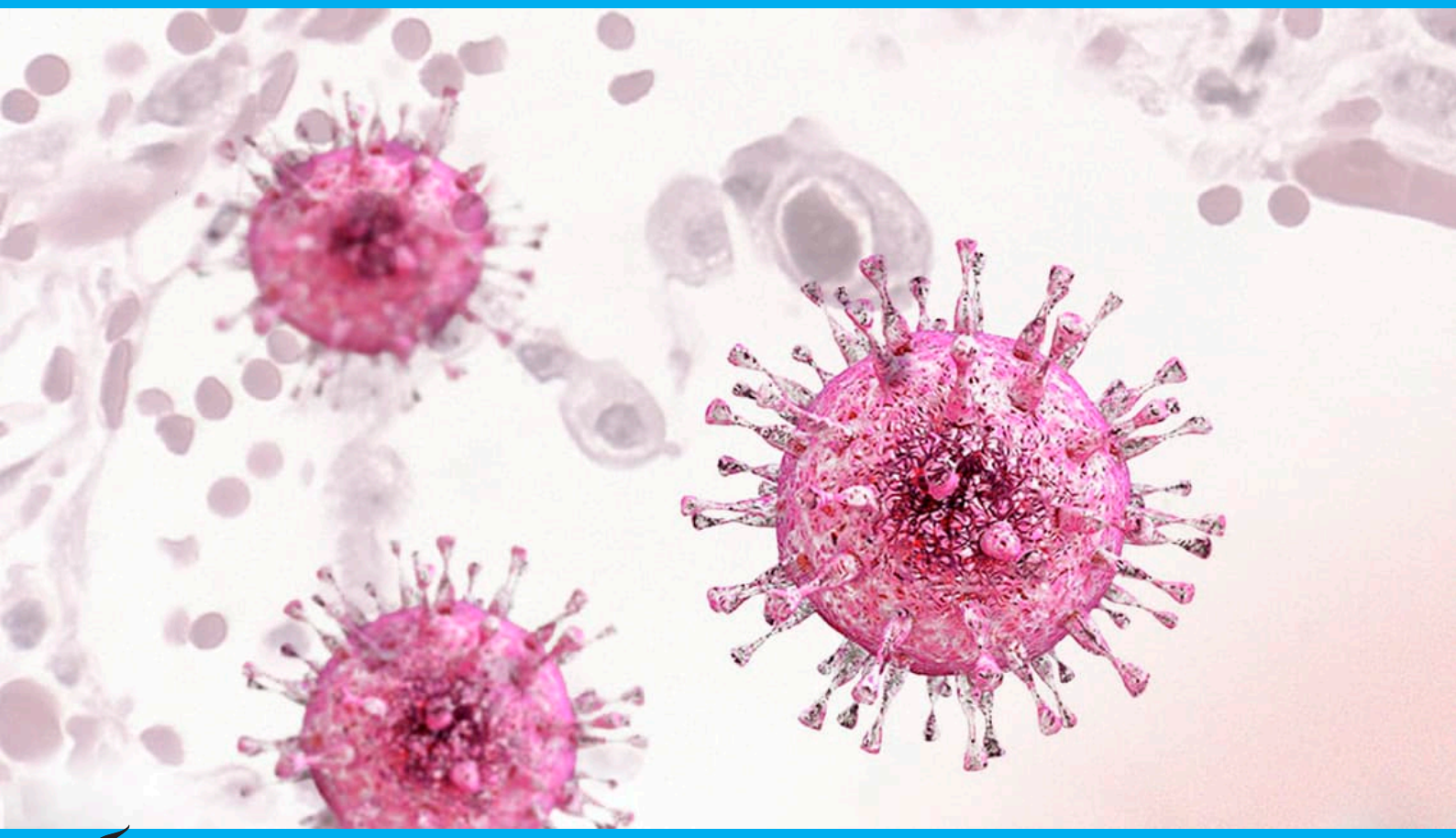
Predicting CMV Serostatus using Donor Data

LJF Hendriks



TU Delft

Predicting CMV Serostatus using Donor Data

by

LJF Hendriks

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday June 27, 2025 at 16:00.

Student number:	4605594	
Project duration:	October 23, 2023 – June 27, 2025	
Thesis Advisor:	M.J.T. Reinders	TU Delft
Daily Supervisor:	E.B. van den Akker	TU Delft
Daily Co-Supervisor:	E.G.J. von Asmuth	LUMC
Medical contact:	A. Lankester	LUMC
External contact:	A. Venter	WMDA

An electronic version of this thesis is available at https://repository.tudelft.nl/.

TU Delft

Contents

Abstract

The Cytomegalovirus (CMV) serostatus, of both donor and patient, plays an important role in allogeneic hematopoietic stem cell transplantation, yet it is only known for 19% of donors in the global database. In this research, the other available data in the global database of the World Marrow Donor Association will be used to predict CMV serostatus for donors who's status is unknown. In a statistical analysis, features such as sex, registry, ethnicity, age and height were found to be informative. A particular focus was put on investigating the relation between Human Leukocyte Antigen (HLA) and CMV. Previous literature on this relation consists of studies on small cohorts, Machida et al., 1998 (N=125) and Hassan et al., 2016 (N=1 849). The large global database (N=8 707 407 allows us to evaluate this relation for many more different HLA groups and validate their findings using a meta-analysis across registries. In this analysis, the majority of HLA groups showed small effects on the CMV serostatus. However, a few groups showed consistent large increases in likelihood to be CMV seropositive. This indicates that there is a biological relation between specific HLA and CMV serostatus. To predict CMV serostatus, a simple logistic regression classifier was iterated on by adding features and using more complex models. The best performing classifier uses XGBoost on all features in the database. This classifier has an AUC performance of 0.70. Although this AUC is too low to rely on it for patient-donor matching, it does show non-trivial predictive power. The classifier could be further improved by using a better embedding for the HLA that includes the similarities between different HLA alleles.

Acknowledgements

First and foremost, I would like to thank Erik van den Akker and Erik von Asmuth for the opportunity to do this research and their guidance during the project. Although there were many bumps in the road during this project, they helped and supported me throughout. I appreciate the discussions we had during our weekly meetings and the insights they let to. Furthermore, I would like to thank Alicia Venter for the support from the WMDA and for allowing me to do this research on their data. Conversations with her and Mark Melchers helped give a deeper understanding of the donor and HLA data. Additionally, I would like to thank Marcel Reinders and Arjan Lankester for their support behind the scenes and the insightful feedback they gave during the important meetings.

Lastly, I would like to thank my friends and family for their ongoing support during my master thesis. In specific, Rebecca and Paul for taking the time to proofread my thesis.

Leo Hendriks
Delft, June 2025

Abbreviations

AHSCT	Allogeneic Hematopoietic stem cell transplantation
AUC	Area Under Curve
CMV	Cytomegalovirus
EBMT	European Society for Blood and Marrow Transplantation
EBV	Epstein-Barr Virus
GRID	Global Registration Identifier for Donors
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigen
IDM	Infectious Disease Marker
ION	Issuing Organisation Number
KIR	Killer-cell immunoglobulin-like receptor
OR	Odds Ratio
PBCC	Point-Biserial Correlation Coefficient
ROC	Receiver-operating characteristic
SD	Standard Deviation
SHAP	SHapley Additive exPlanations
WLS	Weighted Least Squares
WMDA	World Marrow Donor Association
XGBoost	Extreme Gradient Boosting

1 Introduction

Cytomegalovirus (CMV) is a common virus belonging to the herpes virus family. The virus spreads through bodily fluids, including saliva, urine, blood and tears. Once someone is infected with CMV, the virus remains in their body for life. Therefore, seroprevalence is greatly age-dependent: from 36.3% of 6-11-year-olds to 90.8% for those aged above 80 years in the US, according to Staras et al., 2006. Fortunately, in healthy individuals with a functioning immune system, CMV infection usually causes few to no symptoms. However, for allogeneic hematopoietic stem cell transplantation (AHSCT) CMV serostatus is an important factor.

For an AHSCT, the European Society for Blood and Marrow Transplantation (EBMT) recommends to match the CMV serostatus of both donor and patient (Carreras et al., 2019). If the patient is seronegative, the transplantation from a seropositive donor carries the risk of CMV infection to the patient. Conversely, if the patient is seropositive there is a risk of viral reactivation. Since patients undergoing bone marrow transplantation are immuno-compromised, this infection or reactivation can lead to pneumonia and gastroenteritis and a general decrease in overall survival of the patient (Carreras et al., 2019). Therefore, it is important when matching patient and donor to match CMV positive patients to positive donors and CMV negative patients to negative donors. In the global donor database, managed by the World Marrow Donor Association (WMDA), the CMV serostatus is given for only 19% of donors (World Marrow Donor Association, 2023). When the CMV status of the donor is unknown this can result in a delay to transplantation. The donor has to be tested pre-donation. Consequently, it can mean that the donor does not match the patient and another donor has to be found. In this thesis we will investigate correlations between other fields in the donor database and CMV serostatus. Then machine learning will be used to create a predictor for CMV serostatus that can be used for donors with unknown CMV serostatus.

In Cannon et al., 2010, a review of a number of papers is done on the relation between different demographic characteristics and CMV seroprevalence. As expected, a strong correlation between age and CMV serostatus was found. Geographic location also greatly affects the prevalence of CMV in the population. Lastly, they found a relation between ethnicity, sex and CMV serostatus. In this thesis we will verify these results on a more global scale and on a larger dataset. In the global donor database these and more fields are available. The rough geographic location of the donor is indirectly available trough the location of the registry the donor is registered with. A correlation between Epstein-Barr Virus (EBV), another herpesvirus, and CMV serostatus was found in a number of studies (Lazda, 2006). For none of the other Infectious Disease Markers (IDM) a relation to CMV serostatus was found in the literature and therefore they were not included in this study. Finally, the Human Leukocyte Antigens (HLA) and Killer-cell Immunoglobulin-like Receptors (KIR) of the donor are listed in the database.

The HLA genes are the most important matching criteria for AHSCT, according to Carreras et al., 2019. Therefore, an in-depth HLA typing is available for almost all donors in the database. The HLA genes encode the cell-surface proteins that are used by the immune system. The locus refers to the specific location the HLA gene is located. In matching the five most important loci are A, B, C, DRB1 and DQB1, Carreras et al., 2019. Though other loci, like DQA1, are available in the database they are often not listed. Because the HLA genes are related to the function of the immune system, they could affect the likelihood to be CMV positive. Two studies were found that studied the effects of

HLA on CMV serostatus. In Machida et al., 1998 (N=125) two HLA types, HLA-A33 and HLA-Cw4, were found that may be associated with long-term CMV seronegativity in Japanese donors. Hassan et al., 2016 (N=1 849) found that in Irish organ donors the presence of HLA-A1 is associated with reduced susceptibility to CMV infection. This research will attempt to improve on these two studies by investigating the relation between HLA and CMV serostatus on a much larger scale using the WMDA global data set.

Using the information on the HLA alleles in conjunction with available demographic and medical variables, machine learning will be used to predict the CMV serostatus. These variables include age, sex, registry, ethnicity, blood type, rhesus, CCR5, weight and height. For prediction, both logistic regression and XGBoost (Chen and Guestrin, 2016) will be used. One of the challenges is how to make the varied HLA data usable for training. First, the number of unique values will be reduced by converting the HLA data to their G-groups and P-groups and then those groups will be encoded using a Two Hot Encoder that will be created. The classifiers trained on this data will be validated on their performance and analysed for further improvements that could be made.

This research consists of two main chapters. In chapter 2, an extensive statistical analysis will be done on the available data in the donor database and their correlation to CMV serostatus. Particular focus will be put on identifying specific HLA alleles with strong correlations. After finding these features and HLA with strong correlation to CMV serostatus, these will be used to predict CMV serostatus with machine learning in chapter 3. After these two chapters, there will be a main conclusion in chapter 4 which will combine the conclusions from both chapter followed by a discussion of the limitations of this research and options for further research in chapter 5. All the code that was used to generate the results found in this paper is available at https://git.lumc.nl/lhendriks/cmv_thesis.

2

Statistics

2.1. Methods

This chapter introduces the statistical methods that were used to find relations in the donor data. First, some preprocessing steps are discussed that need to be done on the data to prepare it for analysis. Then the the univariate analysis for the non-HLA features is explained followed by the preprocessing that is done to convert the HLA into P and G-groups. Finally, the methods for the meta-analysis across the different registries is addressed.

2.1.1. Preprocessing

Preprocessing is required before the WMDA database is used in statistics and machine learning. In this subsection the preprocessing steps that were done on Age, CMV and EBV serostatus are discussed.

CMV serostatus

CMV serostatus is tested using an antibody test to test for the antibodies that are produced in response to a CMV infection. The IgM antibodies can be detected in the first two weeks after initial exposure to CMV or reactivation. The IgG antibodies are produced several weeks after the initial infection and are measurable for the rest of your life. In table 2.1 each of the different test outcomes, that can be found in the WMDA database, is listed together with their IgG and IgM result, sourced from secion 2.5 of World Marrow Donor Association, 2025.

In the WMDA database the results of the CMV test are listed and we want to convert those to a simple boolean outcome that can be used in the statistics and as an outcome for the classifier. The table lists how we convert each test outcome to a boolean value in the ”CMV status” column. The IgG test outcomes were used to do the conversion since it more reliable. The IgM test outcome is less reliable, since it only tests positive shortly after the initial exposure and because there is a high false positive rate for EBV serpositive patients (Miendje Deyi et al., 2000). For this reason, the M-group was excluded from our research, since it relies on only a positive IgM test and the group is relatively small. For 78.47% of the database the CMV serostatus is unknown, these donors will not be used in the further statistics or the machine learning. There are more CMV negative than positive donors in the database. However, this imbalance is not so large that it is necessary to adjust for it when implementing the machine learning.

Value	Count	Perc.	IgG	IgM	CMV status	Count	Perc.
N	1 984 489	4.91%	negative	negative	negative	5 391 471	13.33%
O	3 406 982	8.42%	negative	unknown			
P	880 053	2.18%	positive		positive	3 315 936	8.20%
B	49 601	0.12%	positive	positive			
G	210 070	0.52%	positive	negative			
H	2 176 212	5.38%	positive	unknown			
M	3 133	0.01%	negative	positive	unknown	31 735 330	78.47%
NULL	31 732 197	78.46%	unknown	unknown			
Total	**40 442 737**	**100.00%**					

Table 2.1: Count of donor data available for each outcome of the antibody CMV test present in the WMDA database and what boolean value it is converted to. For the ”P” outcome the test did not differentiate between IgG and IgM antibodies.

EBV serostatus

As mentioned in chapter 1, Lazda, 2006 have shown that there is a strong relation between EBV and CMV serostatus. To investigate this relation in our dataset we also want to convert the EBV values in the database to boolean values. The conversion for EBV is done analogously to CMV. In table 2.2 each of the different test outcomes is listed and their counts in the donor database. In the counts only donors are included for whom the CMV serostatus is also given. There are only 598 donors in the database for whom EBV serostatus is given but CMV serostatus is not.

Value	Count	Perc.	IgG	IgM	CMV status	Count	Perc.
N	7 123	0.08%	negative	negative	negative	7 302	0.08%
O	179	0.00%	negative	unknown			
P	342	0.00%	positive		positive	47 875	0.55%
B	2 397	0.03%	positive	positive			
G	43 969	0.50%	positive	negative			
H	1 167	0.01%	positive	unknown			
M	79	0.00%	negative	positive	unknown	8 652 230	99.37%
NULL	8 652 151	99.37%	unknown	unknown			
Total	**8 707 407**	**100.00%**					

Table 2.2: Count of donor data available for each outcome of the antibody EBV test for which CMV serostatus is also known in the WMDA database. Again, what EBV status it is converted to is listed and for the ”P” outcome the test did not differentiate between IgG and IgM.

Age

Age has a very strong correlation to CMV serostatus, Staras et al., 2006, therefore we want to calculate the age at testing for each of the donors. We are not interested in the current age of the donors, since it is not known if the donor has become CMV positive in the meantime. In the database the date of birth of the donor and the date of the most recent CMV antibody test is given. We can substract these columns to get the age at testing. For the date of birth just the month is given and the day is removed to further anonimise the data. Therefore, this is an approximation of the age at testing, but this approximation is accurate enough for our purposes.

2.1.2. Univariate Analysis

Metrics

To investigate the effects that single features have on the CMV serostatus we will be using two metrics. The first metric, the odds ratio (OR), is used for boolean or categorical features like Sex, donor registry and EBV serostatus. The odds ratio represents by what factor the donor is more likely to be CMV positive when that characteristic is present. Therefore, an odds ratio of 1 means there is no effect on CMV serostatus. An odds ratio of 2 means the donor is twice as likely to be CMV positive and an odds ratio of 0.5 means the donor is twice as likely to be CMV negative.

To calculate the odds ratio the number of donors with and without the feature are counted and split up by CMV serostatus. When looking at a categorical feature, like donor registry, the group is split by those part of a specific category and all others. The odds ratio is then calculated using these values (table 2.3) and formula 2.1.

	CMV positive	CMV negative
With feature	a	b
Without feature	c	d

Table 2.3: Values needed to calculate the odds ratio.

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc} \tag{2.1}$$

The second metric, the Point-Biserial Correlation Coefficient (PBCC), Lev, 1949, is used for continuous features like Age, Length and Weight. The PBCC is a value between -1 and 1 that represents how strongly the feature is correlated to the CMV serostatus. A value closer to 1 means the features are positively correlated, therefore an increase in the feature will mean the donor is more likely to be CMV positive. A value closer to -1 means a negative correlation and an increase in the value will result in a lower likelihood of CMV positivity. When the value is around 0 there is no relation between the feature and CMV serostatus.

The PBCC is calculated using formula 2.2. In this formula M_1 is the mean of the feature for all CMV positive donors and M_2 is the mean for all CMV negative donors. The standard deviation s_{n-1} over all data is calculated using formula 2.3, with $\bar{X}$ as the mean of the entire feature and X_i for specific feature values. Furthermore, n_1 is the number of positive donors and n_0 the number of negative donors. Finally, n is the total number of donors for whom the feature is given.

$$PBCC = \frac{M_1 - M_0}{s_{n-1}} \sqrt{\frac{n_1 n_0}{n(n-1)}} \tag{2.2}$$

$$s_{n-1} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{2.3}$$

Plots

Next to these two metrics a number of plots were used to investigate the relation between certain features and CMV serostatus. In these plots a continuous feature is plotted, usually age, versus the fraction of the population that is CMV positive. We can then split these plots over certain categories like sex or donor registry to investigate how that categorical feature affects the CMV serostatus distribution. An example of one of those plots can be found in figure 2.6.

