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

Authors: Adarsh Sivasankaran[1,2], Eric Williams[2], Mark Albrecht[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 in the NMDP database 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. The National Marrow Donor Program® (NMDP) provides access to the US Be The Match Registry® as well as several other domestic and international registries. The NMDP database has over 20 million adult donors, and over 450,000 cord blood units.

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.

As a member of a registry only becomes a donor when a stem cell donation is made, and we are interested in predicting the availability of people on the registry, we now use the term member to indicate someone on the registry that we are regarding as a potential donor, and want to investigate their likelihood of being available. Several characteristics are known to be associated with member availability. For example, race and ethnicity are strongly associated characteristics of member availability (G. E. Switzer et al. 2013) . There are also established associations with member age, gender, and other characteristics(G. Switzer et al. 1999; Onitilo et al. 2004). Average availability rates are often extrapolated to represent the availability of each member of a given subgroup. Member characteristics positively associated with availability are often not aligned with medical desirability for selection. For example, stem cells from younger male members are known to provide the best chance of survival for patients (Kollman et al. 2016) even though older female members are known to be generally more available. These univariate associations and contradictory relationships make the task of member selection a very difficult and complex process. Choosing a member 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 member subgroups to determine the number of potentially matched members to request for CT to ensure an available member at transplant. A computational model that predicts member availability during the member 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 member availability based on demographics and non-genetic member related factors. Machine learning involves a *learning model* based on historical data to make predictions on future data based on similar patterns. We use *supervised learning* techniques to train models by providing relevant member information as input and their corresponding responses to CT requests as target outputs. The learning techniques learn a mathematical function based on the historical input-output (member information-CT request response) relationship that can be used to predict responses of future CT requests for every member in the registry.

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 member information (associated with availability) to predict availability. Apart from the

demographic data collected at the time a member joins the registry, the NMDP captures several specific member responses and actions associated with member availability. 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
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 members belonging to a sub-population (for e.g., members who are self-identified European, 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-Americans have historically been much lower than for Europeans. Gragert et al. (2014) reported average availability rate for European registry members to be 51% and for African-Americans it was only 23% when 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 member data (demographics and response to outreach programs) were collected for modeling. A description of the data collected is listed below.

1. *Member Race and Ethnicity:* every member is assigned to one of the following categories based on self-identified race and ethnicity at the time of registration. To aggregate data over the history of recruitment questionnaires we have assigned members to the following broad race/ethnicity categories. The following race groups are retained as levels in a categorical variable for modeling:

   AFA – African-American
   API – Asian and Pacific Islander
   DEC – Declined to Answer
   EUR – European
   HIS – Hispanic
   MLT – Multi-Race groups
   NAM – Native American
   OTH – Others
   UNK – Unknown

2. *Member Age at request:* Calculated from member 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 member registration date and the date the CT request was placed. This information is also used as a numeric variable.

