Title: Machine Learning approach to Predicting Stem-Cell Donor Availability

Authors: Adarsh Sivasankaran[1,2], Eric Williams[2], Galen E. Switzer[3,4], Vladimir Cherkassky[1,5], Martin Maiers[2]

[1] Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, MN 55455
[2] Center for International Blood and Marrow Transplant Research (CIBMTR), Minneapolis, MN 55401
[3] Departments of Medicine, Psychiatry and Clinical and Translational Science, University of Pittsburgh, Pittsburgh, PA 15213
[4] Center for Health Equity Research and Promotion, Veterans Affairs Pittsburgh Healthcare System, Pittsburgh, PA 15213
[5] Department of Electrical and Computer Engineering, University of Minnesota, MN 55455

Keywords: Stem Cell Transplant, Donor Availability, Donor Selection, Machine Learning

0. Abstract:

The success of Unrelated Donor stem-cell transplants depends not only on finding genetically matched donors but also on donor availability. On average 50% of potential donors are unavailable for a variety of reasons after initially matching a patient with significant variations in availability among subgroups (e.g., by race or age). Several studies have established univariate donor characteristics associated with availability. Individual consideration of each applicable characteristic is laborious. Extrapolating group averages to individual donor level tends to be highly inaccurate. In the current environment with enhanced donor data collection, we can make better estimates of individual donor availability. In this study, we propose a Machine Learning based approach to predict availability of every registered donor, to be used during donor selection and reduce the time taken to complete a transplant.

1. Introduction

Stem cell transplants are a curative therapy for a number of malignant and non-malignant hematological disorders. An ideal candidate donor for a stem cell transplant is an HLA (Human Leukocyte Antigen) matched sibling (Ballen *et al.*, 2008). Unfortunately, only 30% of patients can find HLA compatible donors within their families (Besse *et al.*, 2016). The remaining 70% rely on either Unrelated Donors (URDs), Cord Blood Units (CBUs) or Haploidentical Donors for a successful transplant. URD and CBU searches are performed with the assistance of public donor registries. Be The Match registry, maintained by the National Marrow Donor Program (NMDP) is one of the world's largest registries with over 10.5 million adult donors[1], and its cord-blood inventory has over 150,000 units.

[1] While a registered member only becomes a *donor* when a stem cell donation is made, we use 'donor' to identify members on the registry.

HLA compatibility for URD searches is determined by an HLA matching algorithm, such as HapLogic™ (Dehn *et al.*, 2016) used by the NMDP. HLA compatible donors are then displayed to physicians (and/or clinicians) through a donor display system (Traxis℠) along with relevant secondary donor information such as age, gender, weight, and so on for donor selection. Selection of the best donor among matched alternatives for transplant is based on optimal secondary donor characteristics that provide the best chance of survival for patients. Donor availability remains a key consideration during the donor selection process (Krishnamurti *et al.*, 2003). However, until recently, there have been few metrics to gauge donor commitment and availability. For the purposes of this study, we define donor availability as an adult registry member who responds positively to a Confirmatory Typing (CT) request and successfully completes a blood sample collection.

Several characteristics are known to be associated with donor availability. For example, race and ethnicity are strongly associated characteristics of donor availability. There are also established associations with donor age, gender, and other characteristics. Average availability rates are often extrapolated to represent the availability of each member of a given subgroup. Donor characteristics positively associated with availability are often not aligned medical desirability for selection. For example , stem cells from younger male donors are known to provide the best chance of survival for patients (Kollman *et al.*, 2016) even though older female donors are known to be generally more available. These univariate associations and contradictory relationships make the task of donor selection a very difficult and complex process. Choosing a donor who will ultimately decline the CT request causes significant delays and complications to completing a transplant, for patients at a time of critical risk for survival. Until recently, clinicians have had to rely on knowledge of historical average availability rates among donor subgroups to determine the number of potentially matched donors to request for CT to ensure an available donor at transplant. A computational model that predicts donor availability during the donor search process could potentially simplify and improve transplant physicians' decision process and minimize the risk of delays in transplantation.

In this study, we show how a machine learning approach can be leveraged to predict donor availability based on demographics and non-genetic donor related factors. Machine learning is being successfully used in a multitude of applications today. It involves *learning* a model based on historical data to make predictions on future data based on similar patterns. The particular branch of machine learning we use in this study is called *supervised learning*, where a model is trained using *labeled* data.

