distribution of the registry also determines the number of potentially matched donors a patient can likely find in the registry (Gragert *et al.*, 2013, 2014). As we noted in the Data section (Section 2.2) for Donor Selection modeling, a search at the NMDP can have more than 40,000 identically matched donors for a patient with common HLA type. Having multiple copies of the same HLA type can help accommodate for donor availability and typing ambiguity issues. However, maintaining a registry where greater than 50% of the registry have excess copies of the same HLA type is not optimal. Substantial resources are spent maintaining an active donor registry. Only a handful of donors are required to successfully complete a search and a large portion of donors are never utilized. We do not have metrics to determine donor *utility* on a search. It is evident that donors' utility is dependent on multiple factors – genetic and non-genetic factors. Determining the utility of a donor involves manually analyzing all donor related information. Given the size of modern registries, this is an impossible task.

In preceding Chapters (Chapter 2,3) we have modeled donor selection and availability. Both of these are important components for donor utility and both models are based on donors' non-genetic factors. All the donor characteristics are used in these models. In addition to *selection* and *availability*, donors' HLA type is also important to determine utility. For example, a donor who has a really common HLA type has a small likelihood of being used for transplant and in contrast, a donor with a rare HLA type will most likely be used for transplant.

Donors, typically older recruits with lower resolution typing, are associated with multiple HLA types each with a corresponding probability of being the true genotype. The process of assigning multiple genotypes to a donor is call *ambiguous genotype assignment*. Statistical methods such as expectation-maximization (EM) are employed to determine the probabilities of ambiguous HLA types (Kollman *et al.*, 2007; Karnes *et al.*, 2017). This is commonly referred to as *HLA Imputation* (Madbouly *et al.*, 2014). Imputation output for a donor in the registry is shown in Table 4-1. The donor (in Table 4-1) has 5 possible genotypes with ambiguities noted in DRB1 and DQB1 genes. This donor will be identified as a match for patients with any of the 5 genotypes. In such cases, the donor *utility* should account for the genotype ambiguity.

Table 4-1: Imputation Output for a donor based on typed DNA information with five possible phased genotypes and their Haplotype Frequencies

Predicted Race	Predicted Phased Haplotypes	Predicted Frequencies
CAU	A*03:01~C*04:01~B*35:01~DRB1*01:01~DQB1*05:01	0.011755
	A*03:01~C*04:01~B*35:01~DRB1*04:04~DQB1*03:02	3.0923E-4
CAU	A*03:01~C*04:01~B*35:01~DRB1*01:01~DQB1*05:01	0.011755
	A*03:01~C*04:01~B*35:01~DRB1*04:02~DQB1*03:02	2.4768E-4
CAU	A*03:01~C*04:01~B*35:01~DRB1*01:01~DQB1*05:01	0.011755
	A*03:01~C*04:01~B*35:01~DRB1*04:04~DQB1*04:02	5.5973E-6
CAU	A*03:01~C*04:01~B*35:01~DRB1*04:04~DQB1*03:02	3.0923E-4
	A*03:01~C*04:01~B*35:01~DRB1*01:03~DQB1*05:01	1.6042E-4
CAU	A*03:01~C*04:01~B*35:01~DRB1*04:02~DQB1*03:02	2.4768E-4
	A*03:01~C*04:01~B*35:01~DRB1*01:03~DQB1*05:01	1.6042E-4

The NMDP actively looks to identify donors for expensive high-resolution typing, which reduces the genotype ambiguity and improves search results. Efforts are also underway to identify donors for additional DPB1 typing, which is being extensively used for donor selection (Fleischhauer *et al.*, 2012; Shaw *et al.*, 2013, 2014). Both of these, and other registry management tasks, are performed to improve donor search quality and involves significant financial burden on the registry. A utility score that combines donor genetic and non-genetic factors will help in identifying high impact donors for such registry management tasks to focus resources. We develop a mathematical framework that combines the donor selection score, availability score, and HLA information to determine the *utility* of donors.

(Gragert *et al.*, 2014; Dehn *et al.*, 2016) reflect on the importance of diversity in donor registries in terms of finding a match for patients of different races. At the time

(Gragert *et al.*, 2014) study was done, the likelihood of finding a perfectly matched available donor for patients of White European descent was 75% and for Black American patients was between 16% and 19%. Patients of other ethnic groups fall in-between. Enhancing donor diversity will improve these match likelihoods. However, improving registry diversity is a difficult task. Genetic information is only obtained after a donor is added to the registry and typing is performed. Using population genetics to improve diversity has been shown to be ineffective in improving donor diversity (Hurley *et al.*, 2007). We provide a data-driven solution to *intelligent* donor recruitment based on geo-coded information of donors.

4.2 Material and Methods

4.2.1 Utility Scoring

We now develop a mathematical framework to assign a *utility score* to each donor that is based on the three donor factors:

1. Imputed HLA
2. Non-genetic secondary factors
3. Availability

Consider a Donor $\boldsymbol{i}$ who is identified as a match for a patient with genotype $\boldsymbol{k}$. Let the probability of a successful donation for donor $\boldsymbol{i}$ with respect to genotype $\boldsymbol{k}$ be $\boldsymbol{\lambda_{ki}}$. $\boldsymbol{\lambda_{ki}}$ can be defined as:

$$\lambda_{ki} = C_{i|k} * A_i \tag{4.1}$$

where, $\boldsymbol{C_{i|k}}$ is the probability of donor $\boldsymbol{i}$ being *Chosen* for a patient with genotype $\boldsymbol{k}$ and $\boldsymbol{A_i}$is the probability of donor $\boldsymbol{i}$ being *Available* for a CT request. $\boldsymbol{C_{i|k}}$ can be calculated from the donor selection model and $\boldsymbol{A_i}$ is directly available from the donor availability model.