Instead of using a histogram, a more accurate kernel density function approach was implemented. A gaussian kernel density function, Silverman, 1986, was used to estimate the distribution of all CMV positive donors over the continuous feature (f_p). Then, the same distribution is created for all CMV negative donors (f_n) and both are scaled by the number of donors in the distribution, n_p and n_n. Finally, we obtain the distribution of the fraction of CMV positive donors (f) by dividing the positive distribution by the sum of the two distributions, see formula 2.4. The resulting function gives a much more accurate view of the distribution of CMV over the population.

$$f(x) = \frac{n_p f_p(x)}{n_p f_p(x) + n_n f_n(x)} \tag{2.4}$$

The function does not show how many donors are used to create a certain part of the distribution. It could be that there are only a couple of hundred donors that make up a certain part of the distribution, but it could also be hundreds of thousands of donors. The accuracy and the statistical relevance of a section depend heavily on the number of donors that were used to create that section of the distribution.

Therefore, it is important to know how many donors are in a part of the distribution. To solve this problem an kernel density function over all donors in the category was added, which shows the total density of the donors over the feature. This function is scaled by the number of donors in the category divided by the total number of donors and is then plotted in the bottom of the figure, see figure 2.6 for an example. This density function is also used to not plot any values for which the density function falls below a cutoff of 1 000 donors.

Vulcano plots will be used to compare many different categories, like different HLA groups or registries, and investigate their impact on the CMV serostatus. In a vulcano plot, see figure 2.10 for example, the odds ratio vs the P value are plotted for a number of different categories. Both categories are plotted on a logarithmic scale. The odds ratio is plotted on the horizontal axis. Any ratios below $0.5 = 2^{-1}$ and above 2 show a significant impact of that category on the CMV serostatus, therefore two vertical dashed lines were added to easily identify these significant regions. On the vertical axis the P value is plotted. The P value is the probability of obtaining this CMV serostatus result when the odds ratio is 1. The lower this value is the less likely that the found odds ratio is due to random chance. The P values are plotted in a reverse log scale, therefore the higher the data point is in the plot the better the P value is. Since we have a very large dataset most of our P values are very small. For some categories the P value is so small that it becomes less than the minimum float value in Python and gets rounded to 0.0. For these values their P-value are set to 5×10^{-324}, which is the minimum float value in Python, so they can be plotted on the log-scale. For our dataset any P value below 1×10^{-10} is significant. In our vulcano plots we have added a horizontal line at this level, therefore any values above this line in the graph are significant. We have now identified two regions in the graph where we can find the significant values, the top left and the top right region. Any values in these regions are highlighted in a slightly darker colour and are identified by the name of their category followed by the number of donors in that category between brackets.

2.1.3. HLA groups

The HLA data provided by the WMDA is not homogeneous and contains HLA information reported at varying levels of precision. Sometimes the data consists of allele data where just the HLA protein is given like ”A*02:101”, while other times field 3, 4 and a suffix are also given, like ”A*02:101:01:02N”. See figure 2.1 for more details. In other cases not a specific allele is given but a MAC-code is used instead. MAC-codes are used to specify a group of possible allele codes that someone can have. An example is ”A*02:AA” that can be one of the following alleles ”A*02:01/A*02:02/A*02:03/A*02:05”. Finally, for some donors the HLA information is given as a G-group or P-group. A G-group consists of HLA alleles with the same nucleotide sequence across the exons encoding the peptide binding domains. A P-group consist of HLA alleles with the same antigen binding domains (Anthony Nolan Research Institute, 2024). Two examples are respectively ”A*02:01:01G” and ”A*02:01P”. To make the HLA data more homogeneous and better to use in machine learning we try to reduce the HLA data to either a G-group or P-group. We chose to reduce all the HLA data to these groups, since there are many different HLA alleles and this would help us to reduce the search space during the statistical analysis and machine learning, while still keeping the HLA that present different proteins separated.

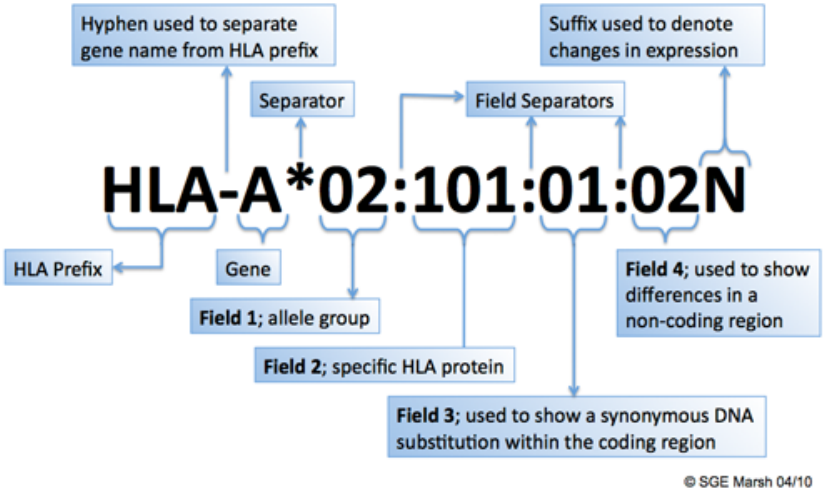


Figure 2.1: HLA naming nomenclature [Anthony Nolan Research Institute, 2024]

First, a number of preprocessing steps are done to speed up the conversion to the groups. The first step is to use the data published by Marsh, 2024a to create two mappings. This file is a list of each of the G-groups with all alleles that are part of that group.

The first mapping is from a combination of the number of fields and HLA alleles to the G-group that allele is a part of. The number of fields is included to speed up the process. To create this mapping we loop over all alleles in the data and add it and all the subgroups from length 2 to the file. The ”N” suffix denotes that the allele is not expressed, which is therefore only relevant for G-groups. This suffix is kept for each subgroup if there is one. For example for ”01:01:01:02N”, each of ”01:01:01:02N”, ”01:01:01N” and ”01:01N” are added to the mapping and all map to the group ”01:01:01G” with their respective lengths of 4, 3 and 2. Since all of these HLA alleles are not expressed, it would also be an option to map all of these alleles to the same value, indicating that they are not expressed. We decided not to use this option, since keeping them as separate columns gives the classifier for predicting CMV serostatus more information that it might be able to leverage when training.

The second mapping is the inverse of the first where a given G group is mapped to its most common allele. The assumption is made that the first allele listed for a group is the most common allele group in that group. This is based on that the alleles are listed in increasing order and they are numbered based on when they were discovered. Therefore, we map each G group to its first allele reduced to a length of 2. This assumption does introduce a bias towards populations that were tested first. Since most testing was done first in western countries (Bull World Health Organ, 1968), this introduces a bias towards European and North-American ethnicities.

These two mappings can be created similarly for P groups using the file by Marsh, 2024b. This process is then repeated for each of the different loci resulting in 4 mappings per loci.

The second preprocessing step is to create a mapping from each of the different MAC-codes to their most common subtype. We make this mapping using the data provided by NMPD, 2024. In this file each of the MAC-codes is listed with their subtypes seperated by a ”/” and a column containing a * if the subtypes should replace the entire HLA allele. An example without a * is ”AA” with the subtypes ”01/02/03/05”. An example with a * is ”GE” with the subtypes ”02:01/32:01” consisting of two fields each. We want to map each of the subtypes to its most common subtype and therefore we will make the same assumption as for the groups that the first subtype listed is the most common option. This again introduces the same bias towards. Using this assumption in our mapping, ”AA” is mapped to ”01” and ”GE” to ”02:01”. The * is also included in the mapping.

Now that we have gone over the preprocessing step we will first go over a number of different common examples. In figure 2.2 it is visualized how to convert each of those common examples to either a G- or P-group by following the arrows. After these common examples, we will discuss a number of edge cases and how they are handled.

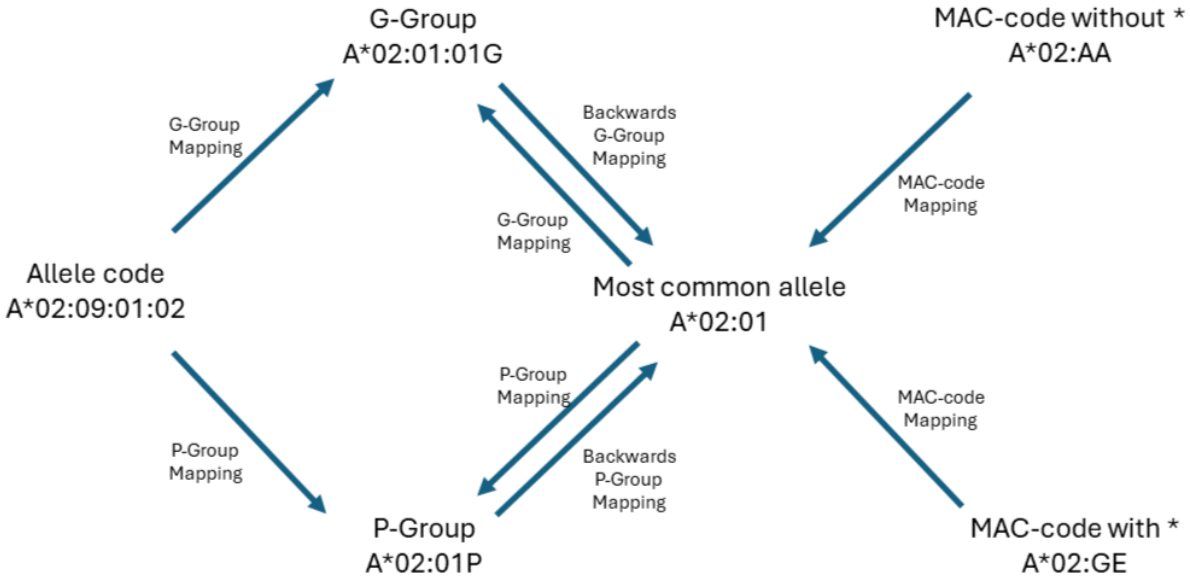


Figure 2.2: This figure shows how most of the different values in the WMDA database can be converted to a G- and P-group.

- **Allele code (A*02:09:01:02N)**: For a specific allele code we lookup the code in the G-group and P-group forward mappings using the code itself and its length to directly find the correct G- and P-group. If the code is not present in one of the mappings we will keep the code itself. Some allele codes, for example, only have a G-group but are not part of a P-Group or vise versa. In that case we want to keep the allele code in the P-group column.
- **G-group (A*02:01:01G)**: When a G-group is given we can directly enter that value for the G-group column. To obtain the P-group we use the backwards G-group mapping to obtain the most common allele in the G-group and then use the P-group mapping to obtain the most likely P-group that that G-group is part off.
- **P-group (A*02:01P)**: When a P-group is given we do the exact same as for a G-group but with the opposite mappings. We map the P-group to its most common allele by using the backwards P-group mapping and then map that to its corresponding G-group.
- **Mac-code without * (A*02:AA)**: For a MAC-code without a star we strip the code from the end and replace it with the most common subtype found in the MAC-code mapping, i.e. for ”A*02:AA” we get ”A*02:01”. We then use this Allele code similar to a regular allele code to find the groups.
- **Mac-code with * (A*02:GE)**: For a MAC-code with a star we remove the entire allele code and replace it with the code found in the MAC-code mapping, i.e. for ”A*02:GE” we get ”A*02:01”. Then we again use the group mappings to find the G- and P-groups.
- **No information provided (None)**: For a number of donors the HLA information is not provided, especially for the less important loci to stem cell donation like DRB5. When ”None” is provided we also return ”None” for both the G- and P-groups.

Finally there are a number of special cases or exceptions which are taken into account as following:

- **Suffixes**: Wherever applicable the suffixes ”L”, ”S”, ”C”, ”A” and ”Q” are removed, since none of them have any influence on the group the allele is part off nor are relevant for our research purposes.
- **DRBX*NNNN**: This code is to designate the absence of the DRB3, DRB4 or DRB5 gene on the chromosome. Therefore this code is mapped to ”None” for both groups.

- **XXXX**: This code is used to designate an unknown value, it is treated as no information provided.
- **XX code (A*02:XX)**: The code ”XX” is used to denote all possible subtypes for that allele code. This usually results in an enormous number of subtypes. These XX codes are treated as missing data since they are very ambiguous. They are usually the result of a different serological HLA test.

The database contains two columns holding the HLA information for each of the loci. For each of those columns these steps will be used to create two new columns holding the G-group and P-group information for that loci. The intention of these steps is to make the data more homogeneous and more easily comparable for the different HLA in our prediction models.

2.1.4. Meta Analysis

The final statistical analysis is a meta analysis. Usually a meta analysis is done to compare and summarize the statistical results of multiple different studies. This meta analysis instead compares the effect of HLA on CMV serostatus within different registries. We want to verify that the effects, that we find using the odds ratio, were not caused by confounding between the registries and the HLA. A logistic regression per registry is done using age, sex and the HLA group. To compare the registries, the coefficient used in front of the HLA group will be the effect size in the meta analysis. To do the meta analysis, Harrer et al., 2021 and the pythons StatsModels package (Seabold and Perktold, 2010) were used. The package gives 4 different estimates for the effect size: fixed effect, random effect, fixed effect weighted least squares (WLS) and random effect WLS. All four of these are displayed in the forest plot that we generate to show the results of the meta analysis. As an example, the following section details the steps used to conduct the fixed-effect meta-analysis. Afterwards, the metrics that were used to investigate heterogeneity and to the generated forest plot are discussed.

In each meta analysis, the effect of the presence of a single HLA group on the CMV serostatus is investigated. Since there are so many different HLA groups, it is necessary to make a selection of HLA groups to investigate. For our research, a meta analysis will be done for the 3 HLA groups with the biggest odds ratios for each of the 5 most important HLA loci (A, B, C, DQB1 and DRB1) in stem cell donation, Carreras et al., 2019. All donors with a specific HLA group are then split into subsets per registry. The subscript k is used to identify the registry with K total registries. The first step in the meta analysis is to do the logistic regression to calculate the effect size for each subset. For the logistic regression, the function in equation 2.5 is used on the variables: age (x_{age}), sex (x_{sex}) and HLA group (x_{HLA}). Polynomial inputs to the third degree for age with an interaction with sex were added to allow for a polynomial relation with both age and sex. This decision is based on the results we found in the univariate analysis in figure 2.6.

In the equation 2.5, c_k^0 through c_k^7 and θ_k are the coefficients that are optimized in the logistic regression. θ_k is the coefficient that will be used as the effect size in the meta analysis.

$$f_k(x_{age}, x_{sex}, x_{HLA}) = c_k^0 + c_k^1 x_{age} + c_k^2 x_{age}^2 + c_k^3 x_{age}^3 + c_k^4 x_{sex} + c_k^5 x_{sex} x_{age} + c_k^6 x_{sex} x_{age}^2 + c_k^7 x_{sex} x_{age}^3 + \theta_k x_{HLA} \tag{2.5}$$

The logistic regression also gives the standard error s_k of each of the coefficients which can be used to calculate the weights for the meta analysis, w_k, using equation 2.6. With these weights we can do the meta analysis and find the fixed effect size θ over all registries using equation 2.7. A θ of around zero shows that there is no interaction between the HLA and CMV serostatus. A positive θ shows that the HLA makes it more likely for the donor to be CMV positive. Inversely, a negative θ makes a CMV negative serostatus more likely.

$$w_k = \frac{1}{s_k^2} \tag{2.6}$$

$$\theta = \frac{\sum_{k=1}^{K} w_k \theta_k}{\sum_{k=1}^{K} w_k} \tag{2.7}$$

After calculating the effect size, two metrics for the heterogeneity between registries are calculated. The first is Q (Cochran, 1954) which can be calculated using equation 2.8. The second metric I^2

(Higgins and Thompson, 2002) is a scaled version of Q calculated using equation 2.9. I^2 is scaled to be between 0 and 1 (a negative value for I^2 is commonly rounded to 0). The closer the value for I^2 is to 1 the more the variability in the meta analysis is explained by between-registry heterogeneity rather than by sampling error.

$$Q = \sum_{k=1}^{K} w_k(\theta_k - \theta)^2 \tag{2.8}$$

$$I^2 = \frac{Q - (K-1)}{Q} \tag{2.9}$$

To the results of the meta analysis will be summarized using a forest plot, like figure 2.12. In these plots the registries are grouped by continent and ordered by the number of donors with the given HLA in that registry. Included in the plot are also the total number of donors in that registry and the total percentage of CMV seropositivity in that registry. To the right of the plot the weight in the random and fixed effect model are shown. Below the title of the plot the calculated Q and I^2 are shown. The p-value shows the statistical significance of the meta analysis.

2.2. Results & Discussion

2.2.1. Univariate Analysis

Metrics

In table 2.4, a selection of the categorical features that are available in the database are listed with their odds ratios. The table shows that there are significantly more female than male donors, this fact also holds for not just the subset of donors for whom CMV serostatus is given. As expected based on Cannon et al., 2010, female donors are much more likely to be CMV seropositive than male donors. However, this effect is much stronger than what is described in the paper. In figure 2.6 we will look more in depth at this feature. For blood type, rhesus and CCR5 the effects seem to be negligible compared to the other odds ratios that were found. The strong correlation between EBV and CMV serostatus that was described in Lazda, 2006 is confirmed with a relatively high odds ratio for EBV seropositive donors. Though the number of donors with both EBV and CMV serostatus is relatively small, this number is significant enough to confirm the correlation

For both registry and ethnicity there are a number of categories that show a significant correlation with CMV serostatus. As in Cannon et al., 2010 seroprevalence in Western Europe and the USA is low. This can be seen by the low odds ratios for Caucasian ethnicity's and for the ZKRD, BBMR, Anthony Nolan, Matchis and Gift of Life registries. Similar to the full donor database by far the largest number of donors are from Caucasian descent. Asian and South American ethnicity's and registries have very high odds ratios. Especially the fact that donors from Southwest Asia are 7.09 times as likely to be CMV seropositive shows a very strong correlation. Based on the odds ratio for Fundacja DKMS we conclude that donors from Eastern Europe are more likely to be CMV seropositive.