4. *Member Gender:* Collected at the time members join the registry. A binary variable is formed to indicate Female or Male members.
5. *Recommitted:* Registry members are sent communications asking them to renew their commitment to the cause of donation. Whether a member responds to outreach, or not, is recorded in the NMDP database. A response to at least one request is associated with higher availability. A binary response variable is generated for modeling indicating a positive value if they were invited to recommit and responded positively, otherwise negative if they were not invited or they were invited but did not respond.
6. *Address Change:* Identifies member-initiated address change requests. A binary indicator variable is used to identify members who have ever initiated a change of address without solicitation.
7. *Do-It-Yourself (DIY):* Primary method of member 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. We have observed that members that register online have a higher availability rate. A binary variable indicates if a member 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. Members answer four questions about the stem cell donation process that have been found, through research (unpublished), to be most relevant to availability. Members are scored according to their responses. Responses from each of the questions were changed to a positive or negative response. A total composite score was also calculated from answers to the four questions and converted to a positive or negative response. An associated binary variable indicates if a registered member was asked to respond to this survey or not. Another indicator variable was used to identify members who responded to the survey. Each of these variables (2 indicators and 4 response components and an overall response score) were used for modeling. See Appendix A for additional details on the survey.
10. *Donor Center (DC):* The NMDP database contains members recruited through multiple donor center networks. In all there were 13 donor centers used for modeling. These were further grouped into 5 donor center networks. See Table 6 in Appendix B for a list of DC and their network associations.
11. *CT Request Response:* Member responses to CT requests were indicated as either yes or no (declined). The CT request response can be categorized as a "no" request for a number of reasons – unable to contact, medically unavailable, or not interested. 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**		**Post Recruitment Survey**	
Race	*No*	*Yes*	*No*	*Yes*	*Live Drive*	*Online*	*No*	*Yes*	*No*	*Yes*
AFA	14,587 (95.5%)	681 (4.5%)	14,708 (96.3%)	560 (3.7%)	12,270 (80.4%)	2,998 (19.6%)	13,588 (89.0%)	1,680 (11.0%)	15,241 (99.8%)	27 (0.2%)
API	11,366 (95.6%)	526 (4.4%)	11,164 (93.9%)	728 (6.1%)	10,453 (87.9%)	1,439 (12.1%)	9,763 (82.1%)	2,129 (17.9%)	11,871 (99.8%)	21 (0.2%)
EUR	91,426 (92.0%)	7,958 (8.0%)	91,709 (92.3%)	7,675 (7.7%)	43,509 (43.8%)	55,875 (56.2%)	82,245 (82.8%)	17,139 (17.2%)	98,880 (99.5%)	504 (0.5%)
DEC	802 (99.3%)	6 (0.7%)	804 (99.5%)	4 (0.5%)	13 (1.6%)	795 (98.4%)	795 (98.4%)	13 (1.6%)	808 (100.0%)	21 (0.0%)
HIS	18,249 (94.2%)	1,128 (5.8%)	18,296 (94.4%)	1,081 (5.6%)	16,675 (86.1%)	2,702 (13.9%)	16,751 (86.4%)	2,626 (13.6%)	19,270 (99.8%)	107 (0.6%)
MLT	10,512 (93.2%)	771 (6.8%)	10,454 (92.7%)	829 (7.3%)	7,339 (65.0%)	3,944 (35.0%)	9,384 (83.2%)	1,899 (16.8%)	11,183 (99.1%)	100 (0.9%)
NAM	1,341 (94.8%)	73 (5.2%)	1,346 (95.2%)	68 (4.8%)	1,019 (72.1%)	395 (27.9%)	1,248 (88.3%)	166 (11.7%)	1,412 (99.9%)	2 (0.1%)
OTH	328 (97.6%)	8 (2.4%)	327 (97.3%)	9 (2.7%)	74 (22.0%)	262 (78.0%)	322 (95.8%)	14 (4.2%)	336 (100.0%)	0 (0.0%)
UNK	18,480 (100.0%)	7 (0.0%)	18,480 (100.0%)	7 (0.0%)	34 (0.2%)	18,453 (99.8%)	18,471 (99.9%)	16 (0.1%)	18,487 (100.0%)	0 (0.0%)
Gender	*No*	*Yes*	*No*	*Yes*	*Live Drive*	*Online*	*No*	*Yes*	*No*	*Yes*
Female	65,862 (91.4%)	6,205 (8.6%)	65,899 (91.4%)	6,168 (8.6%)	42,965 (59.6%)	29,102 (40.4%)	58,875 (81.7%)	13,192 (18.3%)	71,766 (99.6%)	301 (0.4%)
Male	101,229 (95.3%)	4,953 (4.7%)	101,389 (95.5%)	4,793 (4.5%)	48,421 (45.6%)	57,761 (54.4%)	93,692 (88.2%)	12,490 (11.8%)	105,722 (99.6%)	460 (0.4%)

Each of these identified member 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 B. Most member features are displayed on the search reporting system (Traxis℠) so that transplant centers can select the best member among HLA matched alternatives. This display might indicate some measure of bias in member selection, but the selection process does not directly affect the member availability. We also include certain features in the model that are not displayed but are known to affect availability.