2. Data and Methods

Analysis of available data over time has helped experts identify subgroups with higher (or lower) than average availability rates. However, we do not currently have a computational model that unifies all donor information (associated with availability) to predict availability. Apart from the

demographic data collected at the time a donor joins the registry, the NMDP captures several specific member responses and actions associated with donor availability. These include:

a. Response to questions on a Post Recruitment Survey that has been specifically developed through research to measure donor 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
e. 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 is a strong indicator 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 e.g., members of self-identified Caucasian race, or members who have renewed commitment) and extend 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 evaluated CT request data from the period August 1, 2013 to November 30, 2015. The responses from all outreach programs had been collected during this period.A total of 178,249 CT requests were made during this period. Associated donor data (demographics and response to outreach programs) were collected for modelling. 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 research studies a finer classification maybe used. The following race groups are retained as levels in a categorical variable for modelling.

   AFA – African American
   API – Asian and Pacific Islander
   CAU – Caucasian
   DEC – Declined to Answer
   HIS – Hispanic
   MLT – Multi-Race groups
   NAM – Native American
   OTH – Others
   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 modelling.
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 variable is generated from the available data for modellings.
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 (DIY):* Primary method of donor recruitment is via live drives where a representative (or a volunteer) from Be The Match 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 is believed to indicate 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. Binary indicator was used for modelling.
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 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 modelling. A total of 6 features are used to represent donors PRS response.
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 modelling, as different levels in a categorical variable.
11. *CT Request Response:* Donor's request (yes/no) for a CT has been identified. Donors can be *decline* a CT request for a number of reasons – temporarily unavailable, unable to contact, medically unavailable, or not interested. We ignore any temporarily unavailable members in the study. This response is used as the *target* variable.

Table 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**	
	No	*Yes*	*No*	*Yes*	*Live 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

Each of these identified donor demographics and responses are associated with varying levels of availability. Figure 1 shows observed availability rates by race for the analysis period covered. Average availability rates for all of the above listed factors are tabulated in Appendix A. Most donor features are displayed on the search report display system (Traxis^{SM}) so that transplant centers can select the best donor among HLA matched alternatives. This display might indicate some measure of bias in donor selection, but the selection process does not directly impact the donor availability. We also include certain features in the model that are not displayed but are known to affect availability.