The Donor Selection model, as described in Chapter 2, is an SVM based model. SVM projection distances can be used to approximate probabilities of the form $p(y = 1 \,|\boldsymbol{x})$ using a simple logistic transformation (Platt, 1999) , where $y = 1$ represents chosen donors. Let us call $\boldsymbol{\lambda_{ki}}$ to be the *Utility* of Donor $\boldsymbol{i}$ with respect to the patient with Genotype $\boldsymbol{k}$. If there are $\boldsymbol{p}$ number of donors on the registry who match genotype $\boldsymbol{k}$, then the total utility for $\boldsymbol{k}$ is:

$$\lambda_k = \sum_{i=1}^{p} \lambda_{ki} \tag{4.2}$$

By Poisson approximation, the probability of unsuccessful transplant (i.e., no donation happens) is shown below:

$$P_{\overline{transplant}} = e^{-\lambda_k} \tag{4.3}$$

Hence, the probability of a successful transplant is:

$$P_{transplant} = 1 - e^{-\lambda_k} \tag{4.4}$$

If donor $\boldsymbol{i}$ wasn't in the search, the probability of successful transplant is given by:

$$P_{transplant} = 1 - e^{-(\lambda_{k-}\lambda_{ki})} \tag{4.5}$$

Hence, the marginal utility of donor $\boldsymbol{i}$ for Genotype $\boldsymbol{k}$ is

$$\Delta_{i\,|\,k} = [(1 - e^{-\lambda_k}) - (1 - e^{-(\lambda_{k-}\lambda_{ki})})] \tag{4.6}$$

We have seen earlier that the same donor can be a match for multiple patients with different genotypes. Thus, the total *utility* of donor $\boldsymbol{i}$ is the sum of all marginal utilities and is given by (4.7).

$$\Delta_i = \sum_{\substack{over\ all\ \boldsymbol{k} \\ for\ donor\ i}} \Delta_{i\,|\,k} \tag{4.7}$$

The probability of donor $\boldsymbol{i}$ being chosen, represented by $\boldsymbol{C_{i|k}}$, incorporates multiple donor factors. Most importantly, the Donor Selection model includes match grade (10/10, 9/10, or 8/10) and the match probability of donor $\boldsymbol{i}$ being a match for genotype $\boldsymbol{k}$. The value $\boldsymbol{C_{i|k}}$ appropriately calibrates based on these match dependent features.

4.2.2 Intelligent Donor Recruitment

Internal studies at the NMDP have shown that optimal HLA copies (copy count) on the registry is between 3 and 7 for a successful transplant. For this study, we call all genotypes with less than 4 copies in the registry as *rare* and genotypes with more than 4 copies as *common*. A random sample set of 247,938 donors who have a registered address in the database were collected for modeling. Apart from the standard demographic information available, we also gathered geographic specific information based on their addresses. ArcGIS software, developed by ESRI (https://www.esri.com/en-us/home), was used to obtain geo-coded information for donors based on their addresses. This geo-coded information based on reported address has been used as proxy for Race and Ethnicity, Socioeconomic Status of patients and donors for stem-cell related studies previously (Besse *et al.*, 2015; Shaw *et al.*, 2015). These studies use the following two variables from the geo-coded information obtained from ESRI.

1. Designated Market Area (DMA): Each zipcode in the country is assigned to a Market area, which is usually a nearby larger metropolitan area, that is most popularly used by Broadcasting (TV and Radio) audience measurements. Based on donor's zipcode, appropriate DMAs are assigned. We use this information as a categorical variable with 205 different levels.
2. Dominant Tapestry Segment: Tapestry Segmentation, provided by Esri, provided a description of neighborhoods in the country based on demographic and socioeconomic composition. The population is divided in 67 distinctive segments. We use this information as a categorical variable with 67 levels.

(Besse *et al.*, 2015) looks at unmet stem-cell donation requests based on similar geo-coded information.

Problem Formalization: We identify donors who have a rare genotype as defined above. The distribution of genotype frequencies in the dataset is shown in Figure 4-1. Any donor who has genotype of copy count less than 4 is called as rare. This translates to 50% of the collected dataset being labelled as rare.

$$y_i = \begin{cases} 1, & donor\ i\ has\ a\ rare\ genotype \\ 0, & otherwise \end{cases} \tag{4.8}$$

Each donor is represented by a feature vector of dimensionality of 18, which includes donor demographics and geo-coded information.