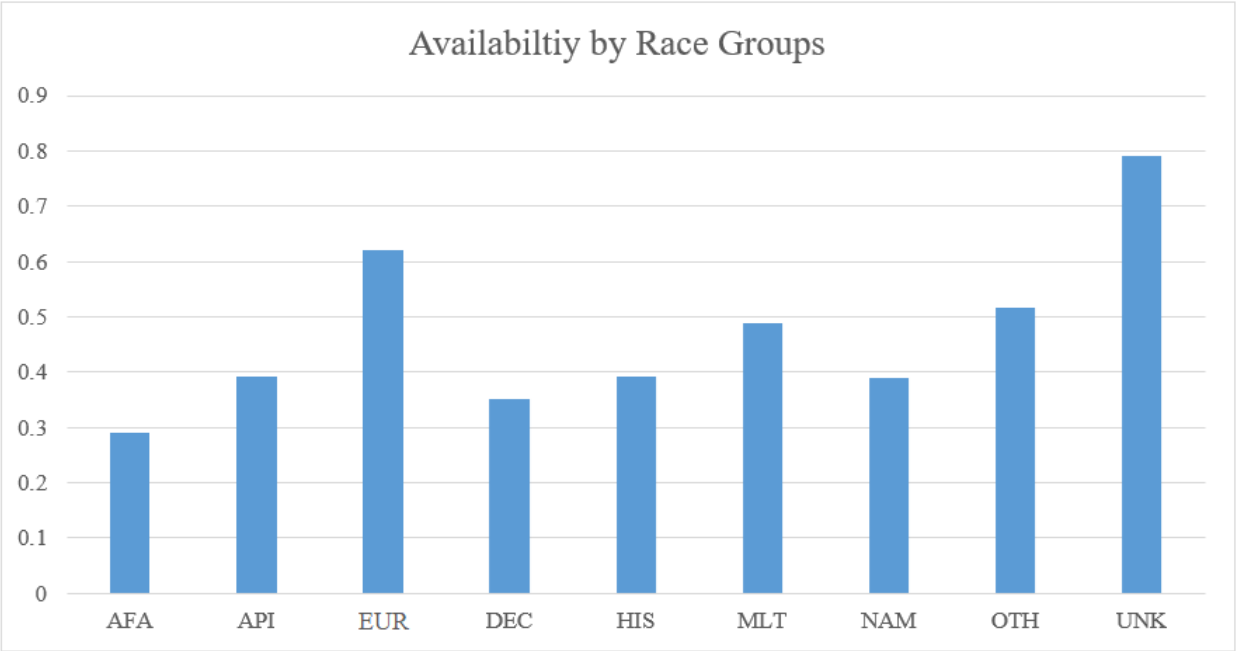

Figure 1: Availability by self-identified broad race groups. Substantial variation is apparent in the average availability rates between groups. Europeans (EUR) have an availability rate of 62% while African-American members have an availability rate of 29%.

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 member 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 *EUR Males* from *DC Code A* who have a *Recommit Response* and subsequently calculating the group averages. The number of such possible sub-groups grows swiftly with the number of variables considered. For example, there are 468 possible interactions between Race, DC Code, DIY and HHQ response variables. Sub-setting data into finer detail might show spurious results. We have thus avoided presenting marginal availability values at this level. We note in Appendix B that different variables have different levels of associated average availability. There is also a problem of missing information for members who are sourced from external donor centers. For example, most international members 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 modeling task is represented as a learning problem. Our goal is to train a machine learning model to predict member availability. We chose to frame the problem as one of *binary classification* into two groups:

1. *Positive response to a CT request (Available)*
2. *Negative response to a CT request (Not Available)*

To develop a patient-focused model we are considering overall availability, not sub-categories of negative responses such as medical deferral or inability to contact.

In the collected dataset, the overall availability rate is 56%, meaning there was a positive response to 56% of the CT requests made between 8/1/2013 and 11/30/2015. In all, we have 16 variables to represent each member, which includes several categorical variables with multiple levels. Several 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. All three techniques are popular learning methods that are commonly used for modeling binary classification formalizations. Logistic Regression works by learning the probability of binary target variable. SVM is a non-probabilistic classification method that learns a decision boundary that separates the two classes (for binary classification). Both SVM and Logistic Regression have been popularly used in multiple biomedical applications including cancer detection and transplant survival analysis studies (Chhatwal et al. 2009; X.-J. Fan et al. 2012; Verplancke et al. 2008; Shiao and Cherkassky 2013; Nematollahi et al. 2017). Gradient Boosting, which falls under a general class of methods called Boosting methods (Breiman 1997; Friedman 2001; Schapire et al. 1998; Freund and Schapire 1997), also learn a decision function. However, unlike the previous two methods, Boosting methods learn multiple classifiers (Decision Trees in this case) on different versions of the training data and eventually combine them to form a single decision rule that has better accuracy than the individual components. Different versions of Boosting techniques have been successfully used in several biomedical studies (Xu et al. 2017; Veta et al. 2015; Ochs et al. 2007). Classification trees (base components in boosting) can handle mixed data (numeric and categorical) and it has practical benefits for our problem.

Modeling was done in R using the *gbm* package (Ridgeway 2006) for Gradient Boosting and the *e1071* (Chang and Lin 2013) and *LiblineaR* (R.-E. Fan et al. 2008) for SVM 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 two parts: training set and testing set. The 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. Modeling 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 are described in Appendix C.

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

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 member 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.

The *gbm* package 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 member availability in the trained model. Only nine of the most important features are shown in the figure.