Feature	Category	Count	Odds Ratio	95% CI
Sex	Female	4 967 892	1.33	±0.00
	Male	3 739 515	0.75	±0.00
Blood Type	O	3 399 450	0.96	±0.00
	A	3 275 912	0.92	±0.00
	B	1 027 445	1.23	±0.01
	AB	404 918	1.17	±0.01
Rhesus	Positive	6 727 078	1.09	±0.00
	Negative	1 314 923	0.92	±0.00
CCR5	Wildtype - homozygous	4 181 389	1.16	±0.01
	Deletion (delta 32) / wildtype - heterozygous	936 921	0.87	±0.00
	Deletion (delta 32) - homozygous	56 308	0.84	±0.01
EBV	Positive	47 875	1.85	±0.10
	Negative	7 302	0.54	±0.03
Registry	DKMS Donor Center (Germany)	3 659 740	0.96	±0.00
	NMDP (USA)	1 403 064	0.97	±0.00
	Fundacja DKMS (Poland)	**710 531**	**3.30**	**±0.02**
	DKMS United Kingdom (UK)	606 414	1.05	±0.01
	ZKRD (Germany)	590 414	0.79	±0.00
	NHS Blood and Transplant - BBMR (UK)	**260 203**	**0.37**	**±0.00**
	Anthony Nolan (UK)	244 302	0.63	±0.01
	Matchis Foundation (The Netherlands)	221 054	0.89	±0.01
	Fundación de Beneficencia Pública DKMS (Chile)	**173 525**	**4.16**	**±0.04**
	Gift of Life Marrow Registry (USA)	**97 976**	**0.45**	**±0.01**
Ethnicity	**Caucasian: Mainland Europe, Greenland, Iceland, Western Russia**	**4 651 240**	**0.47**	**±0.00**
	Unknown	**937 078**	**2.47**	**±0.01**
	Caucasian	869 004	0.69	±0.00
	Hispanic	181 218	1.05	±0.01
	Hispanic: South America	**176 969**	**4.09**	**±0.04**
	Mixed / Multiple	138 769	1.12	±0.01
	Asian: Southwest Asia (Middle East, Turkey)	**137 379**	**7.09**	**±0.10**
	Other	130 740	1.02	±0.01
	Asian: Southern Asia (India, Pakistan, Bangladesh, Sri Lanka, Bhutan, Nepal)	**126 536**	**3.23**	**±0.04**
	African	80 634	1.62	±0.02

Table 2.4: Odds ratios for the categorical features in the database. The categories are sorted by donor count. When there are more than 10 categories only the 10 most common are included in the table. Similar to the forest plot the significant odds ratios of above 2 and below 0.5 have been highlighted.

In table 2.5 we have listed all the continuous features that are available in the database and their Point-Biserial Correlation Coefficients (PBCC). ”Age” is the current age of the donor, ”Age at Test” is the age of the donor at the time of testing and ”Time since Test” is how much time has passed since the donor was tested (”Age at Test” = ”Age” - ”Time since Test”). ”Age at Test” is the feature that should be correlated with CMV serostatus, there is barely any difference with ”Age” and the PBCC for ”Time since Test” is around 0. Therefore, from now on, we will only use ”Age at Test” and this is what we are referring to whenever we use age. The correlation between age and CMV serostatus is a lot smaller than what we initially would expect. We will further investigate this relation in the Plots section.

The ”BMI” feature is calculated using ”Weight” and ”Height” whenever both are given. For neither ”Weight” nor ”BMI” there seems to be a strong correlation with CMV positivity. However, the PBCC for ”Height” is higher than what was expected. This negative correlation might be caused by the fact that the avarage age for Caucasian donors in the database is higher than for Asian donors, while CMV is a lot more common among donors of Asian ethnicity and less common among those of Caucasian descent.

Feature	PBCC	p-value
Age	0.0896	0.0
Age At Test	0.0956	0.0
Time Since Test	0.0104	5.5×10^{-205}
Weight	-0.0297	0.0
Height	-0.1058	0.0
Bmi	0.0305	0.0

Table 2.5: Point-Biserial Correlation Coefficient for the continuous features in the database. For values where the p-value is lower than the minimum float value available in python the value 0.0 was used.

Plots

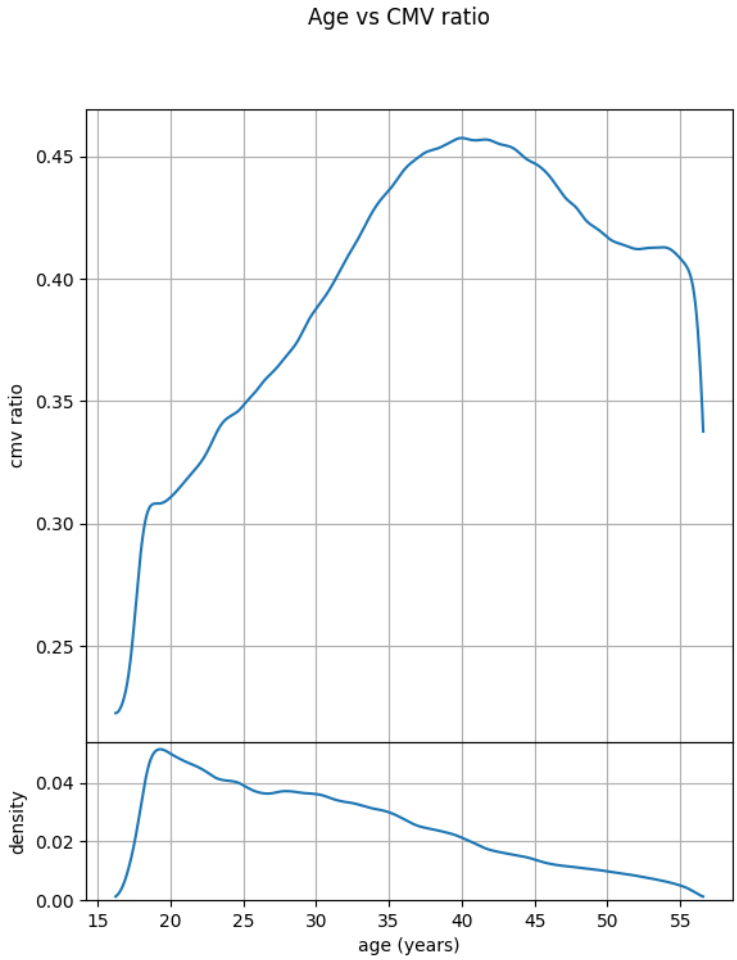


Figure 2.3: Plot of age vs CMV ratio.

In figure 2.3 we have plotted the age against the CMV serostatus. The distribution of the donors over the different ages is as we would expect. Most donor recruitment is done at younger age groups, which is why we see the initial peak and afterwards the steady decline. Up to age 40 the CMV ratio graph looks exactly as we would expect. As the donor age increases the likelihood that the donor is CMV positive increases. However, for donors above the age of 40 this relation does not seem to hold and the CMV seroprevalance decreases. Our first hypothesis for this effect is that there is a bias in the sampling of donors. That most of the donors over 40 are from a certain registry or ethnicity, with a

lower CMV positivity in the population and therefore causing this. We will therefore further investigate this effect by splitting the distribution by registry and ethnicity resulting in figure 2.4 and 2.5 respectively.

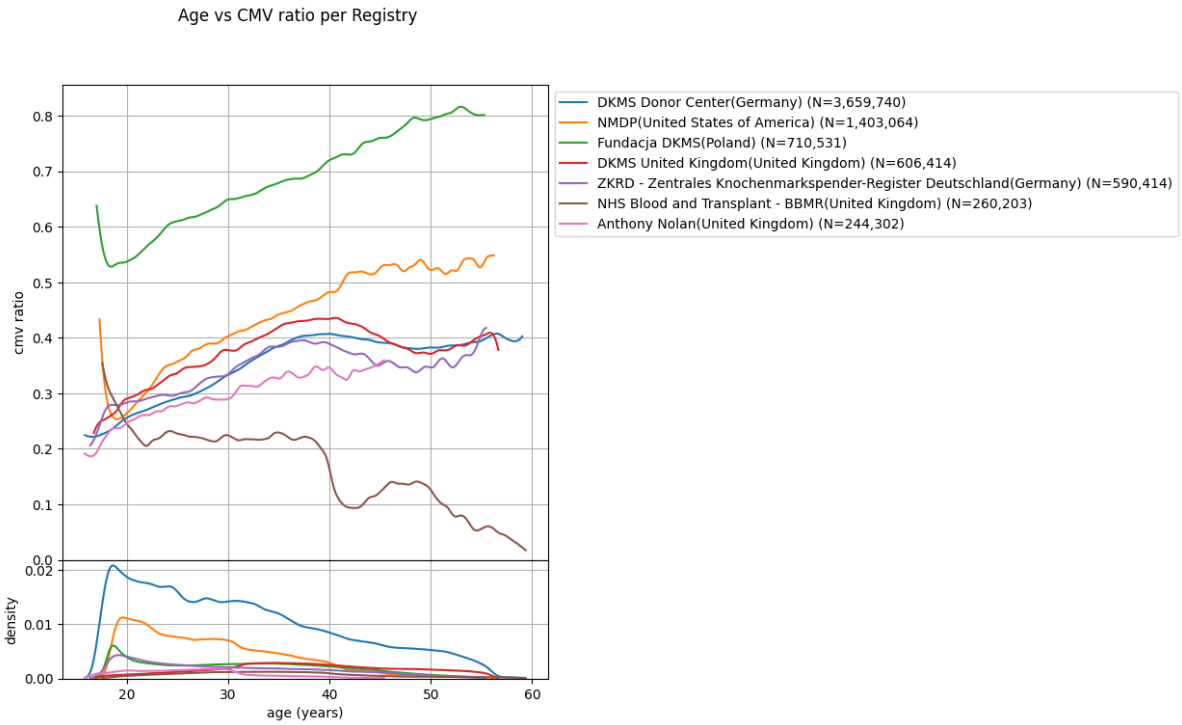


Figure 2.4: Plot of age vs CMV ratio split by registry. The 7 largest registries in the database were plotted.

The first thing to note in figure 2.4, is that almost half the donors with CMV information in the database are from a single registry, ”DKMS Donor Center”. This is one of the biggest donor registries that does a CMV test on a large percentage of their donors. Though there are definitely differences in CMV seroprevalence between registries, this does not seem to be the cause for the effect we found. Since, for some registries, like the NMDP, Fundacja DKMS and Anthony Nolan the CMV ratio curve looks like what we would expect; an ever increasing function of age. However, for other registries, like ”DKMS United Kingdom” and ”ZKRD”, we see the same effect where the CMV seroprevalence decreases for older donors. Therefore, we hypothesize that the decrease for certain registries for donors above 40 is caused by a sampling bias in the general population that is created by the donor recruitment policy of the registries. An example of a measure that could cause this effect is that some registries require older donors to pay to be listed as a donor. This causes donors of lower socioeconomic status to be less likely to be listed as an older donor and according to Cannon et al., 2010 persons of lower socioeconomic status are more likely to be CMV seropositive.

The NHS Blood and Transplant - BBMR is a special case where the CMV seroprevalence seems to decrease instead of increase when age increases. This effect may be caused by the fact that this organization mainly functions as a blood bank and not as a stem cell registry. Therefore, they may have very different donor selection or recruitment policies. We hypothesized that donor registries located in the same country would have similar distributions since they are sampling from the same population. However, in the figure, there is a large differences in the distribution for registries from the same country, for example the three registries from the United Kingdom. Therefore, we conclude that the donors registry is not equivalent to the geographic location of the donor.

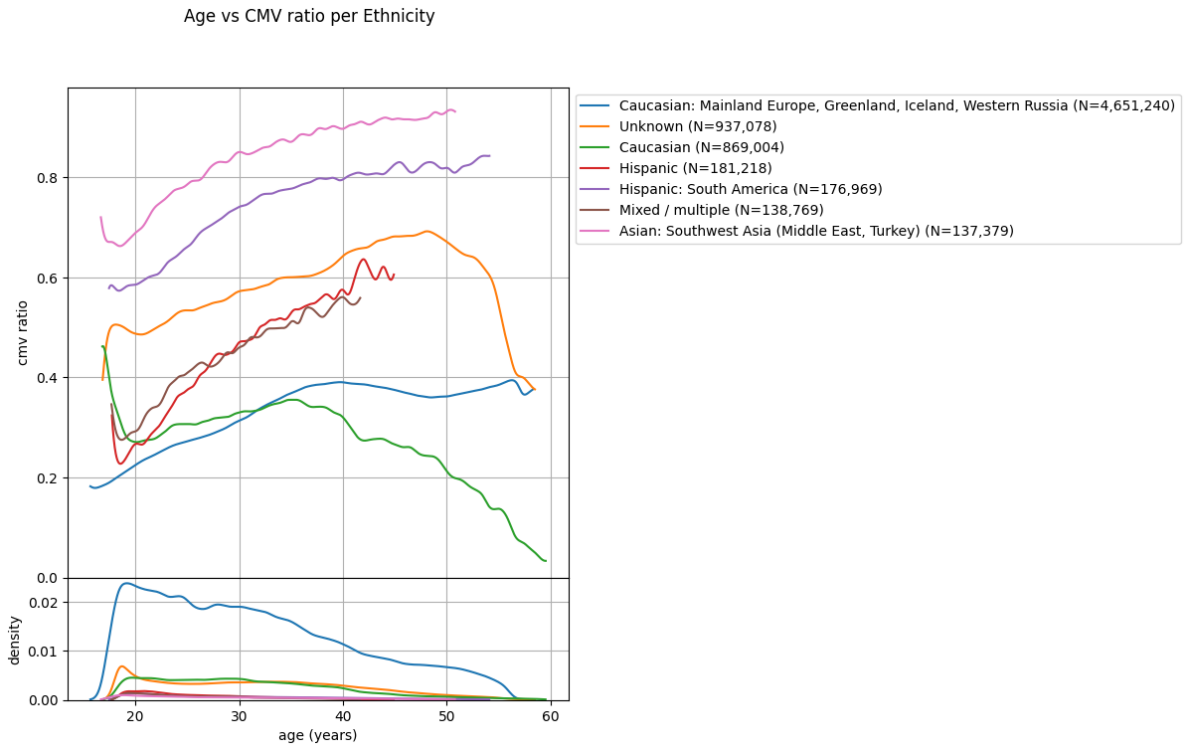


Figure 2.5: Plot of age vs CMV ratio split by ethnicity. The 7 ethnicities that were most represented in the database were plotted.

In figure 2.5, we have plotted the CMV distribution for the seven most common ethnicities. As for the registries, by far the most donors in the database are Caucasian. The differences between the different ethnicities are a lot larger than we expected. For the donors from ”Asian (Southwest Asia: Middle East, Turkey)” ethnicity the average CMV ratio is 80.8%, while for donors from ”Caucasian (Mainland Europe, Greenland, Iceland, Western Russia)” ethnicity the average CMV ratio is only 31.1%. Due to the large differences in the CMV distribution between ethnicities, we expect this feature to be useful during prediction. An interesting avenue for further research is to investigate how the registry and ethnicity of donors interact with eachother. For example is the distribution for Caucasian Donors from Asian registries similarly low.

Again some ethnicities, ”Unknown” and ”Caucasian (Mainland Europe, Greenland, Iceland, Western Russia)”, display the same behaviour that the percentage of CMV positive donors decreases for the oldest donors in the database. Since these donors make up most of the donors with a known CMV serostatus, they are probably the cause of the effect we saw in figure 2.3. Another exception is the ”Caucasian” donors, where we again see the CMV ratio decrease over time, this is most likely caused by 24.5% being part of the NHS registry that showed the same behavior.

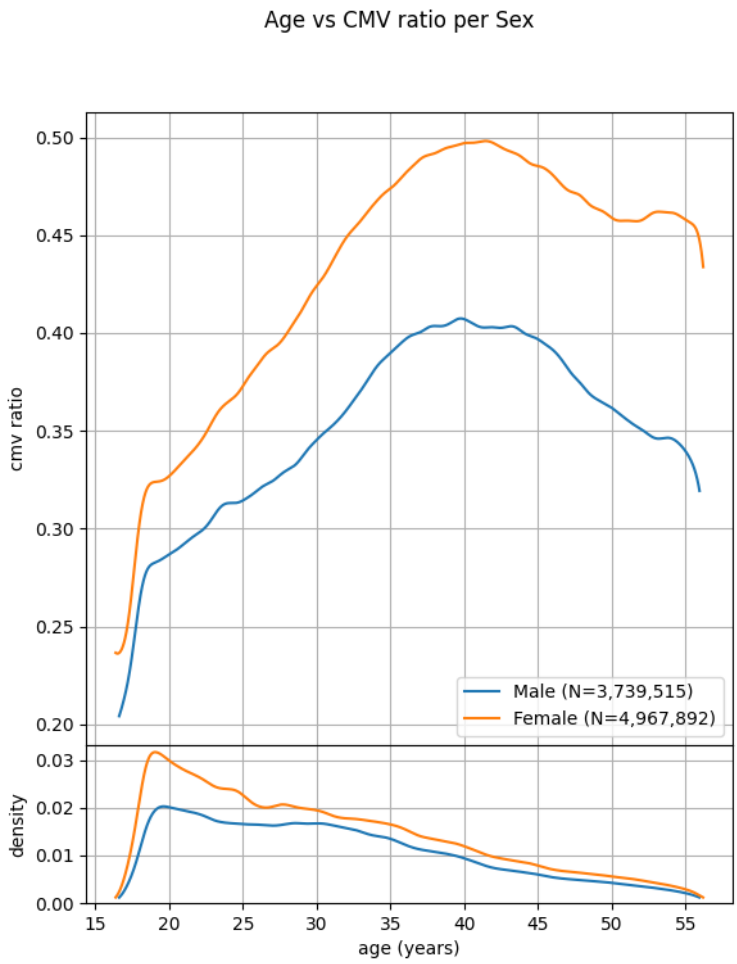


Figure 2.6: Plot of age vs CMV ratio split by sex.

In figure 2.6, we can see that the distributions divided by sex clearly show that female donors are more likely to be CMV positive. For the different ages, the two distributions behave similarly, however the difference in ratios between female and male donors seems to increase with age. The density distributions in the bottom again clearly show that there are significantly more female than male donors in the database.