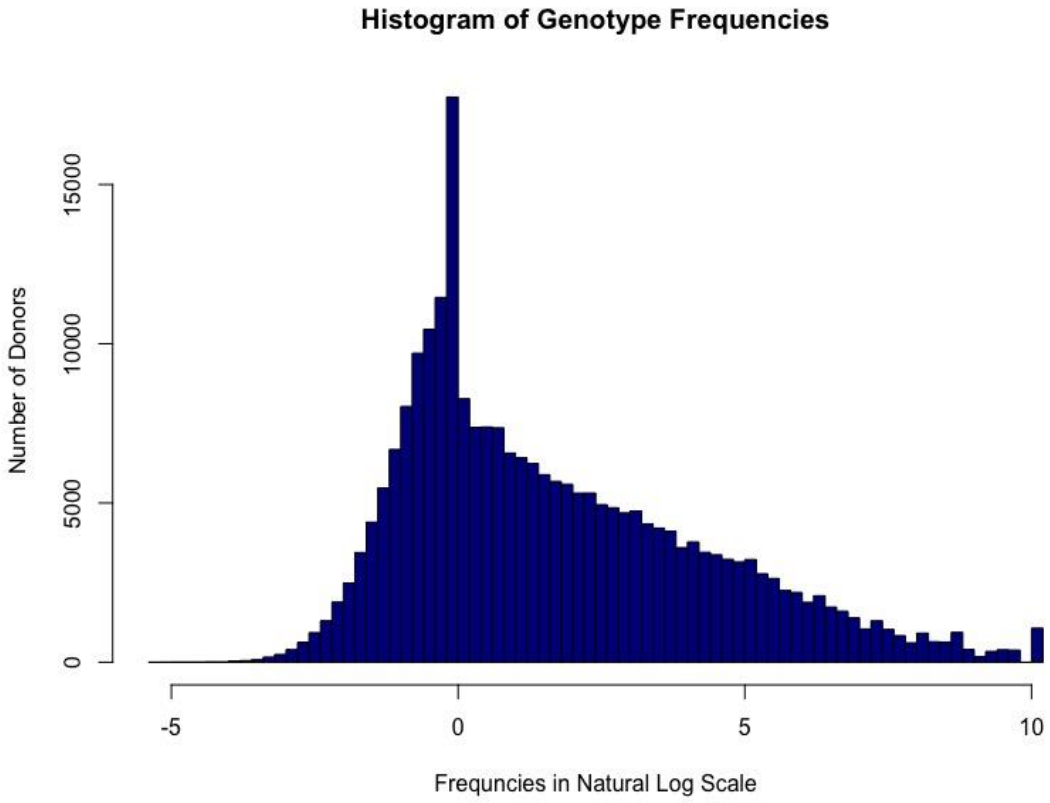


Figure 4-1: Histogram of Genotype Frequencies of Donors represented on a natural log scale.

As in donor availability modeling, we use boosted trees since the data has a large number of categorical variables. However, we use decision stumps (trees with only single split) as the base learner. Hence, individual classifiers in the boosted model can be viewed as indicator functions of the form $g_j(\boldsymbol{x}) = I_{L_\alpha}(C_Z)$, where C_Z is a categorical variable in the data with levels indexed by L_α. Boosted Trees are shown to have the same structure as Additive Models (Hastie, Tibshirani and Friedman, 2001; Cherkassky and Mulier, 2007). This allows us to perform sensitivity analysis to evaluate the impact of input features in predicting the target variable. Experimental procedure here is the same as in availability modeling (see Section 3.5.1).

4.3 Discussion

The developed formalism for evaluating donor utilization combines all aspects of donor characteristics - donor selection score based on matching information and secondary characteristics; donor availability score based on demographic information; and genotype information. The assigned utility score is an indicator of donors who are most likely to be utilized for a successful transplant. Donors who have a rare genotype with favorable secondary characteristics and higher availability will have a higher utility than donors with a really common genotype with unfavorable secondary characteristics and lower availability. Registry wide efforts, such as additional DPB1 typing, can be made cost effective by focusing only on donors with high utility scores.