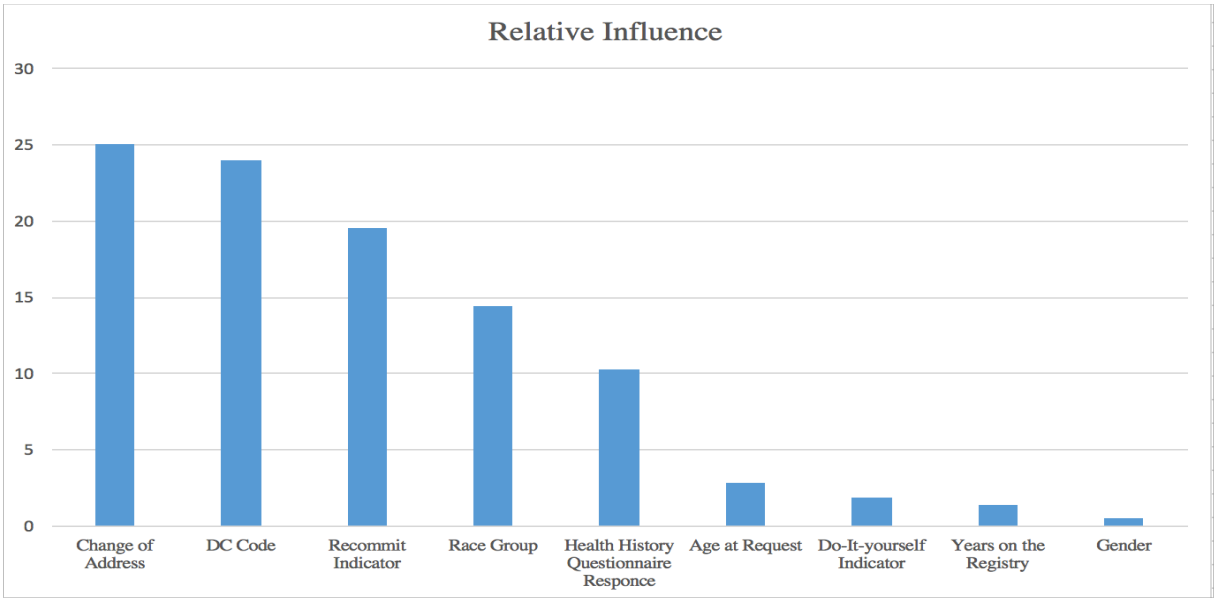

Figure 2: Variable Importance plot for the trained model. Other variables listed in Section 2 had lower influence and are not plotted here.

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 member. We have binned members based on these assigned scores and measured the actual availability rate within each bin. For example, availability among members 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 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 member's availability. The bar for each score range represents the number of members in that range. Average availability and number of members 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.