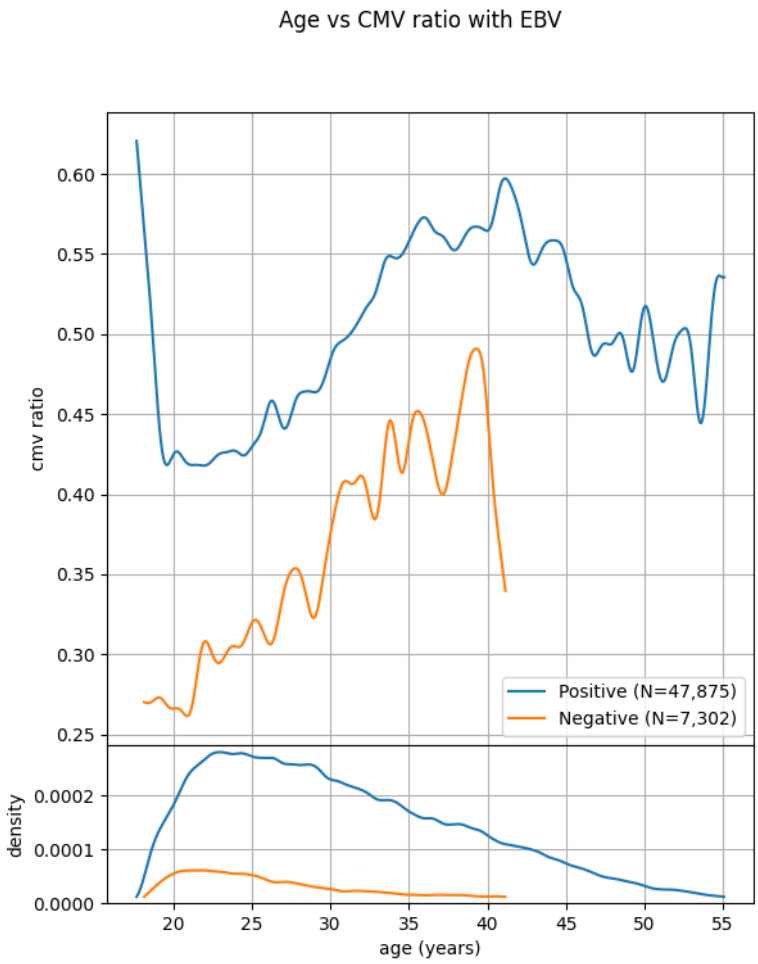


Figure 2.7: Plot of age vs CMV ratio split by EBV serostatus.

Unfortunately, there is not much overlap between donors with given EBV and CMV serostatus, which causes the distributions in figure 2.7 to be noisy compared to the others. There are also significantly fewer donors with a negative EBV serostatus for which the CMV serostatus is known. Despite this noise, there is a clear separation between the two distributions showing that donors with a positive EBV serostatus are more likely to be CMV seropositive, confirming the findings of Lazda, 2006.

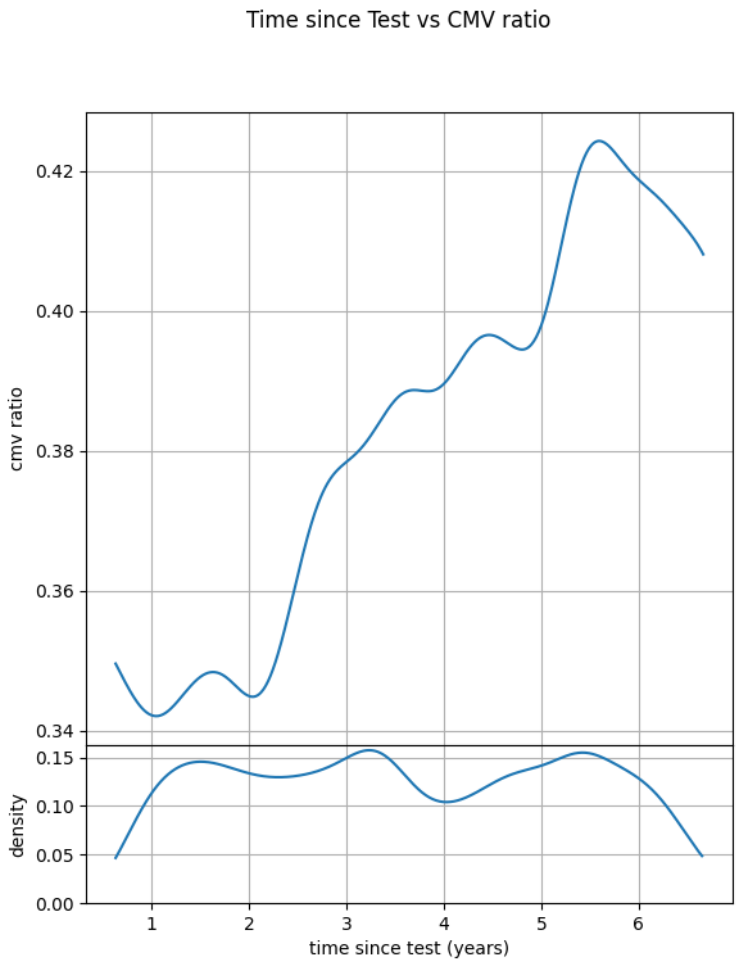


Figure 2.8: Plot of time since test vs CMV ratio. For this plot the density cutoff was increased to 400 000.

For each donor, the date of the CMV test is given in the database. With this we can see how the CMV serostatus in the population has changed over time by plotting the CMV distribution versus the time since test in figure 2.8. For this figure, we decided to increase the cutoff that we use to create the distribution to make sure that the distribution we see here is backed by a significant number of donors. In the figure, we see that the CMV seroprevalence is decreasing over the past 5 years. This change in CMV seroprevalence could be caused by other factors, like more recent donors coming from registries with high CMV seroprevalence. However, on this scale of number of donors that seems unlikely. It could be that the Corona outbreak and surrounding lockdown measures have decreased the spread of CMV over the past years and caused a lower seroprevalence for recently tested donors.

2.2.2. HLA groups

To continue the statistical analysis, we first analyse the effects of preprocessing the HLA features. After using the methods described in section 2.1.3, the number of different HLA values has decreased significantly for all loci when comparing between the original HLA values and the G-groups and P-groups. These changes in the total number of different values in the database can be found in table 2.6. The total number of different values has significantly decreased for all loci. As expected based on their definitions, the P-groups have slightly less distinct categories compared to the G-groups.

To further illustrate how the distribution of the donors over the different HLA descriptions has changed, a frequency plot for each locus has been created. We will look at the frequency plot for locus A, figure 2.9, our findings for the other loci are very similar. Their frequency plots can be found in section A.1 of the appendix. In the frequency plot for ”HLA before”, we see that a large number of donors falls

[H]

	Before	After	
Locus	HLA	G-groups	P-groups
A	40 666	2 453	1 873
B	68 891	2 994	2 294
C	32 020	2 385	1 696
DPA1	721	134	86
DPB1	4 380	753	634
DQA1	916	106	79
DQB1	7 953	755	523
DRB1	18 685	1 286	933
DRB3	2 058	185	157
DRB4	631	93	79
DRB5	585	79	74
Total	177 506	11 223	8 428

Table 2.6: Table with the number of distinct HLA values in the database before and after the HLA grouping step.

in the ”other” category and this category consists of a very large number of different HLA. This large number of categories that consist of a small number of donors make statistical analysis or machine learning difficult. Therefore, our main goal is to reduce the size of the ”other” category. For both G-group and P-group the number of donors in the ”other” category has decreased by almost a 100 times and the number of different HLA contained in the ”other” category has significantly decreased. When comparing the subplots before and after, it looks like the number of donors in certain categories has decreased, since the frequency plot decreases steeper for after. However, this is not the case. This effect is caused by the number of categories decreasing and because of that more smaller categories can be shown before they need to be combined into the ”other” category. This can be seen by looking at the category ”03:02:01G”, third from the right in the HLA before subplot and in the middle for the G-group subplot. The number of donors with this HLA has increased by around 50% from 4.07×10^{-3} before to 6.10×10^{-3} after. Finally the frequency distribution for both G-group and P-group are very similar. From the frequency plots we conclude that this step has been successful in making the data more homogenous.

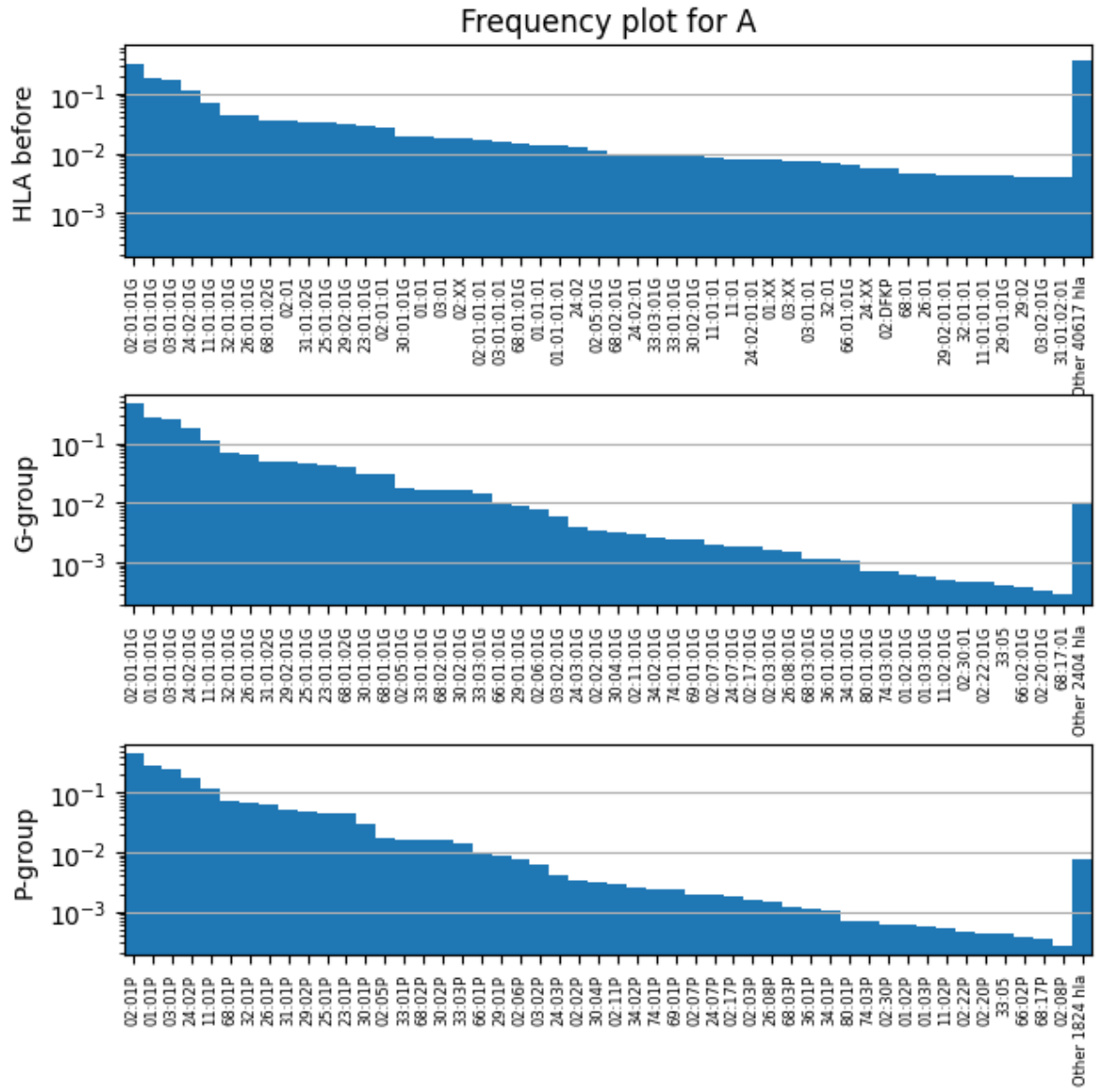


Figure 2.9: Frequency plot for the different HLA descriptions for locus A. The 50 largest HLA were plotted with the percentage of donors with that HLA on a logarithmic scale. All smaller HLA were grouped into a single bar in the far right of each subplot were the label tells you how many categories were summarized in this bar.

With the HLA compressed to less and larger categories, we can now use the G-groups and P-groups to create the volcano plots in figure 2.10 and 2.11 for locus A. For the other loci, the volcano plots can be found in the appendix A.2. For almost all loci, we find there are a number of groups that significantly increase the likelihood to be CMV seropositive. The exception is DQB1 for which there are no G-groups or P-groups in the significant regions. We expect that the increase in likelihood is caused by the fact that these results are at a global scale and that there are large geographic differences between CMV serostatus. By far the largest subset in the database consist of Caucasian donors, with a relatively low CMV seroprevalence. This means that HLA, that are more common for donors living in a region with higher CMV seroprevalance, would have a high odds ratio. We will further investigate this relation in the results of the meta analysis.

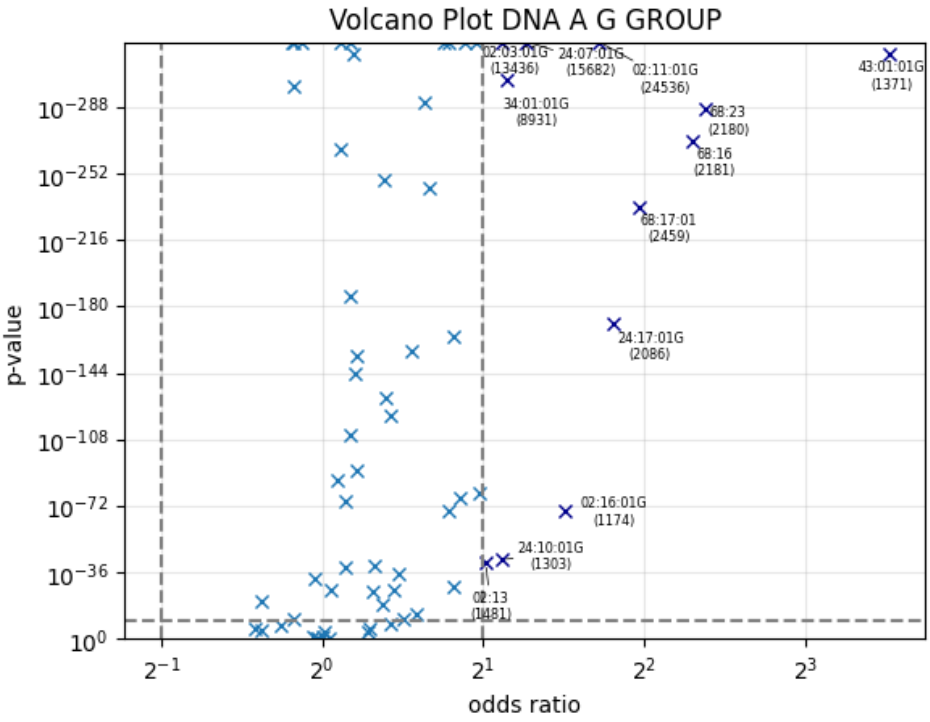


Figure 2.10: Volcano plot for the different HLA G-groups for locus A. For each HLA group in the significant regions the name of the group is visible in the plot followed by the number of donors with the given HLA and a CMV serostatus.

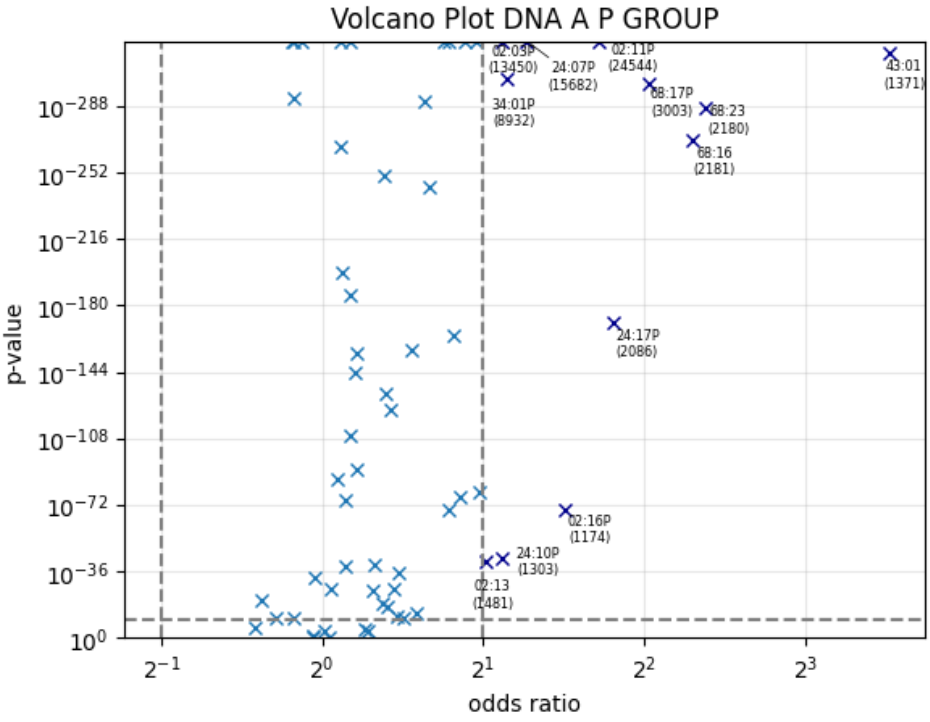


Figure 2.11: Volcano plot for the different HLA P-groups for locus A. For each HLA group in the significant regions the name of the group is visible in the plot followed by the number of donors with the given HLA and a CMV serostatus.

In both the Frequency plots and the Volcano plots we found that there are very few differences between the G-groups and P-groups. Since both HLA groups are so similar we will from now on use

only one of the two. We have decided to use P-groups from now on since it has a slightly smaller number of distinct categories. From now on when we refer to HLA group we are referring to the HLA P-group.

2.2.3. Meta Analysis

The goal of the meta analysis is to see if the increase in CMV positivity caused by certain HLA is consistent when compared between registries, because we want to verify that the HLA effect is not caused by the different registry populations the donors are sampled from. To properly do the meta analysis we only include registries with more than 50 donors with the specific HLA group and only do the meta analysis for HLA with 4 or more registries with enough donors. This results in a meta analysis and forest plot for more than 700 different HLA groups. Therefore, we have decided to only include the forest plots for the 3 groups with the largest Odds Ratio for each of the 5 most important loci (A, B, C, DQB1, DRB1) in the appendix. Out of those 15 plots, the plot for HLA group 40:06P on locus B was moved out of the appendix to further discuss the findings that can be seen in the forest plots. This plot was chosen since it contains a large number of donors spread over different registries around the world.

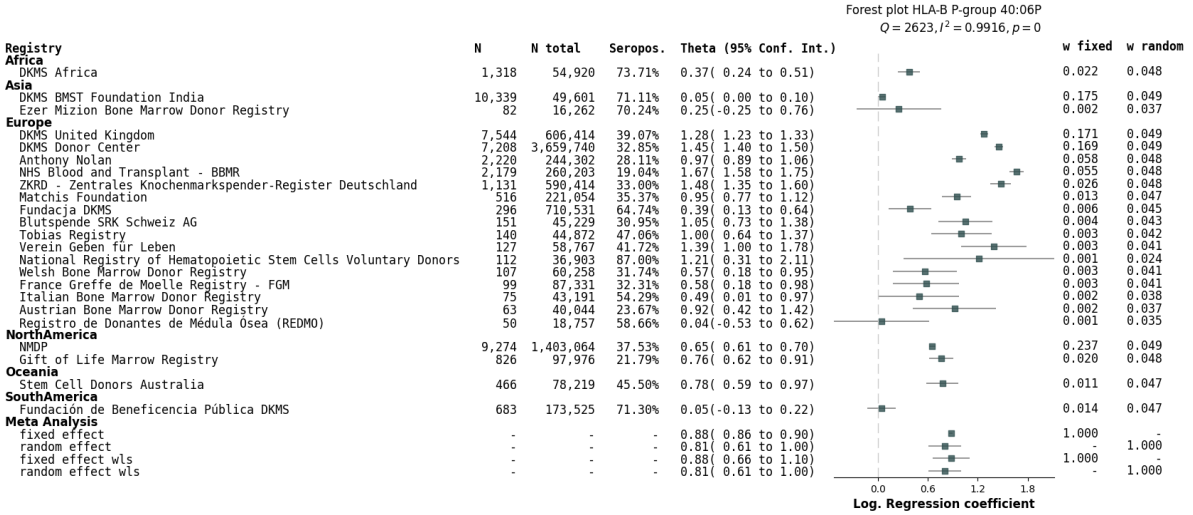


Figure 2.12: Forest plot for the meta analysis of 40:06P on locus B with $OR = 2.84$, $p = 4.94 \times 10^{-324}$ and $N = 45\,310$.