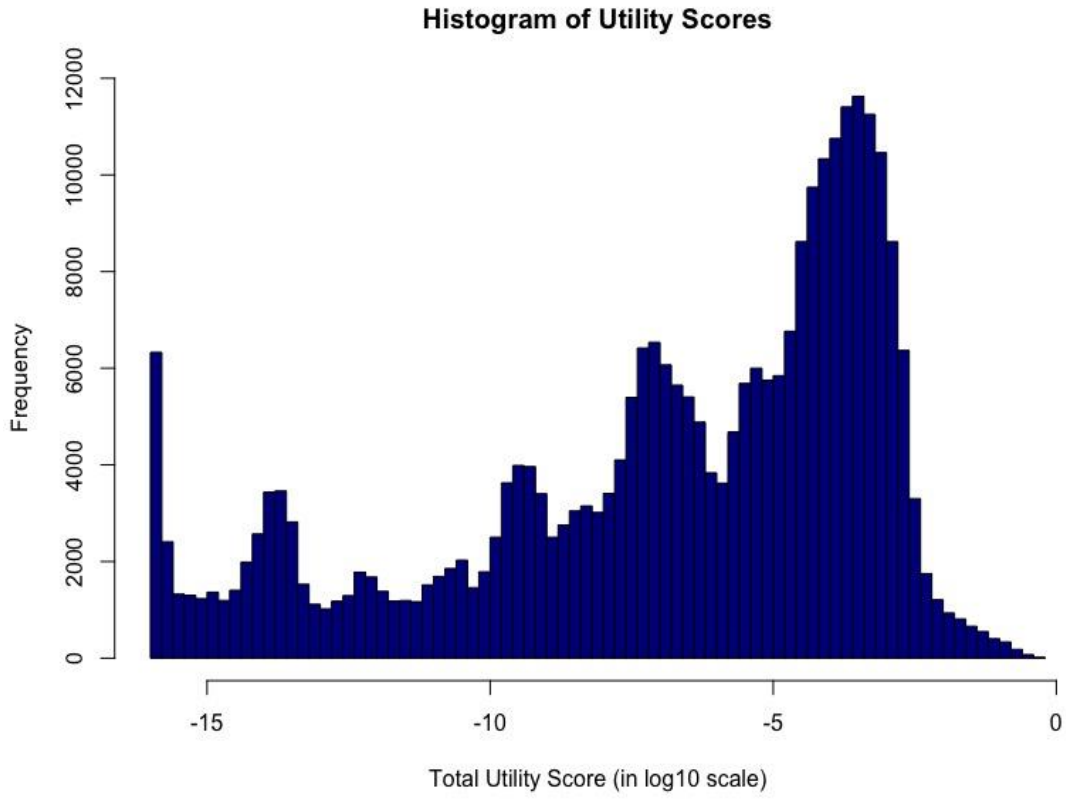


Figure 4-2: Histogram of Donor Utility Scores.

Figure 4-2 shows the utility score distribution of donors in the donor selection model dataset (see Section 2.2). A total of 287,095 donors in the test data were used to generate the histogram. Further investigation showed that donors from smaller searches (smaller number of matched donors) typically have higher utility scores. Based on imputation data, the same formalism can be extended to all donors in the registry to determine their utility.

The NMDP pools donors from other donor registries and actively seeks out new collaborations. This is done to increase the diversity of the registry to help meet the needs of presently underrepresented groups in the registry. Adding donors from another registry involves tremendous costs for the registry. Typically, donors added from another registry have imputed genotype information, unlike donors who are directly recruited. While genotype frequencies vary largely between registries, NMDP also has one of the world's

most diverse registries. As we have seen earlier, we only need a handful of donors to successfully complete a transplant. Currently there are no metrics to measure what the *value* is for adding new members from another registry, i.e., we cannot measure what the incremental utility is for new donors that are being added from other registries. The utility score helps in quantifying additional value a new member adds from another registry. The presented scoring mechanism can be used to identify a shortage of *valuable* donors based on genotypes and determine the impact of adding new donors from other registries.

Donor recruitment is one of the most important functions in maintaining a viable registry. Active recruitment is necessary to provide productive searches with favorable donors. Recruitment is also necessary to improve diversity of the registry. However, following the same practices of recruitment can adversely affect the diversity. Previous studies have shown that there is a correlation between geographic location and genotype frequency distribution. By modeling the rare and common genotypes based on geographic information we can identify specific demographics and locations for targeting donor recruitment. Sensitivity analysis on the trained model is used to compare the effect of individual variables on the output. We adopt this strategy as a means to help with intelligent recruitment. By holding all but one variable constant, we can analyze the effect of a single variable on the final output. This is commonly referred to as partial effects in regression analysis.