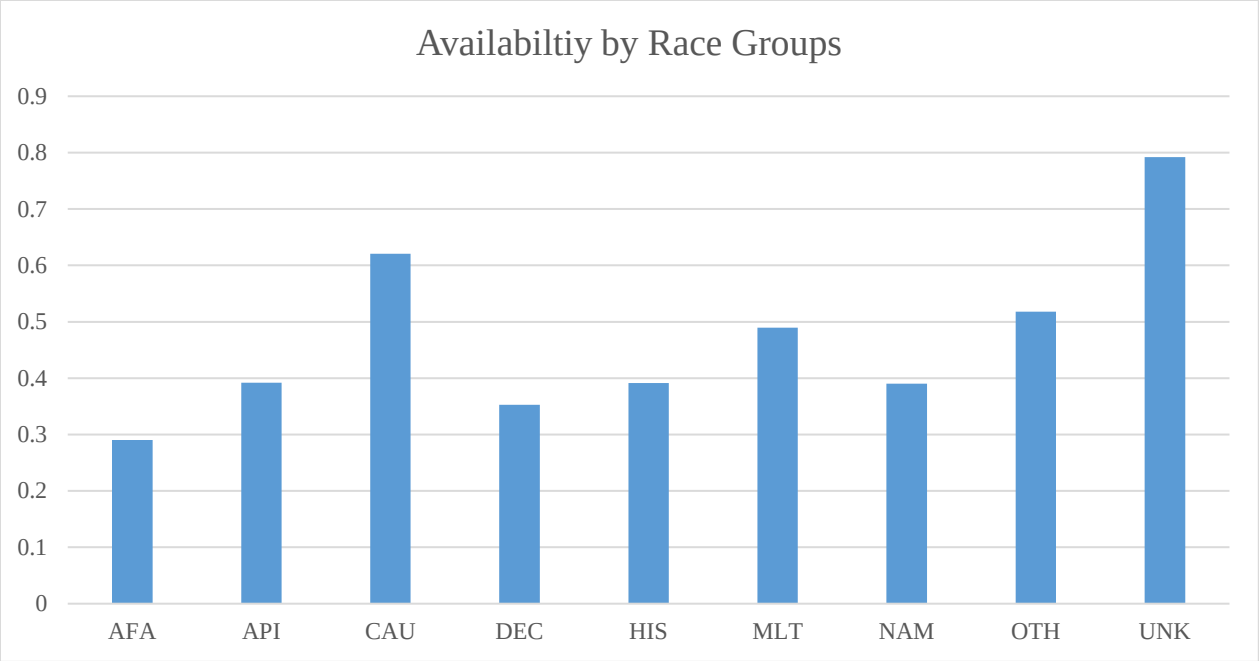

Figure 1: Availability by self-identified broad race groups. Substantial variation is apparent in the average availability rates between groups. Caucasians have an availability rate of 62% while African American donors have an availability rate of 29%.

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). Table 1 shows variations in responses to outreach programs by race groups and gender. We notice donor responses to outreach programs are not uniformly distributed across either gender or race groups. 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 number of considered variables. For example, there are 468 possible interactions between Race, DC Code, DIY and HHQ response variables. Sub-setting data into such finer detail might show spurious results. We have thus avoided presenting marginal availability values at this level. We note in Appendix A 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. Machine learning methods have mechanisms to account for these interactions without users having to specifically define these interactions between variables.

Problem Formalization

An important step in building machine learning models is *Problem Formalization*. Here, based on the available data, the given modeling task is mathematically represented as a learning problem. Since the goal of this effort is to train a machine learning model to predict donor availability, we choose a *binary classification* formalization Binary classification is known as the task of classifying the given data into two groups. In our current problem, the two groups are: 1. *Donors*

who said Yes to a CT request, 2. *Donors who said No to a CT request*, referred to as the *Available* and *Not Available* class respectively. Based on the observed donors' response to CT requests, we assign either of the two classes to each donor in the dataset. As noted earlier, donors can decline a CT request for a number of reasons. To develop a patient-focused model and to avoid modeling sub-categories of negative responses, all declined requests are combined into the same class – *Not Available*.

In the collected dataset, the overall availability rate is 56%, i.e., 56% of the CT requests made between 8/1/2013 – 11/30/2015 were answered positively. In all, we have 15 variables to represent each member, which includes several categorical variables with multiple levels. Several popular Machine Learning algorithms were used for modeling - Gradient Boosted Trees (GBM) (Freund and Schapire, 1997), Support Vector Machines (SVM) (Cortes and Vapnik, 1995) and Logistic Regression.

3. Results and Discussion

The effectiveness of the learning techniques is measured by classification accuracy on an independent test dataset (not used for training the model). The entire dataset is randomly split into 2 parts: training set and testing set. Training set is used for selecting optimal model parameters using cross validation and the test set is used only to measure performance (accuracy) of the optimal model. Classification Accuracy is evaluated by the percentage of data samples classified correctly by the algorithm. Modelling accuracy results measured on the training and test data are tabulated in Table 2 for the algorithms used for modeling. Details of the training process and model selection criteria is described in Appendix B.

Table 2: Training and Testing Accuracy of models measured on the availability dataset

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

We notice Boosting Trees have the best performance. Gradient Boosting, which falls under a general class of methods called Boosting methods (Breiman, 1997; Freund and Schapire, 1997; Schapire *et al.*, 1998; Friedman, 2001) was used. Boosting is a training (model learning) strategy where multiple classifiers (Decision Trees in this case) learn on different versions of the training data and eventually combined to form a single decision rule that has better accuracy than the individual components. Different versions of Boosting techniques have been successfully used for several applications. In biomedical studies Boosting has been used for prediction of effectiveness of drug combinations (Xu *et al.*, 2017), detection of mitosis in breast cancers (Veta *et al.*, 2015), and classification of lung bronchovascular anatomy using CT imaging (Ochs *et al.*, 2007).

Boosting methods typically use Decision Trees as their base model. This is practically beneficial since Trees can handle mixed data (categorical and numeric) and missing values

Binary classifiers are commonly used to assign hard labels (*available* or *not available*) 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. From Table 2, we notice it has better classification accuracies than the other two. Hence, we use this method for further analysis.