In the forest plot 2.12 an overview is given of the meta analysis for 40:06P on locus B. Any effect size above 0 shows that the presence of the HLA 40:06P caused an increase in the likelihood that the donor is CMV seropositive. In the plot we see that for all registries the effect size is above 0 and that there is a lot of heterogeneity between the different registries, since I^2 is close to 1. For some registries with a very high seropositivity, i.e. DKMS BMST Foundation India, the effect size is usually closer to 0 compared to registries with lower seropositivity in their population, i.e. DKMS United Kingdom. This is because for these registries with high seropositivity the logistic regression curve is already higher and therefore the presence of the HLA group can have less of an effect than for other registries. Finally, there are some registries where the confidence intervals for the found effect size are very high. As one would expect, this happens for registries with a small number of donors with the given HLA group.

In the bottom four rows of the forest plot the outcome of the meta analysis is shown. All four outcomes show that the presence of 40:06P on locus B significantly increases the likelihood that the donor is CMV seropositive. All of the found effect sizes also show a high confidence and for none of them 0 is even within the confidence interval.

All forest plots in the appendix A.3 show very similar findings for other HLA groups with high odds ratio as we found for 40:06P on locus B. For all of them the outcome of the meta analysis are positive and 0 is not within their confidence interval. Therefore, we conclude that there are certain HLA that cause donors to be more likely to be CMV positive and that this effect is visible over different registries. When predicting CMV serostatus we will attempt to include the HLA groups of the donor to improve our predictions.

2.3. Conclusion

In the statistics chapter, we have looked in depth at the different available features and how they relate to CMV seropositivity. A univariate analysis was done using odds ratios for categorical features and PBCC for numerical features. Strong relations with CMV were found for the categorical features, sex, EBV, registry and ethnicity, and the numerical features, Age and Height. These relations were investigated more in depth using the distribution plots.

When looking at the distribution of CMV serostatus versus age, it shows the expected steady increase for donors up to age 40. However, for donors older than 40, the percentage of CMV positive donors decreases. When splitting the distribution by registry and by ethnicity, this effect remained only for a couple of large registries, therefore, we concluded that this effect was likely caused by the donor selection criteria that those registries use. By splitting the plots by sex and EBV serostatus, the increased likelihood for female and EBV positive donors to be CMV positive was confirmed. Finally, we found that the percentage of CMV positive donors for recently tested donors, in 2023, has decreased when compared to 5 years ago.

By applying the methods described in section 2.1.3, we were able to significantly decrease the number of unique HLA values in the databases by changing all values in the database to either G-groups or P-groups. We then plotted these groups into volcano plots to see how they affected the CMV serostatus. In these plots, we found that the results for G-groups and P-groups are very similar, therefore, we decided to use P-groups from now on, since the P-groups contain less unique values. In the plots, several different HLA groups were found for all Loci except DQB1 that significantly increase the likelihood that the donor is CMV positive. During the meta analysis, it was investigated whether that effect still holds when the data is split over the different registries.

For each registry, a logistic regression coefficients was trained on age, sex and a specific HLA group. By doing a meta analysis on the coefficient for the HLA group over the different registries. We summarized the results of the meta analysis in a forest plot for each HLA group. We found a large heterogeneity between the registries and that the increase in likelihood to be CMV positive is consistent over the different registries. We conclude that there are certain HLA that are more likely to be CMV positive and that it is valuable to include the HLA of the donor when predicting the donors CMV serostatus.

3

Machine Learning

3.1. Methods

In this chapter, machine learning will be used to predict the CMV serostatus of a donor. Before the different classifiers are discussed, first the metrics and plots, which were used to evaluate and compare the models, are discussed. The first classifier, is a baseline using age, sex and registry of the donor with a logistic regression. This baseline will then be expanded by adding more features and including the HLA information using a two-hot-encoder. Next, our predictions will be improved using a more complex XGBoost model and a grid search to find optimal parameters. In the final subsection of the methods the scoring metrics are discussed. To implement these classifiers and evaluation we have used the Scikit-learn package in Python (Pedregosa et al., 2011). The code implementation can be found at https://git.lumc.nl/lhendriks/cmv_thesis.

3.1.1. Model evaluation

Our classifiers were evaluated using a number of different plots and metrics. To create these evaluation statistics our dataset was split into multiple training and test sets using stratified k-fold cross-validation with the commonly used 10 folds. Though, there is only a small imbalance between our test classes, we decided to use stratified cross-validation to make sure each fold contains an equal number of CMV positive and negative donors. Before creating the folds the database is first shuffled to make sure each fold contains a varied amount of donors from different registries, ethnicities and ages.

The classifiers are evaluated based on the Receiver-operating characteristic curve, henceforth referred to as the ROC curve, and the calibration curve. For each of these plots we plot the average over the 10 folds with a grayed area for the standard deviation.

The first and most important plot to evaluate the performance of a model is the ROC curve. In this plot the false positive rate and true positive rate are plotted against each other for different thresholds. For a perfect classifier there are only true positives. Therefore, the ROC curve will become a square with an Area Under Curve (AUC) of 1.0. When the classifier predicts randomly the ROC curve becomes a diagonal line with AUC of 0.5. Therefore the better the model is the more the ROC curve will be in the top left of the plot and the higher the AUC will be.

Therefore, the most important metric we will use to compare the classifiers is the AUC. This score represents the probability when given a random CMV positive and negative donor the classifier will predict the CMV positive donor to be more likely to be CMV positive than the CMV negative donor. We have decided to use the AUC score, since both false negatives and false positives are important to prevent for donor matching. The donor needs to be matched to the patient and both have negative medical consequences. Secondly, we decided to use AUC over F1-score because our two classes are relatively equal in size. The AUC metric also has the advantage that it is not dependent on the threshold that one uses to predict positive or negative CMV serostatus. Though AUC is the main metric we will use, we have also calculated the accuracy and F1-score for each model.

The second plot is a calibration curve that we use to identify how well our model is calibrated. In a calibration plot the predicted probability is plotted on the x-axis and the observed probability is plotted on the y-axis. The observed probability is the fraction of CMV positive donors in each bin on the x-axis.

For example, the observed probability would be 0.8 at the predicted probability 0.7 if there are 5 donors with a predicted probability around 0.7 and 4 of them are CMV positive. A simple way to make this plot is by separating the predicted probabilities into bins and then calculating the percentage of CMV positive donors in each bin. As we did in section 2.1.2 for the distribution plots, we have instead used a kernel density function to create a smoother plot. When the classifier is perfectly calibrated the calibration curve would follow a diagonal line, i.e. for donors with predicted probability 0.5 exactly 50% of them are CMV positive. Below the calibration curve the density of the predicted probabilities is plotted, again using a gaussian kernel density function, Silverman, 1986. This density curve serves two purposes. Firstly, we can use it to see how well the classifier can separate the positive and negative donors. In an ideal scenario, we would like the classifier to only predict 0 for all CMV negative donors and 1 for all CMV positive donors, resulting in two seperate steep hills on both sides of the density curve. However, in reality it is more difficult to separate the two classes and the predicted probabilities for positive and negative donors are overlapping and closer to 0.5. The second use is to see how many samples are used to create a certain section of the calibration curve. When the calibration curve deviates from the perfect diagonal on a part of the curve where there is a high number of predicted probabilities, this shows an issue in the calibration and design of the model.

3.1.2. Logistic Regression

For our baseline we performed a logistic regression based only on age, sex, and registry. To use the categorical features, except for HLA, a one-hot-encoding was used. For sex, the two columns are combined into a single column, since sex is given for all donors and there are only two categories in the database for sex "M" or "F". For columns with missing values a seperate column is added to indicate that the value is missing. One-hot-encoding can only encode values it has encountered in the training set. When a donor has a value that is not in the training set it will result in zeros for all columns in the one-hot-encoding.

In the distribution plots in section 2.2.1 we saw that there is not a simple linear relation between age and CMV serostatus. Therefore, we have added polynomial features for age up to a degree of 3. Finally, in the plot 2.6 we saw how strongly CMV is related to sex and how the distribution is slightly different between the two sexes. Therefore, we have added an interaction between sex and all the polynomial age columns. This results in the equation 3.1 for the logistic regression.

$$f(x_{age}, x_{sex}, \vec{x}_{reg}) = c^0 + c^1 x_{age} + c^2 x_{age}^2 + c^3 x_{age}^3 + c^4 x_{sex} + c^5 x_{sex} x_{age} + c^6 x_{sex} x_{age}^2 + c^7 x_{sex} x_{age}^3 + \vec{c}^T \cdot \vec{x}_{reg} \tag{3.1}$$

In this equation x_{age}, x_{sex} and $\vec{x}_{reg}$ are the given features for this baseline. The $1 \times k$ vector $\vec{x}_{reg}$, where k is the number of different registries in the training set, corresponds to the one-hot-encoded registry of the donor with a single one in the row corresponding to the donors registry. The logistic regression coefficients are c_0 to c_7 and $1 \times k$ vector $\vec{c}$ for a total of $k + 8$ training coefficients. This logistic regression is very similar to the one done for the meta analysis in section 2.1.4, but uses registry instead of a HLA group. This baseline Logistic Regression will later be improved by adding more features like all the HLA groups.

To optimize the coefficients of the logistic regression we used an Elastic Net (Zou and Hastie, 2005) loss function that uses both L1 and L2 regularization. There are two hyperparameters to control the regularization. Cs is used to control the total strength of regularization. $l1\_ratio$ is used to change the ratio between the L1 and L2 regularization. A $l1\_ratio$ of 1 is equivalent to only using L1 regularization and a $l1\_ratio$ of 0 corresponds to using only L2 regularization. We have optimized these hyperparameters using a traditional grid search (LaValle et al., 2004). The grid consists of 10 logarithmically scaled values for Cs between 10^{-4} and 10^4 and 6 linearly scaled values for $l1\_ratio$ between 0 and 1. To evaluate the different hyperparameters the training set is again split using stratified k-fold cross-validation, this time with 5 folds. The metric we use to identify the best hyperparameters is again the AUC for the same reasons that were discussed previously.

3.1.3. Two-Hot-Encoder

When we want to use our categorical HLA features in our model, we first need to encode them. The default option of using a one-hot-encoding comes with an issue when it comes to our HLA features. The HLA information consists of two values per locus, one for each chromosome of DNA. These two

values are interchangeable and arbitrarily one is put in the first column and the other in the second. If we would do a straight up one-hot-encoding for these two columns this property of the data would be lost. Since the encoding of HLA group $A\_1 = 02 : 01P$ and $A\_2 = 01 : 01P$ would be very different from the identical $A\_1 = 01 : 01P$ and $A\_2 = 02 : 01P$. To solve this problem the two-hot-encoder was implemented. Instead of encoding the two columns per locus separately, they are encoded at the same time.

The goal of the encoder is to create an encoding similar to one-hot-encoding which takes up to two values and puts up to two ones in each row in the columns matching these two values. When fitting the encoder it needs to see all possible categories in the training set, this is done by passing the two stacked columns as a single column to a regular one-hot-encoder. Then the fitted encoder is used on both columns separately resulting in two sparse matrices with the same dimensions with ones in the columns corresponding to the HLA values. By then taking the bitwise OR-operation of these two sparse matrices, the desired matrix is created. The bitwise OR was used instead of simply summing the two matrices, since we wanted a 1 and not a 2 value in the column in case both HLA values are the same.

However, there is one issue with this approach that needs to be fixed. Often, when the donor has the same value for both HLA values the value is given in A_1 and then A_2 is set to $Null$. In the current approach, this would result in two ones in the encoded matrix where we only want one. The extra one is in the column corresponding to the $Null$ value. To fix this issue, the entire column corresponding to the $Null$ value is removed from the final result. This results in unknown HLA (both values are $Null$) encoding to a row consisting of only zeros. This is our desired result, since the donor has none of the HLA values. This does mean that we need to be careful to not drop the first column when using the one-hot-encoding, which is sometimes done to reduce to total number of output columns by one.

Finally, to stop the number of feature columns from increasing exponentially and causing the training to become infeasible, a frequency cutoff was introduced in the encoding. When a HLA value is present for less than 1000 donors it is not given its own encoded column in the output matrix. However, for all these HLA values a separate column is created that indicates that a donor has an infrequent HLA.

3.1.4. XGBoost

After encoding the HLA data, the next goal is to improve the performance of the classifier by using a more complex model than Logistic Regression. We chose to use Extreme Gradient Boosting aka XGBoost (Chen and Guestrin, 2016), because of it has been optimized to train on large datasets and has shown good performance in recent studies. Furthermore, XGBoost also deals well with our missing and sparse data. XGBoost uses an ensemble of decision trees that are used to optimize the loss function which includes L1 and L2 regularization. Decision trees split the data based on a certain condition, like $age < 20$ or $A = 02 : 01P$. The data is then split again with a different condition and this repeats up to the maximum allowed depth of the decision tree. For each of the resulting leaves a score is given, which is positive if those conditions make the donor more likely to be CMV positive and negative if the donor is more likely to be CMV negative. For each tree in the ensemble the score is calculated for each data point and summed. This total score than indicates how likely the donor is to be CMV positive or negative. An example of two decision trees can be found in figure 3.1.

The ensemble of decision trees is build sequentially, starting with one tree and adding trees one by one until the maximum ensemble size of 10000 is reached or the early stopping condition is reached. When training XGBoost uses only 80% of the training set. The other 20% is used to evaluate the ensemble after adding each tree using the AUC as an evaluation metric. When the AUC has not improved in the last 100 added trees the early stopping condition has been reached and the best ensemble of trees is returned.

To find the best decision tree to add to the ensemble it would be intractable to loop over all possible trees and choose the best one. Instead XGBoost creates the new decision tree one split at a time by calculating the gain for each split in the data using the derivative of the used loss function. When the *max_depth* is reached or the gain is smaller than the added regularization and *min_split_loss*, no new split is made. The loss function includes both L1 and L2 regularization. The strength of these two regularizations are set by *reg_alpha* and *reg_lambda* respectively. When in none of the leaf nodes a new split can be made, the tree is added to the ensemble with the weights multiplied by the *learning_rate*. This is to make the boosting process more conservative and to prevent overfitting. All italicized parameters are the hyper parameters that will be optimized in a grid search.

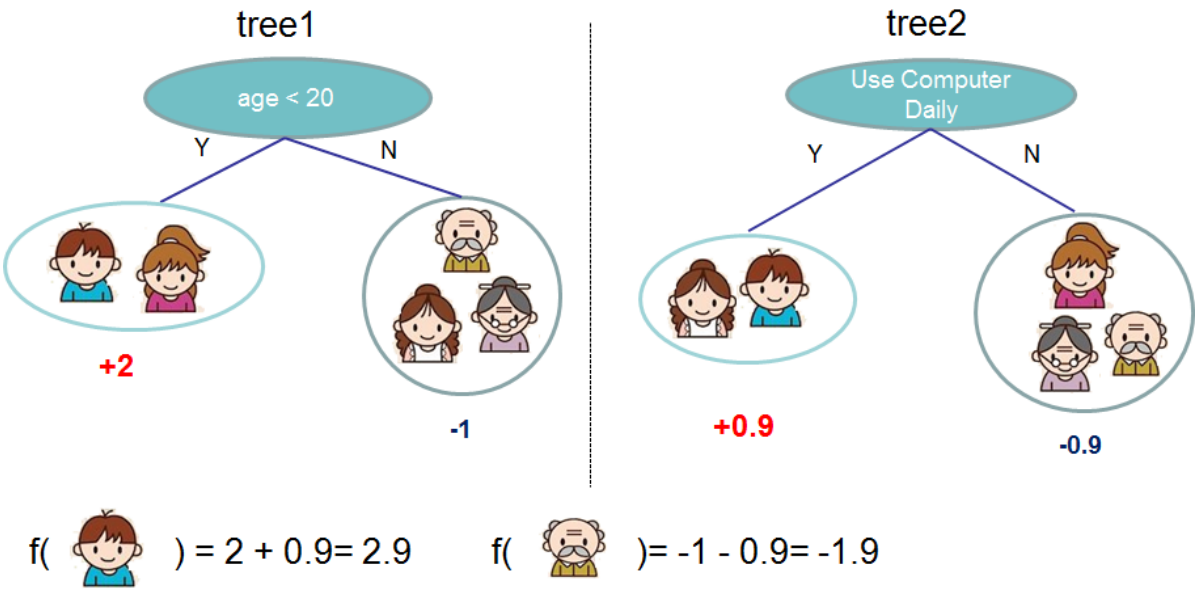


Figure 3.1: Example of an ensemble of two decision trees with depth 1. At the bottom it is shown how the final score is calculated for two data points. Source: Chen and Guestrin, 2016

Grid Search

To optimize the hyper parameters a grid search will be done. The parameters and their values have been listed in table 3.1. To implement the grid search on the cluster it was necessary to separate the code into a caller and segments. The caller will loop over the possible combinations of hyper parameters and check if the scores have been saved for that set of parameters. If they do not exist yet, the caller will call the segment code for that set of hyper parameters. The segment simply runs XGBoost with the given set of hyperparameters on a specific subset of the data and saves the scores in a file. The caller will continue starting segments until a maximum of 250 new segment runs have been started. The caller will then add itself back to the queue with a dependency on all the segment runs it has started and terminate itself. This way the caller will keep running until the scores have been saved for each set of hyper parameters. Then, instead of terminating, the caller finds the maximum of these scores and uses that set of parameters to train itself.

Hyper Parameter	Values
max_depth	$\{1, 2, 3\}$
min_split_loss	$\{0, 0.5, 1, 2, 5\}$
reg_alpha	$\{0, 0.1, 1, 10\}$
reg_lambda	$\{0, 0.1, 1, 10\}$
learning_rate	$\{0.1, 0.3\}$

Table 3.1: The XGBoost hyper parameters used in the grid search and their value ranges. In total there are 480 unique combinations of hyper parameters.