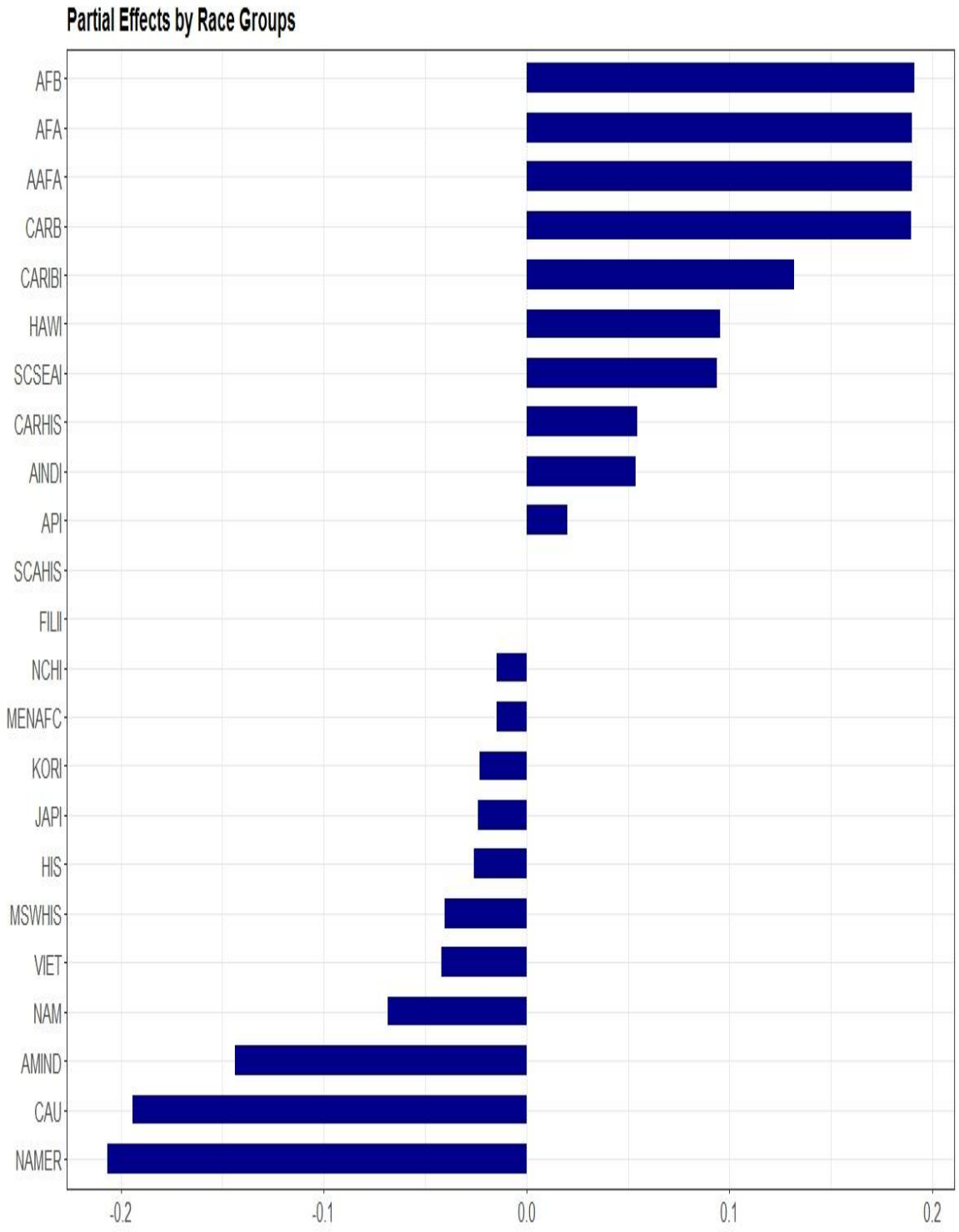


Figure 4-3: Partial Effects of Race Groups. A finer level of Race grouping is used for this modeling.

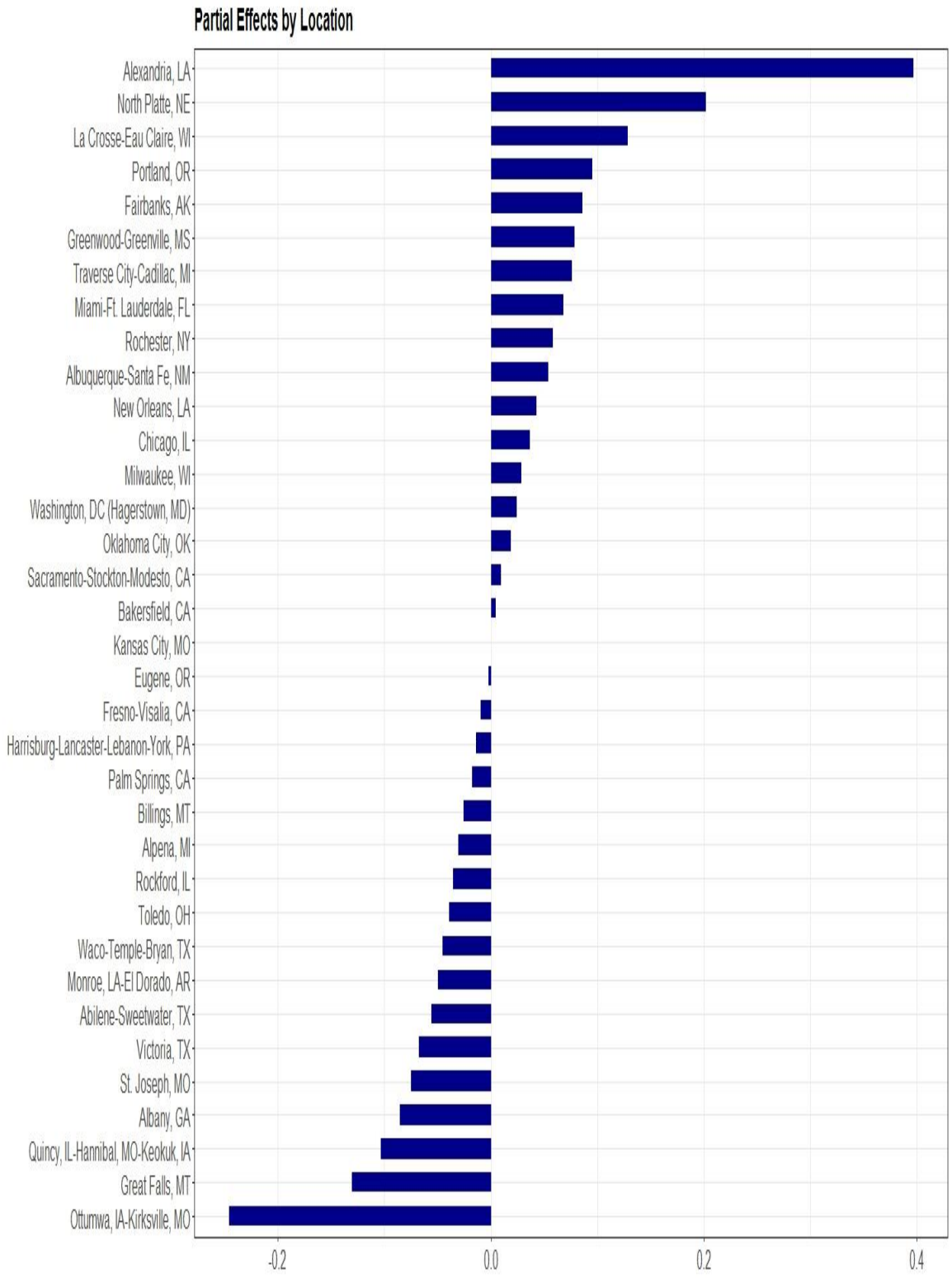


Figure 4-4: Partial Effects by Market Areas. We only display 30 levels from this Variable. Values are centered around the median.