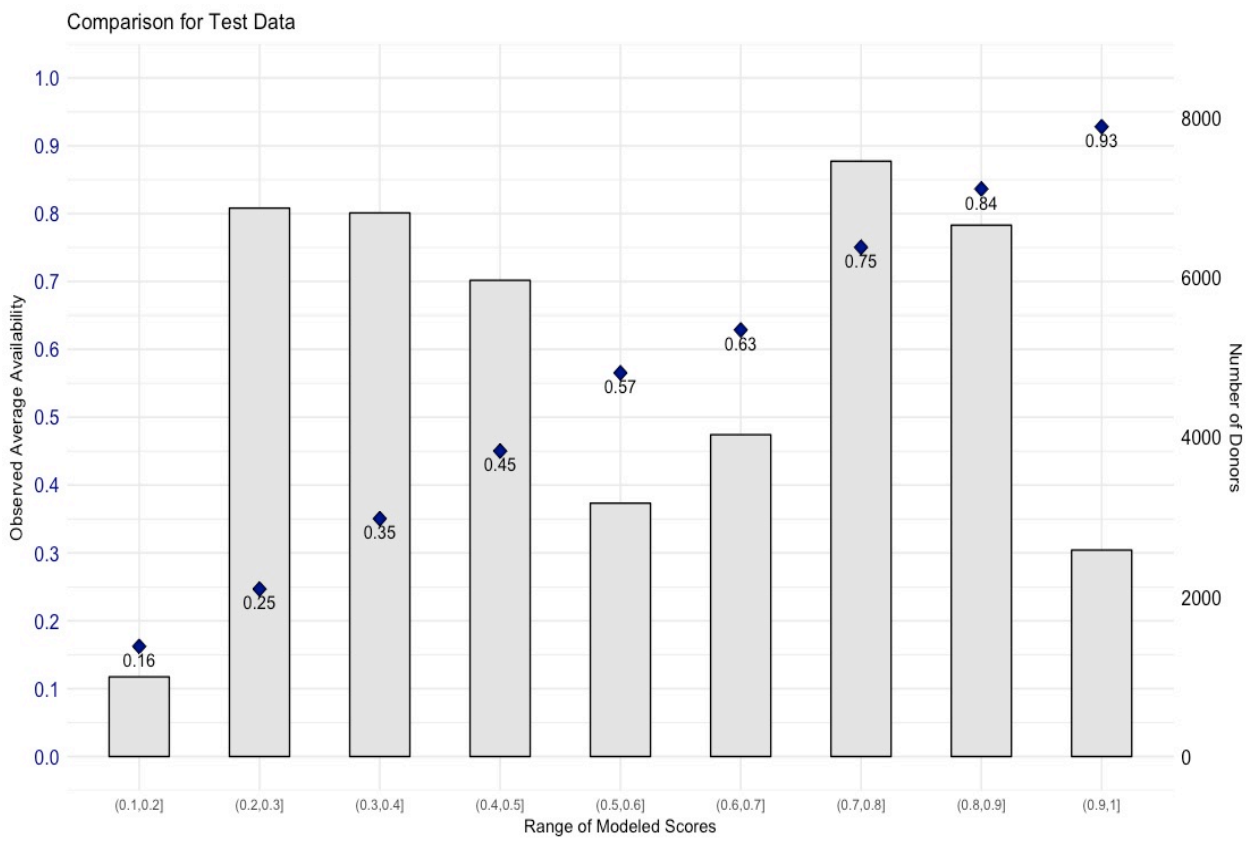

Figure 3: Observed availability rates compared to model assigned scores. The overlaid number in each column represents the average observed availability of members with modeled scores in the corresponding range. The bar graphs are the number of members who are in the brackets noted on the x-axis.

Table 3 Example of two registry members input variables and predicted availability

Model Variable	Member 1	Member 2
Donor Center Category	1	1
Years on Registry	3	3
Age	44	21
Gender	M	F
Race	HIS	AFA
Online Registration	Y	N
Health History Questionnaire	Y	N
Recommitted	Y	N
Address Change	0	0
Post-recruitment Survey Sent	0	0
Post-recruitment Survey Reply	NA	NA
Post-recruitment Survey Score	NA	NA
Predicted Availability	0.95733	0.13387

The NMDP provides access to the US Be The Match Registry as well as several other domestic and international registries. Responses to outreach programs are only recorded for the Be The Match recruits. This results in different levels of information available for modeling availability across donor center networks. The model is capable of adjusting for varied levels of information and assigning a score to every member in the registry.