Just like any of the other classifiers we will verify the grid search by doing a 10 fold stratified cross validation. Furthermore, for each set of hyperparemeters in the grid search we do a 5 fold cross validation to evaluate the performance of that set of hyperparameters. Finally, to use early stopping in the grid search another eval set of 20% of that fold is used to evaluate the performance between rounds for the early stopping condition. This results in only $0.9 * 0.8 * 0.8 = 0.576$ of the full database being used for training during the grid search. However, this should not be an issue, since for our dataset that fraction consists of 5 015 466 donors and after the grid search is done and we have found the optimal set of hyper parameters, we can retrain the XGBoost on the full fold using $0.9 * 0.8 = 0.72$ of the database.

3.2. Results & Discussion

In this section the results of using machine learning to predict CMV serostatus will be discussed. First, a baseline is established using logistic regression on only age, sex and registry. Then the influence of adding the HLA data to this model will be investigated, followed by the results for the XGBoost and its optimal hyperparemeters found through the Grid Search. Finally, the effect of adding the other features in the database is looked at.

In table 3.2 the Accuracy, F1-score and AUC are listed for all four classifiers. Each classifier also has a reference name that was used in the code. These reference names will be used in plots legends to refer to the classifiers. For each of the scores the best performing classifier is the XGBoost with the grid trained hyper parameters. As explained in section 3.1.1, the main score that will be used to compare the classifiers is the AUC.

Classifier	Reference Name	Accuracy	F1-score	AUC
Logistic Regression Baseline	*logreg_baseline*	0.6636 ±0.0003	0.3502 ±0.0006	0.6516 ±0.0006
Logistic Regression with HLA	*logreg_with_hla*	0.6700 ±0.0004	0.4000 ±0.0010	0.6689 ±0.0007
XGBoost Default Hyper params	*xgboost*	0.6763 ±0.0002	0.4263 ±0.0005	0.6812 ±0.0005
XGBoost Gridsearch	*xgboost_parallel_grid*	**0.6765 ±0.0004**	**0.4270 ±0.0010**	**0.6817 ±0.0005**

Table 3.2: Scores for the four different classifiers that were trained. For each score the standard deviation between the folds is included. For each score the best performing classifier was made bolt.

In figure 3.2, the ROC curves for the four different classifiers are plotted. The baseline logistic regression achieved an AUC of 0.64, which could be improved by 0.02 by including HLA. This was further improved by using the more complex XGBoost model, but this showed only a small improvement of 0.01. This improvement is not due to randomness, since it is significantly larger than the standard deviation. By searching for optimal hyper parameters using the Grid Search we were not able to improve the AUC, though this same score was achieved using much smaller trees. All four show an improvement over the reference line, which corresponds to randomly guessing CMV serostatus for a donor. For all classifiers the AUC values are below the desirable range of around 0.8, however they do show non-trivial predictive ability.

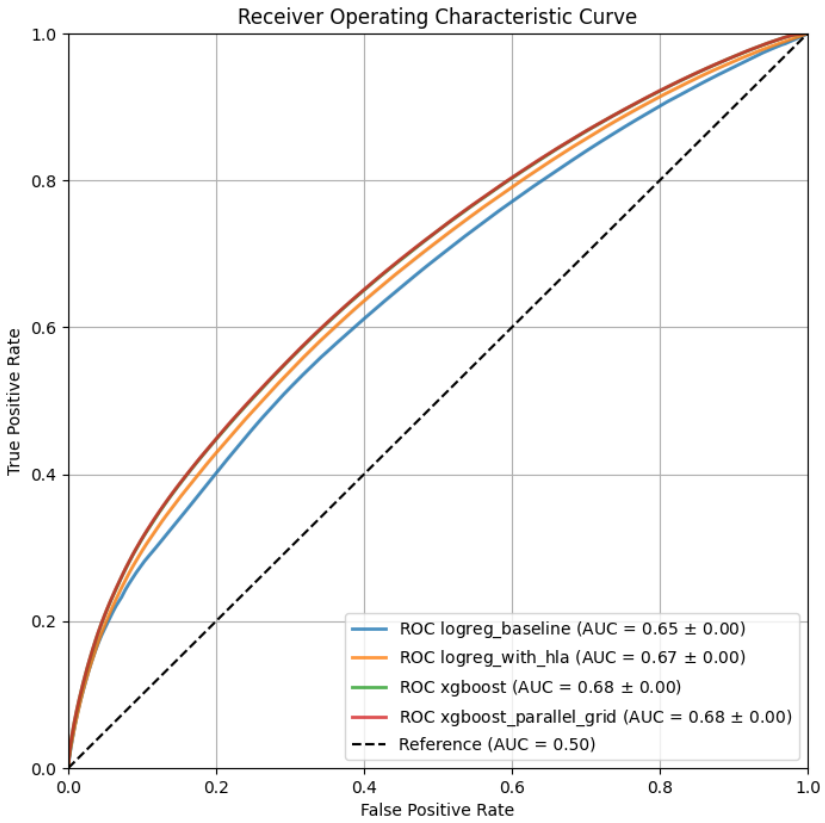


Figure 3.2: ROC curves for all four classifiers with a reference line. In the legend the values for the AUC score can be found with their standard deviation. The line for the default XGBoost is so close to the xgboost_parallel_grid that it can not be seen below it.

The Logistic Regression Baseline shows that a good prediction can be made for the CMV Serostatus using a simple classifier and only three features, age, sex and registry. In the calibration curve in figure 3.3 it can be seen that the curve often deviates from the diagonal. At the two ends of the curves this deviation is supported by only a small number of donors, as can be seen in the density curve in the plot below. However, in the areas where there is a high number of donors there are still visible deviations from the diagonal. These deviations indicate that the problem is too difficult to solve for the current classifier and that can either be improved by adding more features or using a more complex model.

The first improvement made was to include the donor HLA by adding the two-hot-encoded P-groups. In figure 3.3 it shows that the calibration improved by including the HLA features. The deviations from the diagonal have disappeared in the center areas supported by a large number of donors. In the two extremes, the calibration curve still strays away from the diagonal. However, this effect is less than for the baseline.

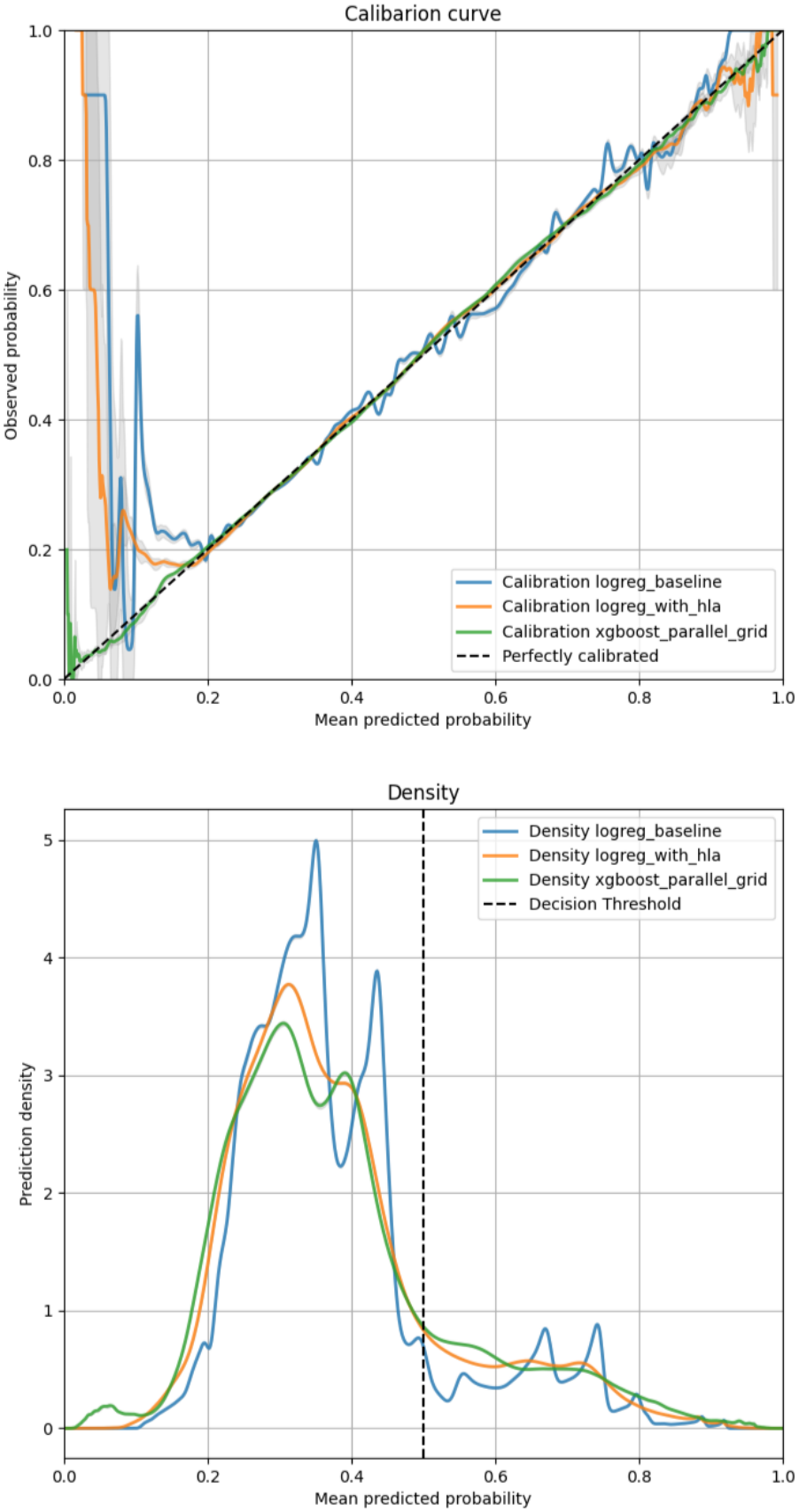


Figure 3.3: This figure shows the calibration curve for the trained classifiers. Below the calibration curves the density of the predictions made by the classifiers has been plotted.

By switching to XGBoost, a more complex model, the calibration is improved in the two extremes. Interesting to note is that there is now a significant group of donors for which the classifier can make a certain prediction that they are CMV negative. This can be seen in the new bump in the density graph for XGBoost at around 0.05. For more clarity, the curves for XGBoost with default hyper parameters have been left out of figure 3.3 since they were again very similar to XGBoost with the grid trained hyper parameters. Separate plots for each of the classifiers can be found in section A.4 of the appendix.

The density curves in figure 3.3 show an issue with the classifiers, which is worth investigating further. In optimal conditions, the density curve would consist of two peaks one around 0 for all the negative samples and one around 1 for all the positive samples. This means that the classifier is very sure about its predictions. Our classifiers show the opposite, barely any predictions are made in the two extremes and there are a lot of predictions that are made right around the threshold. These predictions are very unstable, since a small change will cause the prediction to switch from positive to negative or vice versa. There are three ways to improve the separability of our predictions. The first is to increase the complexity of the model, the second is to shift the threshold for classification to a different value and the third is to add more informative or discriminative features.

Firstly, in table 3.2 we can see that switching to a more complex model from Logistic Regression to XGBoost made very little improvement to the classifier performance. Furthermore, there was also little improvement by tuning the hyper parameters. This leads to the conclusion that increasing the complexity of the model will not help in further improvements to our classifier.

Then, shifting the threshold for classification was investigated. Figure 3.4 shows how the performance of the XGBoost classifier changes for different threshold values. Since the AUC does not depend on the threshold set by the classifier, it is a horizontal line. For the accuracy the maximum is found at the current threshold of 0.5. To verify the correctness of the plot, the plot shows that 38% of the donors are CMV positive, since at threshold 0 it shows an accuracy of 0.38. And, it shows that 62% of the donors are CMV negative, since at threshold 1 it shows an accuracy of 0.62. Adding up to the correct total of 100%. The F1 score can be improved by shifting the decision threshold to 0.3. This threshold improves the balance between precision and recall, to create a higher F1 score. However, the accuracy is significantly lower at this threshold. For donors, we want to correctly predict the CMV serostatus for as many donors as possible, which makes the accuracy more important than the F1 score. Therefore, we conclude that the classifier cannot be improved by shifting the threshold.

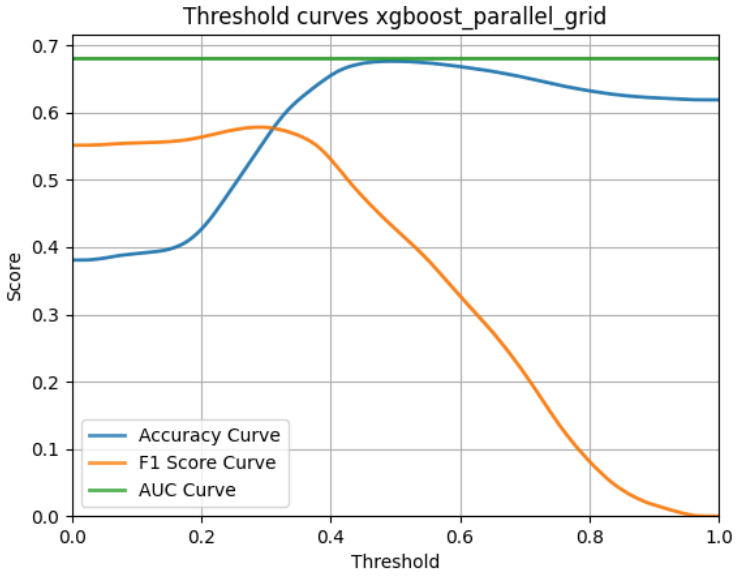


Figure 3.4: This figure shows how the XGBoost classifier with grid-trained hyper parameters performs with different threshold values using the accuracy, F1 and AUC scores.

Because neither increasing complexity nor shifting the threshold improves the classifier, we conclude that more informative and discriminative features need to be added to improve the classifier.

There are still a number of features in the donor database that could improve classification that have not been included yet, like ethnicity, blood type and height. To investigate how effective it would be to add these features, several classifiers were trained each with one of these features added. The AUC scores for these classifiers are listed in table 3.3. Each logistic regression and XGBoost classifier was trained on the baseline features, sex, age and registry, together with the listed feature. Categorical features were encoded using a one-hot-encoder including a column for missing features. For numerical features (height and weight), missing values were imputed using the mean of all values. For each of the available features, the performance of the classifier improved by adding it to the training set. In general, the improvements were slightly larger for the XGBoost model than for the Logistic Regression model. For EBV serostatus and CCR5, the increase in AUC was so small that it falls within the Standard Deviation (SD). This is likely due to that this information is known for only a small part of the donors in the database, especially for EBV serostatus. Including blood group, rhesus and weight, showed small improvements to prediction, though they barely showed any correlation to CMV serostatus in the statistical analysis. Height caused a bigger change to the AUC than weight, likely due to the higher PBCC that was found in section 2.2.1. In addition to the expected great improvement from the HLA information, the biggest improvement came from adding ethnicity to the training data. Likely because the ethnicity helps to identify outliers within registries. Though this should be investigated further.

Finally, both models were trained using all the available data combined. This resulted in the best performing classifier in our work with an AUC of 0.70. Though this classifier did take around 5 times the training time, the increases in performance show that big improvements to CMV prediction can be made through adding more features.

Feature		Logistic Regression			XGBoost		
	Missing	AUC	SD	Change	AUC	SD	Change
Baseline	-	0.6516	±0.0006	-	0.6551	±0.0005	-
HLA	7.70%	0.6689	±0.0007	+0.0173	0.6812	±0.0005	+0.0261
Blood group	6.89%	0.6527	±0.0004	+0.0011	0.6580	±0.0006	+0.0029
Rhesus	7.64%	0.6520	±0.0003	+0.0004	0.6573	±0.0005	+0.0022
CCR5	40.57%	0.6521	±0.0008	+0.0005	0.6558	±0.0005	+0.0007
EBV serostatus	99.37%	0.6517	±0.0006	+0.0001	0.6553	±0.0005	+0.0002
Ethnicity	11.83%	0.6821	±0.0003	+0.0305	0.6923	±0.0005	+0.0372
Height	27.92%	0.6534	±0.0006	+0.0018	0.6620	±0.0007	+0.0069
Weight	21.16%	0.6519	±0.0005	+0.0003	0.6590	±0.0009	+0.0039
All combined	-	**0.6867**	**±0.0004**	**+0.0351**	**0.7023**	**±0.0007**	**+0.0472**

Table 3.3: AUC scores for the classifiers with a different number of features added to the training data. For the baseline the training data only consists of age, sex and registry. In the final row the scores are listed when training on all the available features is combined. The rows in between show the AUC scores when individually adding the feature to the baseline. For each classifier the AUC is given with the SD and the change compared to the baseline. For each feature the percentage of missing values has been added.

Besides adding more features, the encoding of the HLA features could be improved by using an embedding that shows the relations between the different HLA instead of the current two-hot-encoding. This will allow the classifier to better use the HLA since the classifier is currently not aware of the similarities between many HLA alleles. Finally, the classification could be improved by looking for features outside the donor database that have been shown, Cannon et al., 2010, to have predictive power for CMV serostatus, such as socioeconomic status and level of education.

When trying to train XGBoost with default settings, the training would take an extremely long amount of time on the cluster. The longest run that was attempted ran for 10 consecutive days without being able to complete a single of the 10 folds in the cross validation. Interestingly the log shows that the CPU was barely used during this time, indicating that the program might be hanging or waiting on data in- or output. However, when running the code locally with higher verbosity, the learning was not hanging but training very slowly while barely using the CPU. The issue seems to be related to the large number of sparse features caused by including the HLA information in the input. When training without the HLA included the training would be finished within a minute. In the end, after investigating thoroughly, we could not identify the exact cause for the extremely long training time. However, when training on the GPU instead of the CPU the training would complete within 2 minutes per fold. It is unclear why the

switch from CPU to GPU is making such a large difference in the training time. However, this training time is fast enough to allow optimizing the hyper parameters by doing a parallel grid search.

3.3. Conclusion

In the machine learning chapter, an investigation was conducted to predict CMV serostatus using classification. An initial baseline was established using logistic regression on age, sex, and registry with an AUC of 0.65. Small improvements were made on this baseline by adding the HLA P-groups to achieve an AUC of 0.67. Switching to the XGBoost model and grid training the hyperparameters resulted in an AUC of 0.68. Although these absolute AUC values are not high, they consistently exceed random prediction, indicating that the classifiers capture a meaningful signal in the data.