Partial effects for two of the most important and actionable variables is shown in Figures 4-4, 4-5. As one would expect, donor race has a significant effect in determining the contribution to the diversity of the registry. The information used in the model is self-identified Race and Ethnicity. From the Figure 4-5 we note that African Americans are 50% more likely to contribute to registry diversity. Hence recruiting in Black American communities can positively affect registry diversity. In Figure 4-5 we show only 30 of the 206 different market areas for presentation purposes. This helps in identifying locations where recruitment can be targeted to diversify the registry. For example, recruits from Alexandria, Louisiana are 55% more likely to have a rare genotype than recruits from the region between Ottumwa, Iowa and Kirksville, Missouri.

Chapter 5 Conclusion and Future Work

The methods presented in this thesis can help in streamlining several aspects of stem-cell transplants and management of volunteer donor registries. We have provided data-analytic solutions to specific problems that are currently encountered.

In donor selection modelling, we developed a cost-sensitive based model that helps in identifying donors with the most favorable secondary characteristics. The task of finding 3-5 donors among thousands of identically matched donors can be very challenging to even experienced search experts. A significant amount of time is spent in sifting through matched donors and is done while patients are under critical care. Incorporating the proposed model in the existing donor display interface will aid in making this decision faster and more consistent. We have shown that 75% of donor searches (in the out-of-sample data) had all their chosen donors listed within position 45 in the proposed sorting method. This will be of particular help to Transplant Centers (TCs) that lack expertise in donor selection. Additionally, this will also help in monitoring the quality of donor selections and allow for intervention where necessary.

However, we do make certain assumptions during modelling. The most important assumption made in modelling is that we assume the data is generated from a single source, that is, we assume decision made by different physicians all follow similar rules. This may not be true. Currently, we do not have enough historical donor search data for each TC to model them separately. Once enough data is archived, a more sophisticated formalization such as Multi-Task Learning can be used, with each Transplant Center represented as a separate task. TC clustering can be performed based on selection preferences. For example, certain TCs might have specific match probability cutoffs which need to be explored in finer detail once sufficient historical data is archived.

In the next modeling effort, we have proposed a model to predict donor availability. The proposed model can predict a positive CT request response based on donor

demographics and outreach data for each donor. This is a marked improvement over the current system of extrapolating group averages, which tend to be highly erroneous. Knowing possible response to donation request will help in reducing repeated donor searches. However, the current model only accounts for donor request at CT request. Similar modeling strategy can be extended to model donor responses at Work-up stage. The model also needs to be updated when data distribution changes. For example, when the study was performed, only 10% of the registered members had a Recommit request. Currently, all donors who have an email listed in the database have been sent a Recommit request. When such changes in data distribution are observed, models need to be re-trained.

In Chapter 4, we provide a way to combine the two previous modeling efforts with donor HLA type to identify donors for future registry wide improvements. The proposed framework will also help in quantifying the utility of new donors and registries being added to the current system. We also provide recommendations for improving HLA diversity by intelligent recruitment. Following proposed methods for targeted recruitment will change the distribution of HLA types gradually, which alters the *rare* and *common* genotypes in the registry. When a shift in distribution is observed, models need to be readjusted with the new *rare* and *common* HLA definitions to make appropriate suggestions.

Publications from this Thesis

- A.Sivasankaran, E. Williams, G.E. Switzer, V. Cherkassky, M. Maiers, “Machine Learning approach to Predicting Stem-Cell Donor Availability”, DOI: https://doi.org/10.1101/242719 [preprint]

- A.Sivasankaran, E. Williams, V. Cherkassky, M. Maiers, “Unrelated Donor Selection for Stem Cell Transplants using Predictive Modelling”, DOI: https://doi.org/10.1101/242735 [preprint]

- A. Sivasankaran**,** M. Albrecht, E. Williams, M. Maiers, V. Cherkassky, “Donor Selection for Hematopoietic Stem Cell Transplant using Cost Sensitive SVM”, **ICMLA**, 2015, DOI:https://doi.org/10.1109/ICMLA.2015.166

- A. Sivasankaran, E. Williams, M. Albrecht, "Key Driver Analysis of HLA Diversity: Analytically focused recruitment strategies for improving registry quality", **WMDA Council meeting,** 2014

- Analyzing Utility of Donors in Adult Volunteer Stem-Cell Registry [In Preparation]

Chapter 6 References

Besse, K. *et al.* (2016) 'On Modeling Human Leukocyte Antigen–Identical Sibling Match Probability for Allogeneic Hematopoietic Cell Transplantation: Estimating the Need for an Unrelated Donor Source', *Biology of Blood and Marrow Transplantation*. Elsevier, 22(3), pp. 410–417. doi: 10.1016/j.bbmt.2015.09.012.

Besse, K. L. *et al.* (2015) 'Estimating demand and unmet need for allogeneic hematopoietic cell transplantation in the United States using geographic information systems.', *Journal of oncology practice*. American Society of Clinical Oncology, 11(2), pp. e120-30. doi: 10.1200/JOP.2014.000794.

Bochtler, W. *et al.* (2016) 'A comparative reference study for the validation of HLA-matching algorithms in the search for allogeneic hematopoietic stem cell donors and cord blood units', *HLA*, 87(6), pp. 439–448. doi: 10.1111/tan.12817.