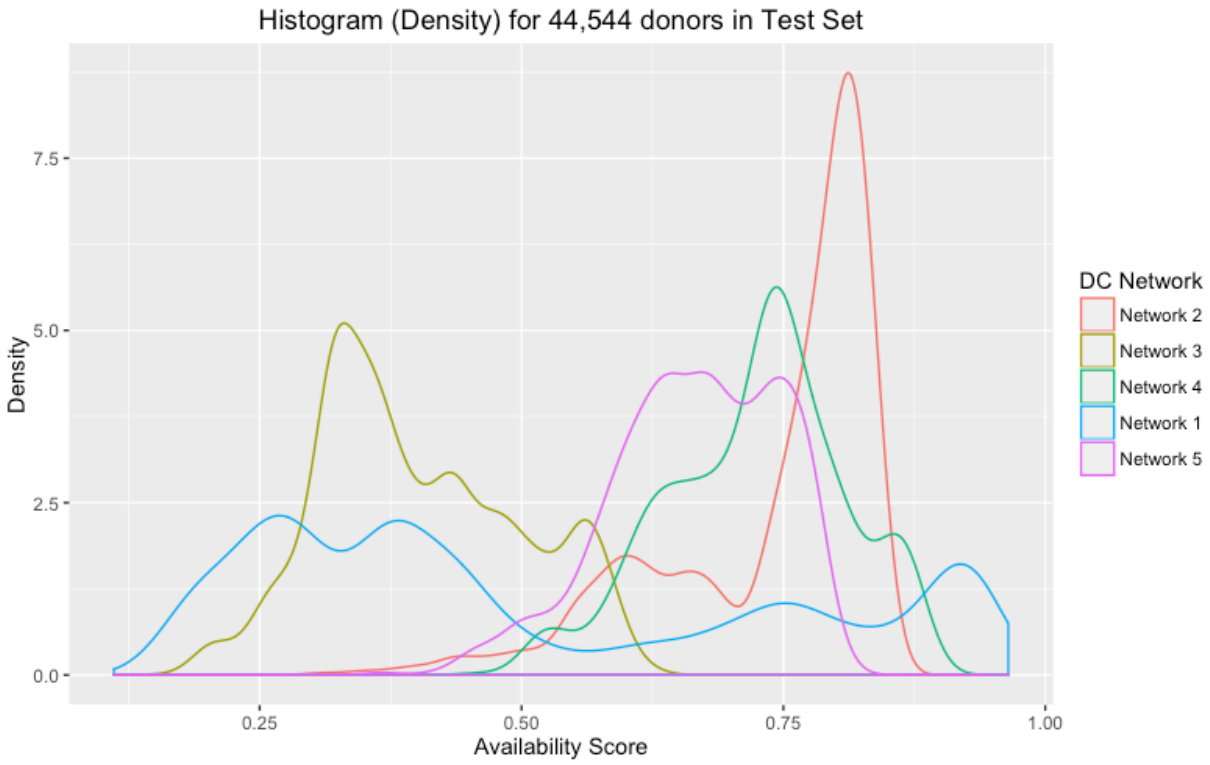

Figure 4: Density of member availability for different donor networks

Figure 4 shows the density of availability scores across each of the five DC networks. Since we have access to more model inputs for the registry members in Be The Match (Network 1) the model is able to provide greater differentiation. For other networks, the lack of inputs results narrower distribution of values.

4. Conclusion

Typically, experts have to rely on historical averages for availability estimates. This results in inaccuracy in estimating an individual member's availability. This also limits the ability of clinicians to accurately estimate how many members should be contacted in order to ensure the patient will be supplied with a bone marrow transplant product when needed. The proposed system helps by providing an individual estimate for each member based on the set of factors available for estimating their availability. Creating a single score for each potential member can significantly simplify the member selection process and reduce the time taken to successfully complete the transplant.

Acknowledgement

The methods used for this analysis were developed through a research grant funded by the US Office of Naval Research (N00014-17-1-2388). We thank Bronwen Shaw, Juliet Barker, Stephen Spellman, Jason Dehn, and Michael Wright for their valuable comments and suggestions on this manuscript.

References

Ballen, Karen K., Roberta J. King, Pintip Chitphakdithai, Charles D. Bolan, Edward Agura, Robert J. Hartzman, and Nancy A. Kernan. 2008. "The National Marrow Donor Program 20 Years of Unrelated Donor Hematopoietic Cell Transplantation." *Biology of Blood and Marrow Transplantation* 14 (9 SUPPL.): 2–7. https://doi.org/10.1016/j.bbmt.2008.05.017.

Besse, Kelsey, Martin Maiers, Dennis Confer, and Mark Albrecht. 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* 22 (3). Elsevier: 410–17. https://doi.org/10.1016/j.bbmt.2015.09.012.

Breiman, Leo. 1997. "Arcing The Edge." *Statistics* 4: 1–14. https://doi.org/10.1.1.367.9480.

Chang, Chih-chung, and Chih-jen Lin. 2013. "LIBSVM : A Library for Support Vector Machines." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2: 1–39. https://doi.org/10.1145/1961189.1961199.

Cherkassky, Vladimir. 2013. *Predictive Learning*. VCtextbook.

Cherkassky, Vladimir, and Filip Mulier. 2007. *Learning from Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470140529.

Chhatwal, Jagpreet, Oguzhan Alagoz, Mary J. Lindstrom, Charles E. Kahn, Katherine A. Shaffer, and Elizabeth S. Burnside. 2009. "A Logistic Regression Model Based on the

National Mammography Database Format to Aid Breast Cancer Diagnosis." *American Journal of Roentgenology* 192 (4): 1117–27. https://doi.org/10.2214/AJR.07.3345.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20 (3): 273–97. https://doi.org/10.1023/A:1022627411411.

Dehn, Jason, Michelle Setterholm, Kelly Buck, Jane Kempenich, Beth Beduhn, Loren Gragert, Abeer Madbouly, Stephanie Fingerson, and Martin Maiers. 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* 22 (11). Elsevier Inc.: 2038–46. https://doi.org/10.1016/j.bbmt.2016.07.022.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. "LIBLINEAR: A Library for Large Linear Classification." *Journal of Machine Learning Research* 9: 1871–74. https://doi.org/10.1038/oby.2011.351.

Fan, X-J, X-B Wan, Y Huang, H-M Cai, X-H Fu, Z-L Yang, D-K Chen, et al. 2012. "Epithelial–mesenchymal Transition Biomarkers and Support Vector Machine Guided Model in Preoperatively Predicting Regional Lymph Node Metastasis for Rectal Cancer." *British Journal of Cancer* 106 (11): 1735–41. https://doi.org/10.1038/bjc.2012.82.

Freund, Yoav, and Robert E Schapire. 1997. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting." *Journal of Computer and System Sciences* 55 (1): 119–39. https://doi.org/10.1006/jcss.1997.1504.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29 (5): 1189–1232. https://doi.org/DOI 10.1214/aos/1013203451.