An analysis of the calibration and density curve shows that the classifiers are currently unable to create separation between the positive and negative donors. Increasing the complexity of the classifier by switching from Logistic Regression to XGBoost only made a minor improvement of 0.01 to the AUC. Furthermore, moving the decision threshold did not improve the accuracy. Therefore, the training data needs to be enhanced for better prediction. An initial investigation shows that especially including ethnicity improves classification. Training on all features in the database resulted in the best performance of 0.70. More improvements could be made by better embedding of the HLA features.

4

Conclusion

In this study, both statistical and machine learning analyses were conducted to investigate prediction of CMV serostatus in stem cell donors. The statistical analysis revealed several features with strong associations to CMV seropositivity. Categorical features such as sex, EBV status, registry, and ethnicity, along with numerical features like age and height, were found to be informative. In-depth exploration of the age distribution highlighted an unexpected decrease in CMV positivity for donors over 40 years, which we suspect are due to registry-specific selection criteria. Additionally higher likelihood to be CMV positive for Female and EBV positive donors were confirmed from literature.

Through the use of G-groups and P-groups the number of unique alleles was significantly reduced. P-groups were selected for further use due to comparable results and less unique values. Several HLA groups, across all loci except DQB1, showed significant associations with CMV seropositivity. These findings were confirmed through a meta-analysis across registries, which demonstrated consistent predictive value of specific HLA groups. These findings support the inclusion of HLA information in machine learning for predicting CMV serostatus.

In the machine learning analysis, classification models were trained to predict CMV serostatus. A baseline logistic regression using age, sex, and registry yielded an AUC of 0.65. Adding HLA P-group features and using XGBoost led to only modest improvements. The best performing model, with an AUC of 0.70, is XGBoost trained on all features in the database. Although these AUC values are not high in absolute terms they do show non-trivial predictive power.

Increasing the complexity of the classifier and tuning the hyper parameters only lead to a minor improvement of 0.01 to the AUC. Adding more features to the training data, especially ethnicity, helped to further separate CMV positive from negative donors. Further improvements are discussed in section 5.

5

Limitations and Further Research

In this chapter, limitations with regards to this research and angles for future research will be discussed. These are ideas that emerged while working on this thesis but fell outside its scope. Or, options to further improve the performance of the classifier that were found while analyzing the results.

One of the limitations is that we were not able to find an explanation for the drop in CMV seroprevalence for donors above 40. Our hypothesis is that this effect is caused by differences in donor selection criteria between the registries. This hypothesis should be further investigated by diving deeper into the data. Interviews with experts from the affected registries could help to clarify this hypothesis or find another underlying reason for this effect.

To begin with, the features available in the donor database should be explored more thoroughly. The initial four classifiers, which were investigated in depth, mainly focused on age, sex, registry and HLA. These were chosen based on the statistical analysis. In table 3.3 we found that adding more features, especially ethnicity, to the XGBoost classifier resulted in the best performance. For this classifier simply all features available in the database were combined. It might be valuable to further investigate whether a specific subset of features could lead to a better final performance. For all these features the effect they have on the classification has only been evaluated through their AUC score. These classifiers should be further analyzed to gain more insight into which features are valuabe, for example by looking at their ROC curve and calibration curve.

One subset of interest is the combination of registry and ethnicity, since there is a lot of overlap between them. Therefore, it is likely that the combination of these two features helps the classifier to identify certain outliers. These interactions between registry and ethnicity should be further investigated in a statistical analysis. Furthermore, to use these interactions in the logistic regression, it is necessary to add interactions columns between ethnicity and registry in the preprocessing of the data. The XGBoost classifier can use this interaction inherently as long as the *max_depth* of the trees is larger than 1.

Besides ethnicity, height shows the biggest gain in performance. Though the high PBCC found for this feature is likely due to confounding. Considering that height is available for a large number of donors it is still valuable to include it for training.

In literature, Lazda, 2006 and our statistical analysis a clear positive correlation between CMV and EBV serostatus was found. However, including this feature in the training resulted in a very small change in performance. Though this unexpected finding is likely due to the small number of donors for whom both EBV and CMV serostatus is known. This should be confirmed by further analysis of the lower than expected improvement.

The current embedding of the HLA introduces a large bias towards western countries. Ambiguous alleles are mapped to their first subtype, which is the first discovered subtype. This makes it more likely to be a western subtype, since allele testing was done first for western donors, Bull World Health Organ, 1968. It would be interesting to investigate further if this had effects on our statistical analysis and classifiers for predicting CMV serostatus. The effect could be analysed by separately scoring the classifier on the different ethnicities in the test set and checking for major differences between the performance for different ethnicities. This limitation could be solved by embedding directly on the HLA alleles.

There are a number of other limitations to the current embedding of the HLA features using the two hot encoder and G-groups and P-groups. Some fine details of the HLA are lost by reducing the HLA alleles to just their groups. Furthermore, there are still a large number of unique HLA groups and each of the groups needs a single column in the two-hot-encoder. This large and sparse training data results in slow training. The encoding also does not capture the similarities between the different HLA alleles. Finally, the encoding of the ”N”-suffix for the ”Null” alleles does not explicitly show that these alleles are not expressed.

The first option is to encode the HLA using a frequency based embedding on the protein sequence of each HLA allele. This method has the advantage that it reduces the number of columns to just the number of different proteins and shows similarities between similar HLA with similar protein sequences. The downside of this method is that it loses the specificity of individual alleles. Two HLA alleles with the same number of total proteins in very different orders would result in the same embedding though they behave very differently.

Embedding the k-mers or the full HLA sequence using either Word2Vec or another transformer model, would capture the individual alleles and show similarities between HLA. The downside of this embedding is that it would take a lot of resources to implement and tune properly. However, it is uncertain whether such a representation captures variance relevant for predicting differences in immunological endpoints, such as CMV serostatus. Conversely, embeddings could be made on the basis of predicted binding capacities by HLA's. For instance, NetMHCpan (Reynisson et al., 2020), is a model that predicts peptide binding probabilities for HLA molecules using artificial neural networks. Using this model, we could embed the HLA directly on its predicted binding probability of a list of peptides for known CMV antigens. This would likely result in the best embedding maintaining individuality of alleles while also showing similarities between similar alleles.

Besides improving the performance of the classifier, this embedding could be used for another research spin-off. This embedding allows us to further analyze the HLA alleles that were identified to strongly correlated to CMV serostatus. An analysis into the similarities between these alleles and differences with other alleles, could help to give more insight into why these alleles positively correlate to CMV serostatus.

The final available feature that we have not discussed yet are the killer-cell immunoglobulin-like receptors (KIRs), Parham, 2005. The KIRs play an important role in the regulation of the natural killer cells and therefore likely have an effect on CMV infection and serostatus. Similar to HLA there are a large number of different KIRs and we would need to make an embedding to be able to use them in our machine learning pipeline. With this KIR data there is more opportunity for research. Besides CMV serostatus, there might also be biological interactions between KIRs, CCR5, EBV and Human Immunodeficiency Virus (HIV).

In this research, we were limited to using only data from the WMDA's global donor database. This data is highly anonymised and contains only data that is relevant for AHSCT. As seen in this thesis, the data set contains some data relevant to CMV serostatus. However, there are a number of other features that are not available to the WMDA, but are available to the individual registries. Examples of these features are the socio-economic status of the donor and whether the donor lives in a densely populated area through the exact address of the donor. To protect the privacy of the medical information of the donors, this data would have to be treated very carefully. Obtaining this data poses significant challenges as registries maintain strict confidentiality regarding their donor information.

In addition to the registries, the EBMT has an extensive collection of donor data for only European donors. This data also includes a more in-depth history of the donors CMV tests. Unlike the WMDA database, which just includes the result and date of the last test. This history would allow us to further validate the classifier by taking donors that were negative and recently tested positive to see if our classifier predicts them as being more likely to be CMV positive.

In the machine learning section of our research Logistic Regression and XGBoost were used. It would be interesting to investigate other models and see if they perform better on this dataset, like models based on artificial neural networks. The XGBoost model that we used could be further specialized for this problem. Due to time constraints, the best XGBoost model using all combined features uses default hyper parameters. These hyper paramaters should be optimized similar to what was done in this work, though it will likely again result in only a marginal increase in performance. Another option would be to look into training some of the hyper parameters that were left as default in the current grid search. For XGBoost it is possible to calculate SHapley Additive exPlanations (SHAP) values,

Lundberg and Lee, 2017, which show how strongly certain features influence the probability to be CMV positive. It would be interesting to calculate these values, analyze them and see if they agree with our findings in chapter 2. Furthermore, the reason for the slow training on the CPU and fast training on the GPU was never found, which should be investigated further.

Finally, it would be interesting to investigate further validation of the classifier. To ascertain that there is no bias towards specific countries. It would be possible to do a cross validation over the different registries in our dataset instead of regular cross validation. In this cross validation, the donors for one registry would be left out as a test set and the donors for all other registries would be used for training. This cross validation would then show how well our classifier generalizes between donors from different registries.

Bibliography

Anthony Nolan Research Institute. (2024). *HLA alleles*. Retrieved July 18, 2024, from https://hla.alleles.org

Bull World Health Organ. (1968). Nomenclature for factors of the HL-a system. *Bulletin of the World Health Organization*, *39*(3), 483–486.

Cannon, M. J., Schmid, D. S., & Hyde, T. B. (2010). Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. *Reviews in Medical Virology*, *20*(4), 202–213. https://doi.org/10.1002/rmv.655

Carreras, E., Dufour, C., Mohty, M., & Kröger, N. (Eds.). (2019). *The ebmt handbook: Hematopoietic stem cell transplantation and cellular therapies*. Springer Cham. https://doi.org/10.1007/978-3-030-02278-5

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, *abs/1603.02754*. http://arxiv.org/abs/1603.02754

Cochran, W. G. (1954). Some methods for strengthening the common χ2 tests. *Biometrics*, *10*(4), 417–451. Retrieved August 14, 2024, from http://www.jstor.org/stable/3001616

Harrer, M., Cuijpers, P., A, F. T., & Ebert, D. D. (2021). *Doing meta-analysis with R: A hands-on guide* (1st). Chapman & Hall/CRC Press.

Hassan, J., O'Neill, D., Honari, B., Gascun, C. D., Connell, J., Keogan, M., & Hickey, D. (2016). Cytomegalovirus infection in ireland: Seroprevalence, hla class i alleles, and implications. *Medicine*, *95*(6), e2735.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/https://doi.org/10.1002/sim.1186

LaValle, S. M., Branicky, M. S., & Lindemann, S. R. (2004). On the relationship between classical grid search and probabilistic roadmaps. *The International Journal of Robotics Research*, *23*(7-8), 673–692.

Lazda, V. (2006). Evaluation of epstein-barr virus (ebv) antibody screening of organ donors for allocation of organs to ebv serostatus matched recipients. *Transplantation Proceedings*, *38*(10), 3404–3405. https://doi.org/https://doi.org/10.1016/j.transproceed.2006.10.066

Lev, J. (1949). The Point Biserial Coefficient of Correlation. *The Annals of Mathematical Statistics*, *20*(1), 125–126. https://doi.org/10.1214/aoms/1177730103

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. *CoRR*, *abs/1705.07874*. http://arxiv.org/abs/1705.07874

Machida, S., Takahashi, T., Nomoto, K., & Murata, S. (1998). Hla antigens are associated in seronegativity to cytomegalovirus(cmv). *Nihon rinsho. Japanese journal of clinical medicine*, *56*(1), 129–133.

Marsh, S. G. E. (2024a). *HLA G Groups*. Retrieved July 18, 2024, from http://hla.alleles.org/wmda/hla_nom_g.txt

Marsh, S. G. E. (2024b). *HLA P Groups*. Retrieved July 18, 2024, from http://hla.alleles.org/wmda/hla_nom_p.txt

Miendje Deyi, Y., Goubau, P., & Bodéus, M. (2000). False-positive igm antibody tests for cytomegalovirus in patients with acute epstein-barr virus infection. *European Journal of Clinical Microbiology amp; Infectious Diseases*, *19*(7), 557–560. https://doi.org/10.1007/s100960000317

NMPD. (2024). *HLA MAC codes*. Retrieved July 22, 2024, from https://hml.nmdp.org/mac/files/alpha.v3.zip

Parham, P. (2005). Mhc class i molecules and kirs in human history, health and survival. *Nature reviews. Immunology*, *5*, 201–14. https://doi.org/10.1038/nri1570

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., & Nielsen, M. (2020). Netmhcpan-4.1 and netmhciipan-4.0: Improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic Acids Research*, *48*(W1), W449–W454. https://doi.org/10.1093/nar/gkaa379

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). Chapman; Hall.

Staras, S. A. S., Dollard, S. C., Radford, K. W., Flanders, W. D., Pass, R. F., & Cannon, M. J. (2006). Seroprevalence of Cytomegalovirus Infection in the United States, 1988–1994. *Clinical Infectious Diseases*, *43*(9), 1143–1151. https://doi.org/10.1086/508173

World Marrow Donor Association. (2023). *Donor database statistics*. Retrieved November 3, 2023, from https://statistics.wmda.info/density/

World Marrow Donor Association. (2025). *Data manager - search match service data submission information*. Retrieved June 3, 2025, from https://share.wmda.info/pages/viewpage.action?pageId=210665521

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

A

Further plots

A.1. Frequency Plots per Loci

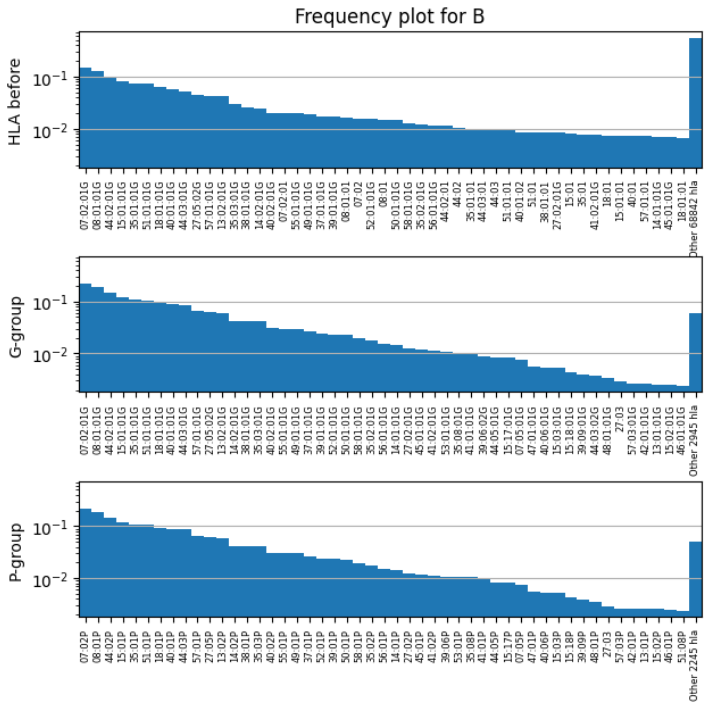


Figure A.1: Frequency plot for the different HLA descriptions for locus B. The 50 largest HLA were plotted with the percentage of donors with that HLA on a logarithmic scale. All smaller HLA were grouped into a single bar in the far right of each subplot were the label tells you how many categories were summarized in this bar.

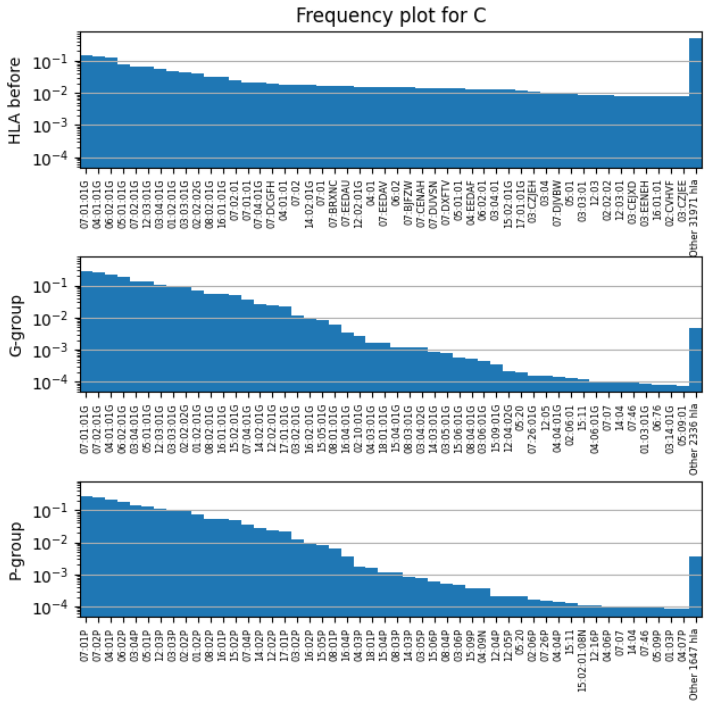


Figure A.2: Frequency plot for the different HLA descriptions for locus C. The 50 largest HLA were plotted with the percentage of donors with that HLA on a logarithmic scale. All smaller HLA were grouped into a single bar in the far right of each subplot were the label tells you how many categories were summarized in this bar.

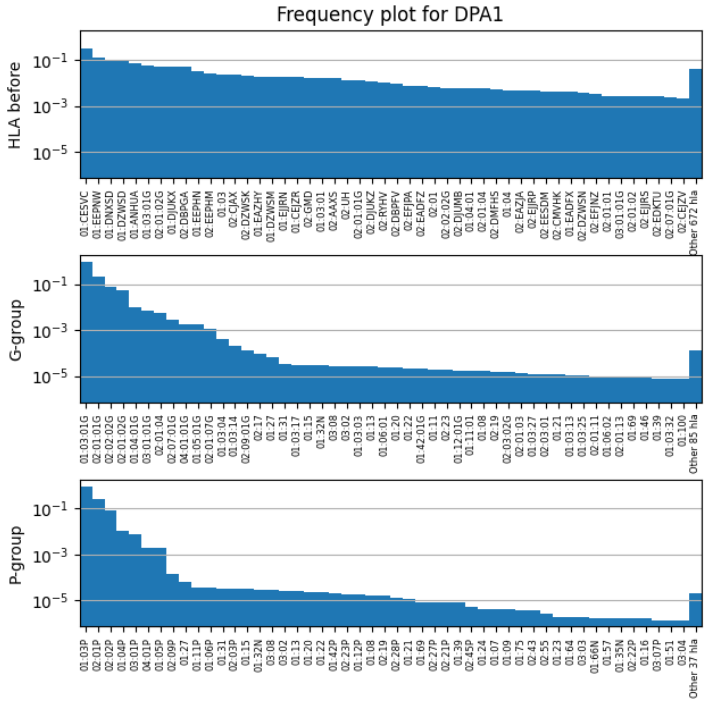


Figure A.3: Frequency plot for the different HLA descriptions for locus DPA1. The 50 largest HLA were plotted with the percentage of donors with that HLA on a logarithmic scale. All smaller HLA were grouped into a single bar in the far right of each subplot were the label tells you how many categories were summarized in this bar.