Boeckh, M. and Nichols, W. G. (2004) 'The impact of cytomegalovirus serostatus of donor and recipient before hematopoietic stem cell transplantation in the era of antiviral prophylaxis and preemptive therapy', *Blood*, 103(6), pp. 2003–2008. doi: 10.1182/blood-2003-10-3616.

Breiman, L. (1997) 'Arcing The Edge', *Statistics*, 4, pp. 1–14. doi: 10.1.1.367.9480.

Buck, K. *et al.* (2016) 'High-Resolution Match Rate of 7/8 and 9/10 or Better for the Be The Match Unrelated Donor Registry', *Biology of Blood and Marrow Transplantation*, 22(4), pp. 759–763. doi: 10.1016/j.bbmt.2015.12.012.

Caruana, R. and Niculescu-Mizil, A. (2006) 'An empirical comparison of supervised learning algorithms', *Proceedings of the 23rd international conference on Machine*

learning, C(1), pp. 161–168. doi: 10.1145/1143844.1143865.

Chang, C. and Lin, C. (2013) 'LIBSVM : A Library for Support Vector Machines', *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2, pp. 1–39. doi: 10.1145/1961189.1961199.

Chapelle, O. *et al.* (2002) 'Choosing multiple parameters for support vector machines', *Machine Learning*, 46(1–3), pp. 131–159. doi: 10.1023/A:1012450327387.

Chawla, N. V. *et al.* (2002) 'SMOTE: Synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, 16, pp. 321–357. doi: 10.1613/jair.953.

Cherkassky, V. (2013) *Predictive Learning*. VCtextbook.

Cherkassky, V. and Dhar, S. (2010) *Simple Method for Interpretation of High-Dimensional Nonlinear SVM Classification Models.*

Cherkassky, V. and Dhar, S. (2015) 'Interpretation of black-box predictive models', in *Measures of Complexity: Festschrift for Alexey Chervonenkis*, pp. 267–286. doi: 10.1007/978-3-319-21852-6_19.

Cherkassky, V. and Mulier, F. (2007) *Learning from Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi: 10.1002/9780470140529.

Copelan, E. A. (2006) 'Hematopoietic Stem-Cell Transplantation', *New England Journal of Medicine*, 354(17), pp. 1813–1826. doi: 10.1056/NEJMra052638.

Cunha, R. *et al.* (2014) 'Impact of HLA mismatch direction on outcomes after umbilical cord blood transplantation for hematological malignant disorders: a retrospective Eurocord-EBMT analysis', *Bone Marrow Transplantation*. Nature Publishing Group,

49(1), pp. 24–29. doi: 10.1038/bmt.2013.120.

Dehn, J. *et al.* (2016) 'HapLogic: A Predictive Human Leukocyte Antigen–Matching Algorithm to Enhance Rapid Identification of the Optimal Unrelated Hematopoietic Stem Cell Sources for Transplantation', *Biology of Blood and Marrow Transplantation*. Elsevier Inc., 22(11), pp. 2038–2046. doi: 10.1016/j.bbmt.2016.07.022.

Elkan, C. (2001) 'The foundations of cost-sensitive learning', *IJCAI International Joint Conference on Artificial Intelligence*, pp. 973–978. doi: doi=10.1.1.29.514.

Fan, R.-E. *et al.* (2008) 'LIBLINEAR: A Library for Large Linear Classification', *Journal of Machine Learning Research*, 9, pp. 1871–1874. doi: 10.1038/oby.2011.351.

Fleischhauer, K. *et al.* (2012) 'Effect of T-cell-epitope matching at HLA-DPB1 in recipients of unrelated-donor haemopoietic-cell transplantation: A retrospective study', *The Lancet Oncology*, 13(4), pp. 366–374. doi: 10.1016/S1470-2045(12)70004-9.

Freund, Y. and Schapire, R. E. (1997) 'A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting', *Journal of Computer and System Sciences*, 55(1), pp. 119–139. doi: 10.1006/jcss.1997.1504.

Friedman, J. H. (2001) 'Greedy function approximation: A gradient boosting machine', *Annals of Statistics*, 29(5), pp. 1189–1232. doi: DOI 10.1214/aos/1013203451.

Gragert, L. *et al.* (2013) 'Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry', *Human Immunology*. American Society for Histocompatibility and Immunogenetics, 74(10), pp. 1313–1320. doi: 10.1016/j.humimm.2013.06.025.

Gragert, L. *et al.* (2014) 'HLA Match Likelihoods for Hematopoietic Stem-Cell Grafts in

the U.S. Registry', *New England Journal of Medicine*, 371(4), pp. 339–348. doi: 10.1056/NEJMsa1311707.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc. (Springer Series in Statistics).

Hollenbach, J. A. *et al.* (2015) 'Race, ethnicity and ancestry in unrelated transplant matching for the national marrow donor program: A comparison of multiple forms of self-identification with genetics', *PLoS ONE*, 10(8), pp. 1–15. doi: 10.1371/journal.pone.0135960.

Hurley, C. K. *et al.* (2007) 'Overview of registries, HLA typing and diversity, and search algorithms', *Tissue Antigens*. Blackwell Publishing Ltd, 69(s1), pp. 3–5. doi: 10.1111/j.1399-0039.2006.758_2.x.

Irene, R. *et al.* (2017) 'Donor Selection for Allogenic Haemopoietic Stem Cell Transplantation : Clinical and Ethical Considerations', 2017, pp. 1–33.