Gragert, Loren, Mary Eapen, Eric Williams, John Freeman, Stephen Spellman, Robert Baitty, Robert Hartzman, et al. 2014. "HLA Match Likelihoods for Hematopoietic Stem-Cell Grafts in the U.S. Registry." *New England Journal of Medicine* 371 (4): 339–48. https://doi.org/10.1056/NEJMsa1311707.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hollenbach, Jill A., Aliya Saperstein, Mark Albrecht, Cynthia Vierra-Green, Peter Parham, Paul J. Norman, and Martin Maiers. 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): 1–15. https://doi.org/10.1371/journal.pone.0135960.

Kollman, Craig, Stephen R. Spellman, Mei Jie Zhang, Anna Hassebroek, Claudio Anasetti, Joseph H. Antin, Richard E. Champlin, et al. 2016. "The Effect of Donor Characteristics on Survival after Unrelated Donor Transplantation for Hematologic Malignancy." *Blood* 127 (2): 260–67. https://doi.org/10.1182/blood-2015-08-663823.

Krishnamurti, L, S Abel, M Maiers, and S Flesch. 2003. "Availability of Unrelated Donors for Hematopoietic Stem Cell Transplantation for Hemoglobinopathies." *Bone Marrow Transplantation* 31 (7): 547–50. https://doi.org/10.1038/sj.bmt.1703887.

Nematollahi, M, R Akbari, S Nikeghbalian, and C Salehnasab. 2017. "Classification Models to Predict Survival of Kidney Transplant Recipients Using Two Intelligent Techniques of Data Mining and Logistic Regression." *International Journal of Organ Transplantation Medicine* 8 (2). Avicenna Organ Transplantation Institute: 119–22. http://www.ncbi.nlm.nih.gov/pubmed/28959387.

Ochs, Robert A., Jonathan G. Goldin, Fereidoun Abtin, Hyun J. Kim, Kathleen Brown, Poonam

Batra, Donald Roback, Michael F. McNitt-Gray, and Matthew S. Brown. 2007. "Automated Classification of Lung Bronchovascular Anatomy in CT Using AdaBoost." *Medical Image Analysis* 11 (3): 315–24. https://doi.org/10.1016/j.media.2007.03.004.

Onitilo, A.A., Y.H. Lin, E.C. Okonofua, L.B. Afrin, J. Ariail, and B.C. Tilley. 2004. "Race, Education, and Knowledge of Bone Marrow Registry: Indicators of Willingness to Donate Bone Marrow among African Americans and Caucasians." *Transplantation Proceedings* 36 (10): 3212–19. https://doi.org/10.1016/j.transproceed.2004.10.019.

Ridgeway, Greg. 2006. "Generalized Boosted Regression Models." *Documentation on the R Package "Gbm", Version* 1: 7.

Schapire, Robert E., Yoav Freund, Peter Bartlett, and Wee Sun Lee. 1998. "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods." *Annals of Statistics* 26 (5): 1651–86. https://doi.org/10.1214/AOS/1024691352.

Shiao, Ht, and Vladimir Cherkassky. 2013. "SVM-Based Approaches for Predictive Modeling of Survival Data." *Icdm*. http://world-comp.org/p2013/DMI8050.pdf.

Switzer, Galen E, Jessica G Bruce, Larissa Myaskovsky, Andrea DiMartini, Diana Shellmer, Dennis L Confer, Linda K Abress, et al. 2013. "Race and Ethnicity in Decisions about Unrelated Hematopoietic Stem Cell Donation." *Blood* 121 (8): 1469–76. https://doi.org/10.1182/blood-2012-06-437343.

Switzer, GE, MA Dew, AA Stukas, JM Goycoolea, J Hegland, and RG Simmons. 1999. "Factors Associated with Attrition from a National Bone Marrow Registry." *Bone Marrow Transplantation* 24 (3): 313–19. https://doi.org/10.1038/sj.bmt.1701884.

Verplancke, T, S Van Looy, D Benoit, S Vansteelandt, P Depuydt, F De Turck, and J Decruyenaere. 2008. "Support Vector Machine versus Logistic Regression Modeling for Prediction of Hospital Mortality in Critically Ill Patients with Haematological Malignancies." *BMC Medical Informatics and Decision Making* 8 (1): 56. https://doi.org/10.1186/1472-6947-8-56.

Veta, Mitko, Paul J. van Diest, Stefan M. Willems, Haibo Wang, Anant Madabhushi, Angel Cruz-Roa, Fabio Gonzalez, et al. 2015. "Assessment of Algorithms for Mitosis Detection in Breast Cancer Histopathology Images." *Medical Image Analysis* 20 (1). Elsevier: 237–48. https://doi.org/10.1016/j.media.2014.11.010.

Xu, Qian, Yi Xiong, Hao Dai, Kotni Meena Kumari, Qin Xu, Hong Yu Ou, and Dong Qing Wei. 2017. "PDC-SGB: Prediction of Effective Drug Combinations Using a Stochastic Gradient Boosting Algorithm." *Journal of Theoretical Biology* 417 (March). Academic Press: 1–7. https://doi.org/10.1016/j.jtbi.2017.01.019.