Modelling is done in R using the *gbm* package (Ridgeway, 2006) for Gradient Boosting and the *e1071* (Chang and Lin, 2013) 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. A detailed theory about this calculation is provided in (Hastie, Tibshirani and Friedman, 2001). Figure 2 shows the relative influence of variables in predicting donor availability in the trained model. Only 9 of the most important features are shown in the figure.

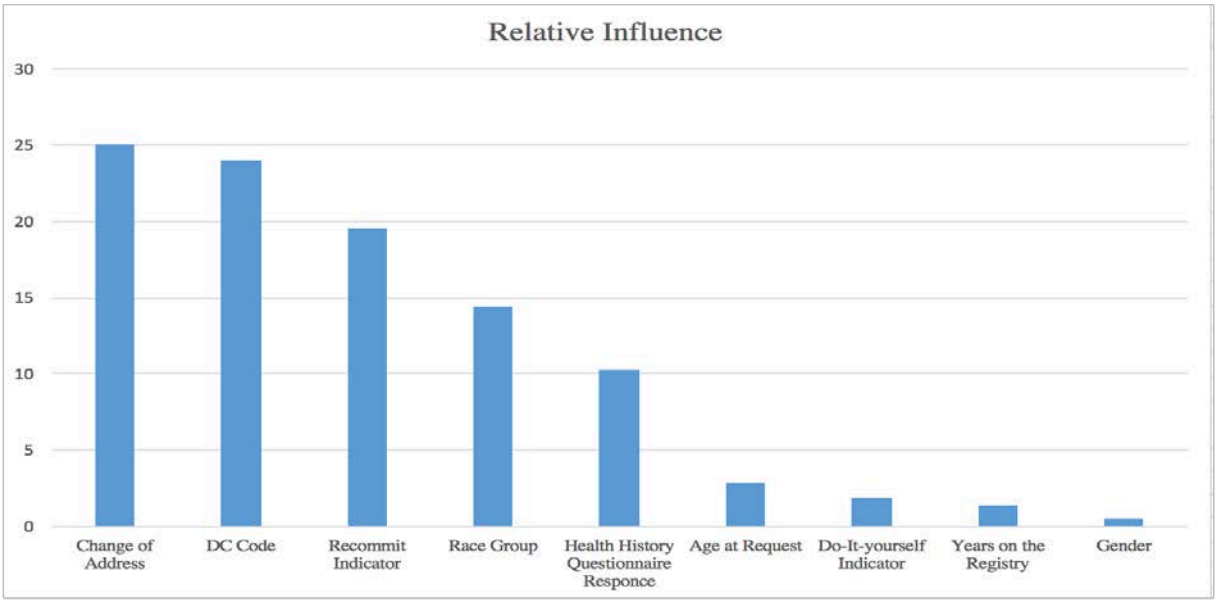

Figure 2: Variable Importance plot for the trained model. Other variables listed in Section 2 had lower influence and are not plotted here.

Comparing actual Availability to model outcomes

Figure 3 shows actual availability rates broken down by model assigned score ranges. The gradient boosting model assigns a number between 0 and 1 to every donor. We have binned donors based on these assigned scores and measured the actual availability rate within each bin. For example, availability among donors who had a score between 0.9 and 1.0 was 93%. Availability rates in each score range are marked by the rhombus shaped points in the figure. This (almost) linear relation between average availability and model assigned scores can be observed across all

brackets, as shown in Figure 3. This indicates the model assigned scores linearly correspond to the observed availability. Consequently, model assigned scores can be used as a direct indicator of a donor's availability. The bar for each score range represents the number of donors in that range. Average availability and number of donors in each score range are represented on different scales on the left and right side of the figure respectively. The linear trend between model scores and actual availability suggests that the model assigned scores are very strong indicators of availability.