Juanjuan, W. *et al.* (2007) 'Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding', in *International Conference on Signal Processing Proceedings, ICSP*. doi: 10.1109/ICOSP.2006.345752.

Kanda, J. *et al.* (2015) 'Impact of HLA Mismatch Direction on the Outcome of Unrelated Bone Marrow Transplantation: A Retrospective Analysis from the Japan Society for Hematopoietic Cell Transplantation', *Biology of Blood and Marrow Transplantation*. Elsevier Inc, 21(2), pp. 305–311. doi: 10.1016/j.bbmt.2014.10.015.

Karnes, J. H. *et al.* (2017) 'Comparison of HLA allelic imputation programs', *PLOS ONE*. Edited by J. Tang. Public Library of Science, 12(2), p. e0172444. doi: 10.1371/journal.pone.0172444.

Kollman, C. *et al.* (2004) 'Assessment of optimal size and composition of the U.S. National Registry of hematopoietic stem cell donors.', *Transplantation*, 78(1), pp. 89–95. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15257044 (Accessed: 5 November 2017).

Kollman, C. *et al.* (2007) 'Estimation of HLA-A, -B, -DRB1 Haplotype Frequencies Using Mixed Resolution Data from a National Registry with Selective Retyping of Volunteers', *Human Immunology*, 68(12), pp. 950–958. doi: 10.1016/j.humimm.2007.10.009.

Kollman, C. *et al.* (2016) 'The effect of donor characteristics on survival after unrelated donor transplantation for hematologic malignancy', *Blood*, 127(2), pp. 260–267. doi: 10.1182/blood-2015-08-663823.

Lee, S. J. *et al.* (2007) 'High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation', *Blood*, 110(13), pp. 4576–4583. doi: 10.1182/blood-2007-06-097386.

Madbouly, A. *et al.* (2014) 'Validation of statistical imputation of allele-level multilocus phased genotypes from ambiguous HLA assignments', *Tissue Antigens*. Blackwell Publishing Ltd, 84(3), pp. 285–292. doi: 10.1111/tan.12390.

Petersdorf, E. W. *et al.* (2007) 'Clinical significance of donor–recipient HLA matching on survival after myeloablative hematopoietic cell transplantation from unrelated donors', *Tissue Antigens*, 1(SUPPL. 1), pp. 17–24. doi: 10.1111/j.1399-0039.2006.759.

Petersdorf, E. W. (2015) 'HLA mismatching in transplantation', *Blood*, 125(7), pp. 1058–1059. doi: 10.1182/blood-2014-12-619015.

Platt, J. (1999) 'Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods', *Advances in large margin classifiers*, 10(3), pp. 61–74.

doi: 10.1.1.41.1639.

Ridgeway, G. (2006) ‘Generalized boosted regression models’, *Documentation on the R Package ‘gbm’, version*, 1, p. 7.

Shaw, B. E. *et al.* (2013) ‘Translating the HLA-DPB1 T-cell epitope-matching algorithm into clinical practice’, *Bone Marrow Transplantation*. Nature Publishing Group, 48(12), pp. 1510–1512. doi: 10.1038/bmt.2013.91.

Shaw, B. E. *et al.* (2014) ‘HLA-DPB1 matching status has significant implications for recipients of unrelated donor stem cell transplants HLA-DPB1 matching status has significant implications for recipients of unrelated donor stem cell transplants’, 107(3), pp. 1220–1226. doi: 10.1182/blood-2005-08-3121.

Shaw, B. E. *et al.* (2015) ‘Analysis of the Effect of Race, Socioeconomic Status, and Center Size on Unrelated National Marrow Donor Program Donor Outcomes: Donor Toxicities Are More Common at Low-Volume Bone Marrow Collection Centers’, *Biology of Blood and Marrow Transplantation*. Elsevier, 21(10), pp. 1830–1838. doi: 10.1016/j.bbmt.2015.06.013.

Shouval, R. *et al.* (2014) ‘Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in SCT’, *Bone Marrow Transplantation*. Nature Publishing Group, 49(3), pp. 332–337. doi: 10.1038/bmt.2013.146.

Spellman, S. R. *et al.* (2012) ‘A perspective on the selection of unrelated donors and cord blood units for transplantation’, *Blood*, 120(2), pp. 259–265. doi: 10.1182/blood-2012-03-379032.

Spellman, S. R. *et al.* (2012) ‘A perspective on the selection of unrelated donors and cord blood units for transplantation A perspective on the selection of unrelated donors and cord

blood units for transplantation', 120(2), pp. 259–265. doi: 10.1182/blood-2012-03-379032.

Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc.

Weiss, G., McCarthy, K. and Zabar, B. (2007) 'Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?', *Dmin*, pp. 1–7. Available at: http://storm.cis.fordham.edu/~gweiss/papers/dmin07-weiss.pdf.

Worel, N. (2016) 'ABO-Mismatched Allogeneic Hematopoietic Stem Cell Transplantation', *Transfusion Medicine and Hemotherapy*, 43(1), pp. 3–12. doi: 10.1159/000441507.

Zadrozny, B., Langford, J. and Abe, N. (2003) 'Cost-sensitive learning by cost-proportionate example weighting', *Third IEEE International Conference on Data Mining*, pp. 435–442. doi: 10.1109/ICDM.2003.1250950.