Appendix A

Post Recruitment Survey Inputs

The four post recruitment survey questions

1. Imagine that we are calling you today to ask you to give a blood sample for further testing to see if you are the best match for a patient. On a scale of 1 to 5 where 1 is very likely and 5 is not at all likely, how likely are you to agree to give a blood sample?

2. If you are selected as a match for a patient, the total time commitment for testing and actual donation is about 40 hours over a 4-6 week period. On a scale of 1 to 5, how willing would you be to make that time commitment to donate?

3. If further testing showed that you were the best match for a patient, on a scale of 1 to 5, how likely is it that you would actually donate?

4. On a scale of 1 to 5 where 1 is very willing and 5 is not at all willing, how willing would you be to donate to someone you don't know?

   Responses of 7 (don't known) and 8 (left blank) were also possible

Encoding responses to questions

Responses of 1 and 2 were encoded as positive
Responses of 3-5 were encoded as negative.
Responses of 7 and 8 were encoded as negative

Composite Scoring

Calculating the composite score from the four Post Recruitment Survey (PRS) questions:

PRS response is positive if the composite score is 4.
PRS response is negative if the composite score >4 (maximum score would be 8).

If response to a question is 1 or 2, it is counted in the composite calculation below as a 1.
If response to a question is 3, 4 or 5, it is counted in the composite calculation below as a 2.

If two or more responses are missing then the PRS response is negative.
If there is only one response missing, then that response is a 1 in the composite calculation below.
A response of 7 or 8 is the same as a missing score.

PRS composite score = (Converted as indicated above)
Sum of converted responses to questions 1 through 4.

Appendix B

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 modeling for their assumed effect on availability.

Table 4: Marginal Availability rate by Gender.

Gender	***Availability Rate***	***Count***
Male	55.8%	106,182
Female	55.9%	72,067

Table 5: Marginal Availability by Health History Questionnaire Response

Questionnaire Response	***Availability Rate***	***Count***
Responded	87.1%	10,961
Not Responded	53.8%	167,288

Table 6: Marginal Availability Rates by Donor Centers

Donor Center Code	***Network Name***	***Network Code***	***Availability Rate***	***Count***
A	NMDP	1	50.4%	100,057
B	INT	4	67.6%	470
C	INT	4	62.1%	863
D	DOD	3	39.9%	22,140
E	DKMS	2	79.6%	31,332
F	INT	4	69.3%	990
G	INT	4	72.6%	212
H	INT	4	76.5%	243
I	US - NN	5	64.8%	3,062
J	DKMS	2	58.9%	13,708
K	US - NN	5	48.8%	1,695
L	DKMS	2	75.5%	3,152
M	INT	4	68.7%	323

Table 7: Marginal Availability Rate by Post Recruitment Survey Response

Post Recruitment Survey	***Availability Rate***	***Count***
Response	75.2%	761
No Response	55.8%	177,488

Table 8: Marginal Availability by Recommit Response

Recommit Request	***Availability Rate***	***Count***
Response	93.5%	11,158
No Response	53.4%	167,091

Table 9: Marginal Availability Rates for DIY record

Do-It-yourself Registration	***Availability Rate***	***Count***
Online	64.6%	86,863
Others (Donor Drive)	47.6%	91,386

Table 10: Marginal Availability Rate for Change of Address requests

Change of Address Request	***Availability Rate***	***Count***
Self-Initiated	79.8%	25,682
No Action	51.2%	152,567

Appendix C

We use binary classifiers (as we have two categories: *available* and *not available*) to model the availability. Classifiers estimate a function $f(\boldsymbol{x})$ that separates the input space into two regions corresponding to the two classes, where $\mathbf{x}$ is a feature vector that is used to represent measurements/observations for a data instance. For our problem, $\mathbf{x}$ is a vector representing member characteristics listed in Section 2. Each member 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, $f(\boldsymbol{x})$ is the model predicted output for sample $\mathbf{x}$.

Classifiers typically have tuning parameters that need to be adjusted by the user according to the data being modeled; 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 modeling 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*. However, the quality of prediction is determined by the classifier performance on the test set. None of the samples in the testing set is used for model selection; the 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 modeling problem. We refer readers to (Cherkassky 2013; Cherkassky and Mulier 2007) for a detailed discussion on model selection procedure and further details on modeling techniques.

Data Pre-Processing

Raw data is processed before model training and is an important part of the modeling 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 B with only two possible values, are encoded as binary variables. Care is taken to identify missing values and miscellaneous responses to outreach requests.