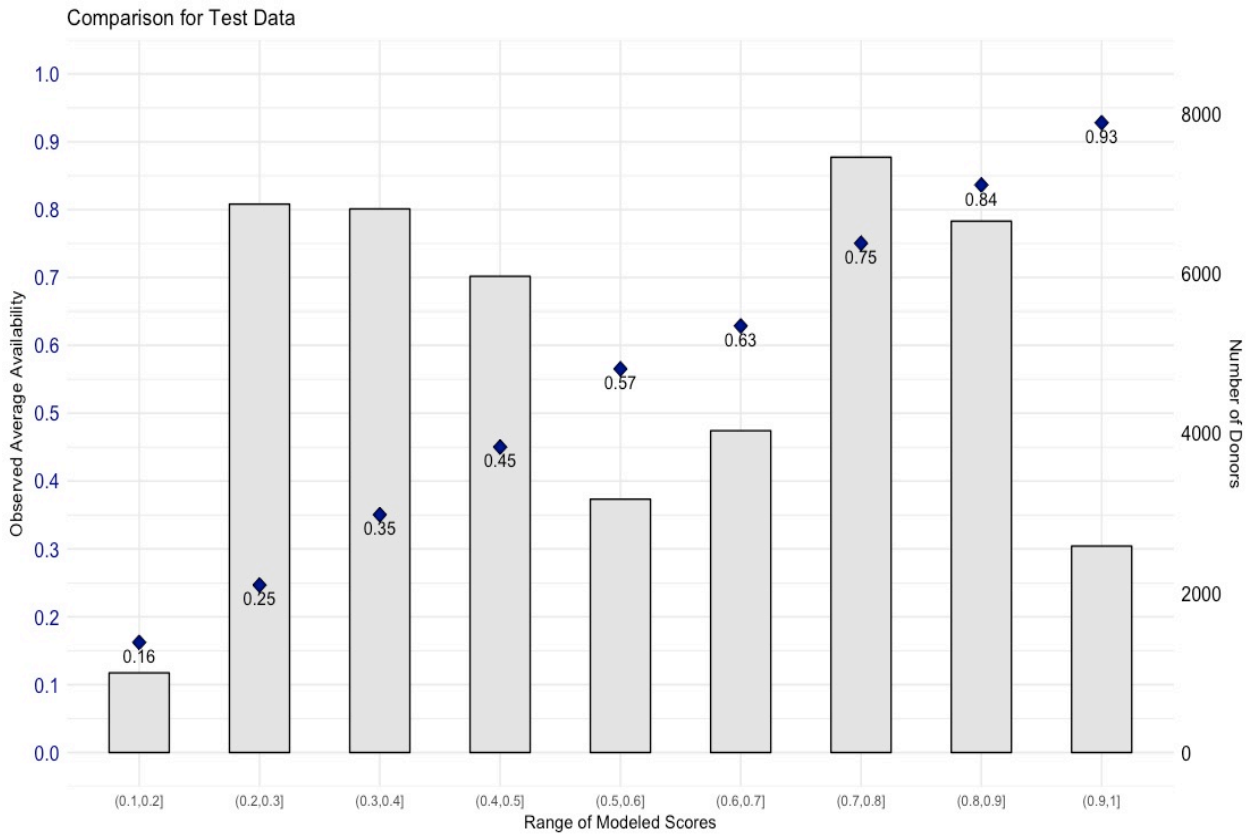

Figure 3: Observed availability rates compared to model assigned scores. The overlaid number in each column represents the average observed availability of donors with modelled scores in the corresponding range. The bar graphs are the number of donors who are in the brackets noted on x-axis.

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 of information, instead of having missing information in the modelling set. The model is capable of adjusting for varied levels of information and assigning a score to every member in the registry.

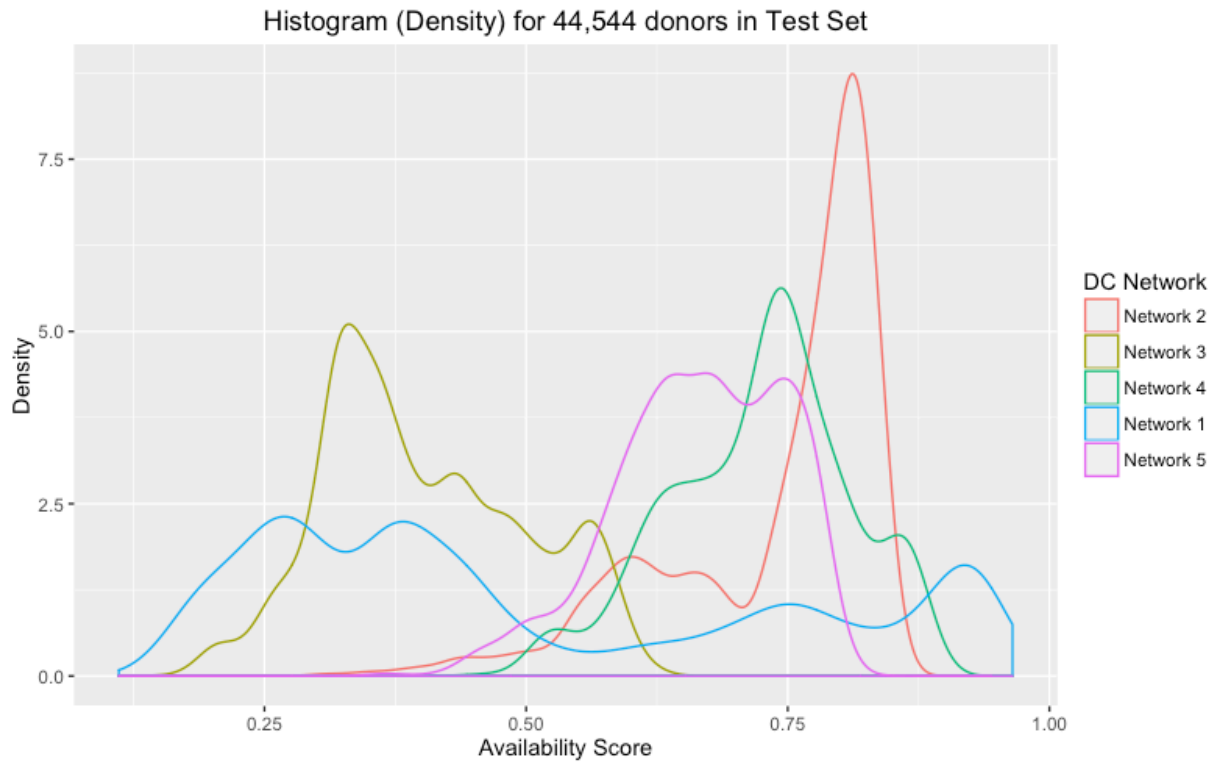

Figure 4: Density of donor availability for different donor networks

Figure 4 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 potentially results in a wider spread in the assigned scores. While higher peaks do reflect higher average availability, lack of information makes it hard to differentiate between donors. Having a wide spread density plot means the model can differentiate well between donors who are *available* and *not available*.

4. Conclusion

Typically, experts have 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 individually based 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 simplify the donor selection process and reduce the time taken to successfully complete the transplant.

Acknowledgement

We thank Stephen Spellman, Jason Dehn, and Michael Wright for their valuable comments and suggestions on an early draft of this manuscript.

References

Ballen, K. K. *et al.* (2008) 'The National Marrow Donor Program 20 Years of Unrelated Donor Hematopoietic Cell Transplantation', *Biology of Blood and Marrow Transplantation*, 14(9 SUPPL.), pp. 2–7. doi: 10.1016/j.bbmt.2008.05.017.

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.

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

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.

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

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

Cortes, C. and Vapnik, V. (1995) 'Support-Vector Networks', *Machine Learning*, 20(3), pp. 273–297. doi: 10.1023/A:1022627411411.

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.

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.

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.* (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.

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.

Krishnamurti, L. *et al.* (2003) ‘Availability of unrelated donors for hematopoietic stem cell transplantation for hemoglobinopathies’, *Bone Marrow Transplantation*, 31(7), pp. 547–550. doi: 10.1038/sj.bmt.1703887.
Ochs, R. A. *et al.* (2007) ‘Automated classification of lung bronchovascular anatomy in CT using AdaBoost’, *Medical Image Analysis*, 11(3), pp. 315–324. doi: 10.1016/j.media.2007.03.004.
Ridgeway, G. (2006) ‘Generalized boosted regression models’, *Documentation on the R Package ‘gbm’, version*, 1, p. 7.
Schapire, R. E. *et al.* (1998) ‘Boosting the margin: A new explanation for the effectiveness of voting methods’, *Annals of Statistics*, 26(5), pp. 1651–1686. doi: 10.1214/AOS/1024691352.
Veta, M. *et al.* (2015) ‘Assessment of algorithms for mitosis detection in breast cancer histopathology images’, *Medical Image Analysis*. Elsevier, 20(1), pp. 237–248. doi: 10.1016/j.media.2014.11.010.
Xu, Q. *et al.* (2017) ‘PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm’, *Journal of Theoretical Biology*. Academic Press, 417, pp. 1–7. doi: 10.1016/j.jtbi.2017.01.019.

Appendix A

Marginal Availability rates for different factors listed in Section 2 are tabulated below. Marginal availability rates are affected by sample size for certain factors. We include all of these in the modelling for their assumed effect on availability.

Table 3: Marginal Availability rate by Gender.

Gender	***Availability Rate***
Male	55.8%
Female	55.9%

Table 4: Marginal Availability by Health History Questionnaire Response

Questionnaire Response	***Availability Rate***
Responded	87.1%
Not Responded	53.8%

Table 5: Marginal Availability Rates by Donor Centers

Donor Center Code	***Availability Rate***
A	50.4%
B	67.6%
C	62.1%
D	39.9%
E	79.6%
F	69.3%
G	72.6%
H	76.5%
I	64.8%
J	58.9%
K	48.8%
L	75.5%
M	68.7%

Table 6: Marginal Availability Rate by Post Recruitment Survey Response

Post Recruitment Survey	***Availability Rate***
Response	75.2%
No Response	55.8%

Table 7: Marginal Availability by Recommit Response

Recommit Request	***Availability Rate***

Response	93.5%
No Response	53.4%

Table 8: Marginal Availability Rates for DIY record

Do-It-yourself Record	***Availability Rate***
Online	64.6%
Others (Donor Drive)	47.6%

Table 9: Marginal Availability Rate for Change of Address requests

Change of Address Request	***Availability Rate***
Self-Initiated	79.8%
No Action	51.2%

Appendix B

We use binary classifiers (since we have two categories: *available* and *not available*) to model the availability. Classifiers estimate a function ***f(x)*** that separates the input space into two regions corresponding to the two classes, where **x** is a feature vector that is used to represent measurements/observations for a data instance. For our problem, **x** is a vector representing donor characteristics listed in Section 2. Each donor is represented by a vector (row in the dataset). The estimated function is then used to assign labels based on the feature vector. The quality of a classifier is measured by a loss function. If a classifier predicts all the given data instances perfectly the loss will be 0. The loss function measures the discrepancy between predictions by the classifier and the actual desired output. A standard loss function used is the 0/1 function shown in Equation 1.

$$L\big(y, f(\boldsymbol{x})\big) = \begin{cases} 0 & if\ y = f(\boldsymbol{x}) \\ 1 & if\ y \neq f(\boldsymbol{x}) \end{cases} \tag{1}$$

y is the actual observed output, ***f(x)*** is the model predicted output for sample **x**.

Classifiers typically have tuning parameters that need to be adjusted by the user according to the data being modelled. This is achieved by cross validation. The entire dataset is typically randomly separated into two groups, one for training the model and the other for testing the accuracy of the model. Depending on the size of the dataset available, the split ratio is determined. For our problem, we use a 75:25 split (training: testing). A k-fold cross validation is then performed on the training set to tune model parameters. A grid of parameters is searched and the best parameter is chosen based on lowest average validation across all folds. We use a 5-fold cross validation to select the best parameter. Depending on the modelling method, users need to tune one or two parameters. After the best parameter is determined, a model is trained on the entire training data split. Classifier accuracy is then measured on the same set and is recorded as the *Training Accuracy/ Error*. But, the quality of prediction is determined by the classifier performance on the test set. Care should be taken to not use any of the samples in the testing set for model selection. Test set is only used for assessing model accuracy, *Testing Accuracy/Error*. Error is measured as the average error across all samples in the dataset as show in Equation 2.

$$Error = \frac{1}{N}\sum_{i=1}^{N} L(y_i,\ f(\boldsymbol{x}_i)) \tag{2}$$

This is the average error for a dataset with *N* samples and the loss function is as defined in Equation 1. While the above-mentioned metric is the most commonly used, care should be taken to use an appropriate metric that addresses specific requirements of the modelling problem. We refer readers to (Cherkassky and Mulier, 2007; Cherkassky, 2013) for a detailed discussion on model selection procedure and further details on modelling techniques.

Data Pre-Processing

Raw data is processed before model training and is an important part of the modelling process. Apart from following standard procedures for pre-processing it is also important to reflect domain knowledge in the pre-processing.

Age at request and Years on registry are scaled on a 0-1 range. The maximum and minimum are recorded to process future samples to the same scale. Categorical predictors (Race and Donor Center Code) with multiple levels are used as *factor* variables in R software. Other predictors, as listed in Appendix A with only two possible values, are encoded as binary variables. Care is taken to identify missing values and miscellaneous responses to outreach requests.