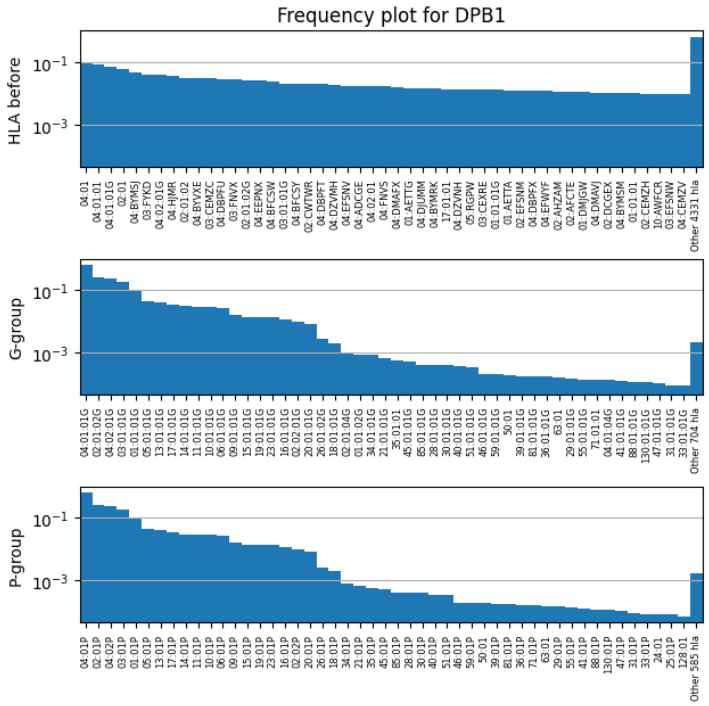


Figure A.4: Frequency plot for the different HLA descriptions for locus DPB1. The 50 largest HLA were plotted with the percentage of donors with that HLA on a logarithmic scale. All smaller HLA were grouped into a single bar in the far right of each subplot were the label tells you how many categories were summarized in this bar.

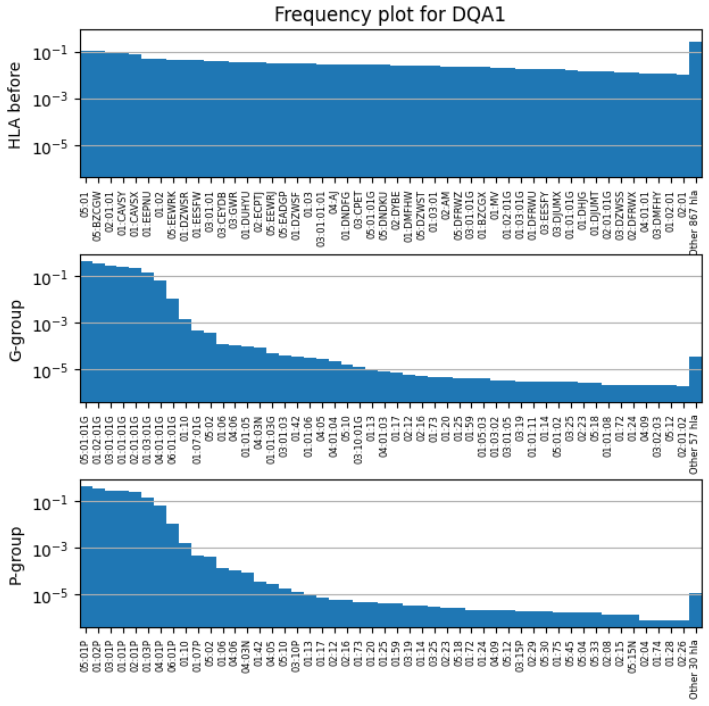


Figure A.5: Frequency plot for the different HLA descriptions for locus DQA1. The 50 largest HLA were plotted with the percentage of donors with that HLA on a logarithmic scale. All smaller HLA were grouped into a single bar in the far right of each subplot were the label tells you how many categories were summarized in this bar.

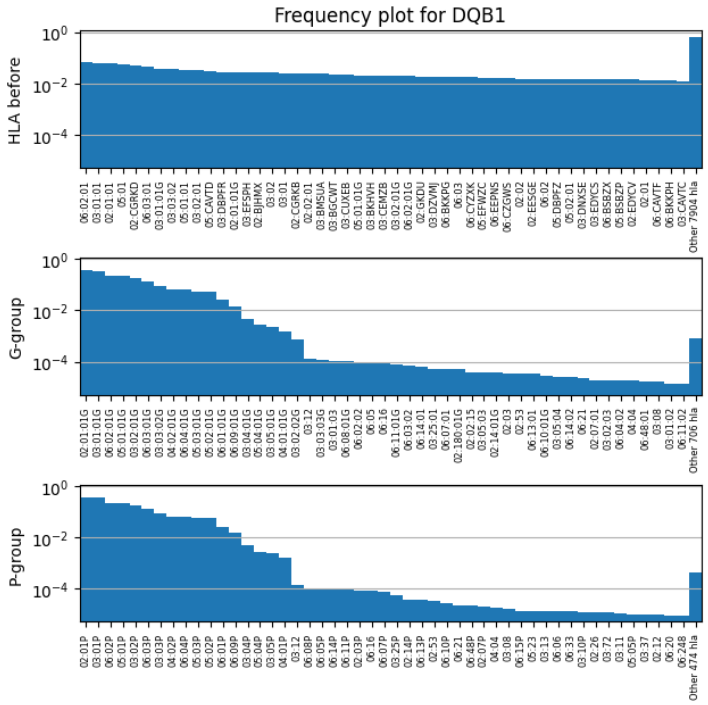


Figure A.6: Frequency plot for the different HLA descriptions for locus DQB1. The 50 largest HLA were plotted with the percentage of donors with that HLA on a logarithmic scale. All smaller HLA were grouped into a single bar in the far right of each subplot were the label tells you how many categories were summarized in this bar.

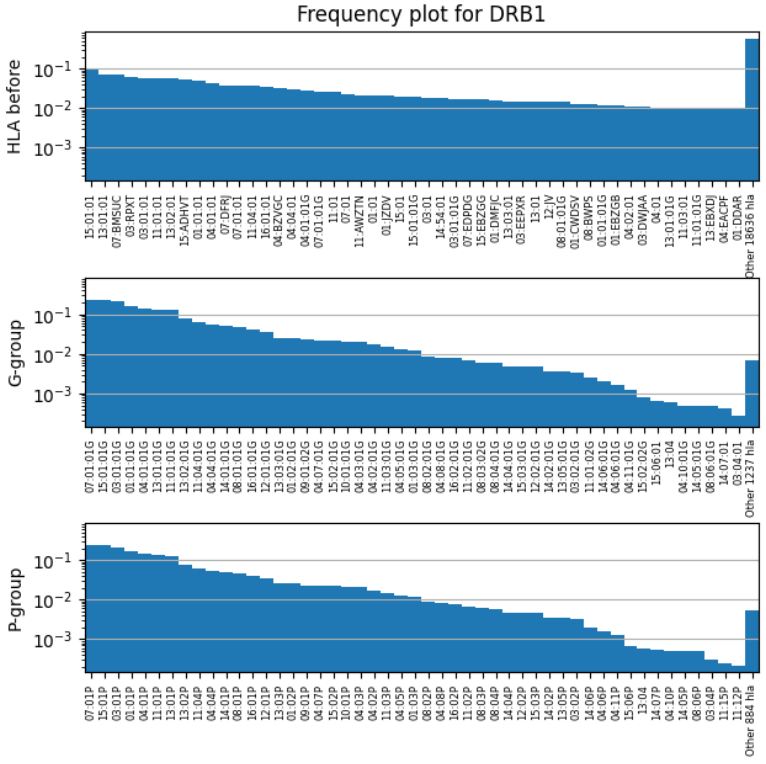


Figure A.7: Frequency plot for the different HLA descriptions for locus DRB1. The 50 largest HLA were plotted with the percentage of donors with that HLA on a logarithmic scale. All smaller HLA were grouped into a single bar in the far right of each subplot were the label tells you how many categories were summarized in this bar.

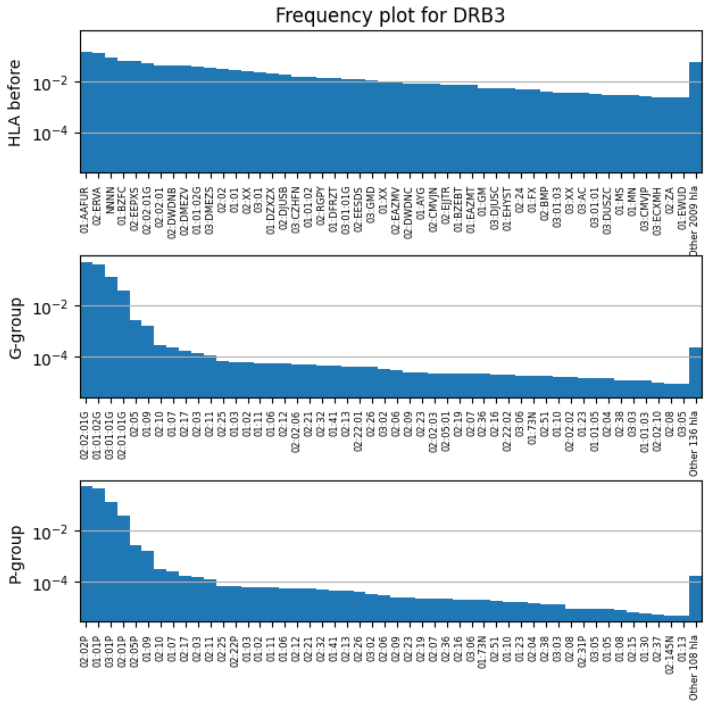


Figure A.8: Frequency plot for the different HLA descriptions for locus DRB3. The 50 largest HLA were plotted with the percentage of donors with that HLA on a logarithmic scale. All smaller HLA were grouped into a single bar in the far right of each subplot were the label tells you how many categories were summarized in this bar.

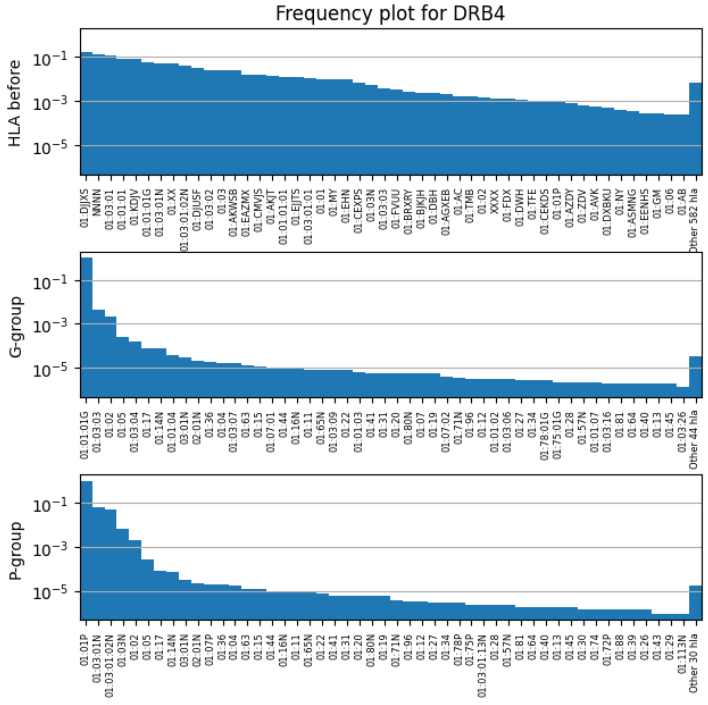


Figure A.9: Frequency plot for the different HLA descriptions for locus DRB4. The 50 largest HLA were plotted with the percentage of donors with that HLA on a logarithmic scale. All smaller HLA were grouped into a single bar in the far right of each subplot were the label tells you how many categories were summarized in this bar.

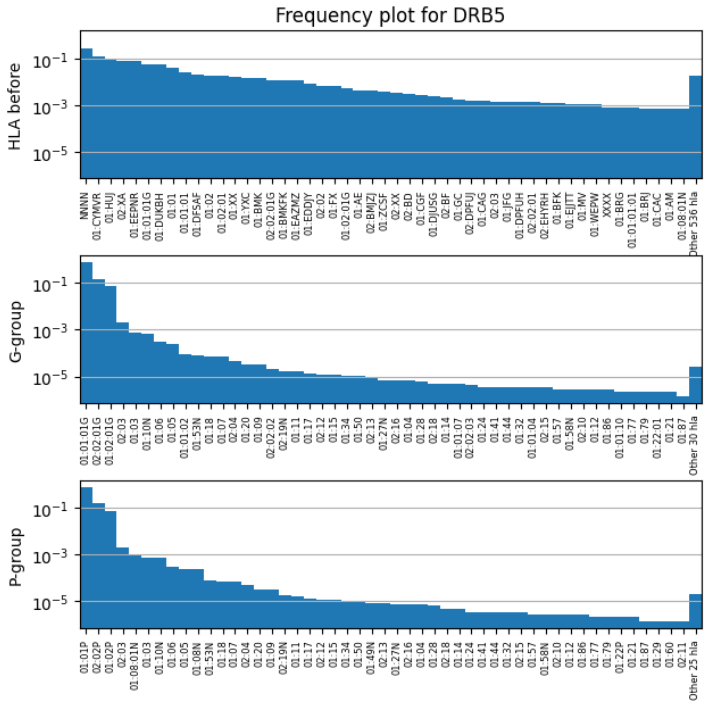


Figure A.10: Frequency plot for the different HLA descriptions for locus DRB5. The 50 largest HLA were plotted with the percentage of donors with that HLA on a logarithmic scale. All smaller HLA were grouped into a single bar in the far right of each subplot were the label tells you how many categories were summarized in this bar.

A.2. Vulcano Plots per Loci and Group

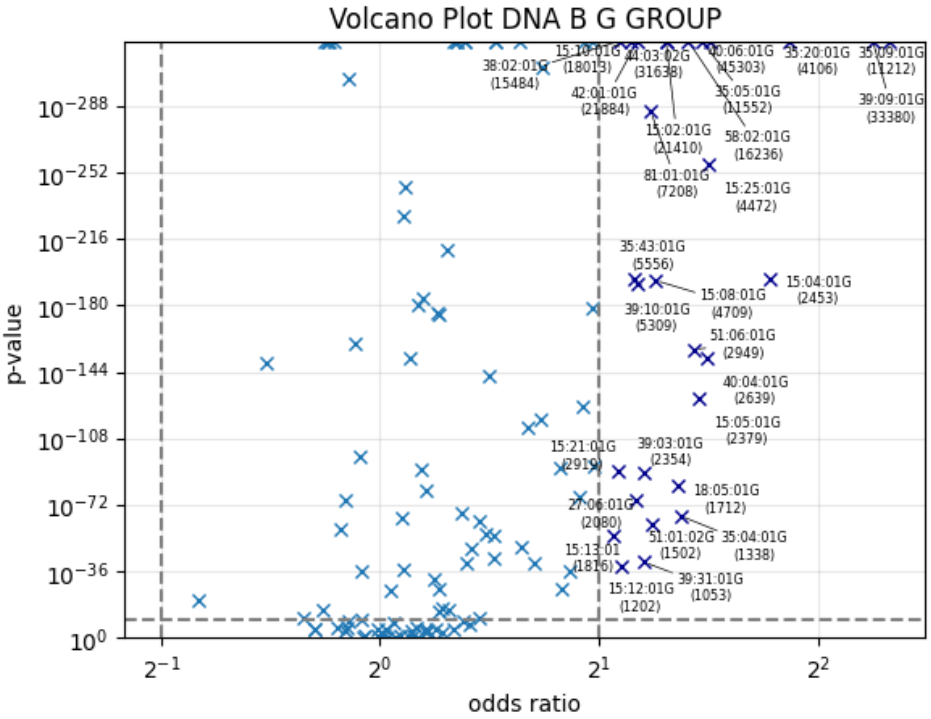


Figure A.11: Volcano plot for the different HLA G-groups for locus B. For each HLA group in the significant regions the name of the group is visible in the plot followed by the number of donors with the given HLA and a CMV serostatus.

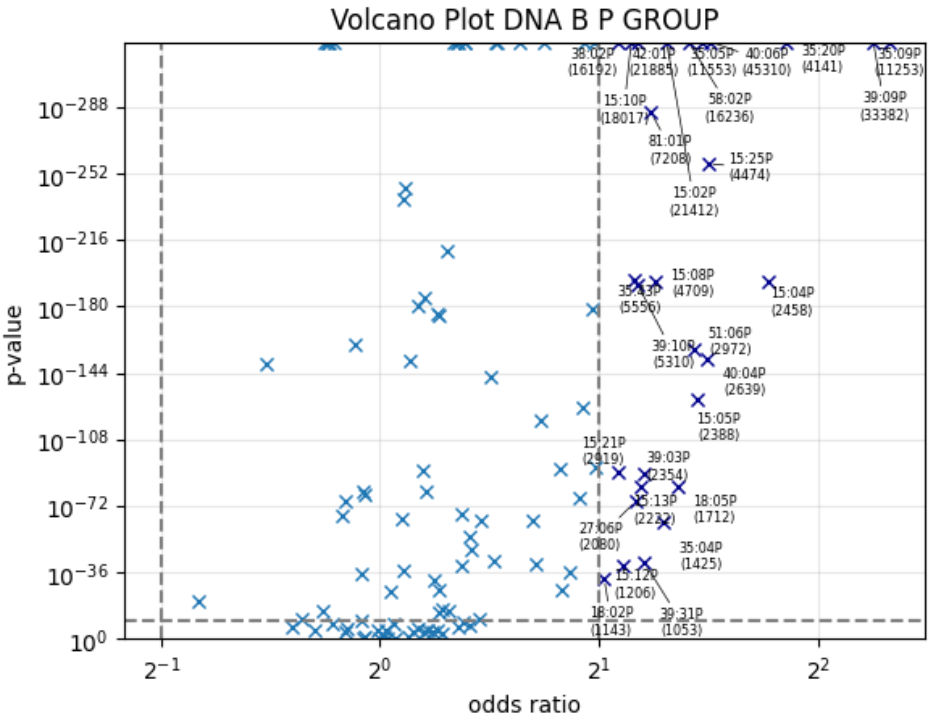


Figure A.12: Volcano plot for the different HLA P-groups for locus B. For each HLA group in the significant regions the name of the group is visible in the plot followed by the number of donors with the given HLA and a CMV serostatus.

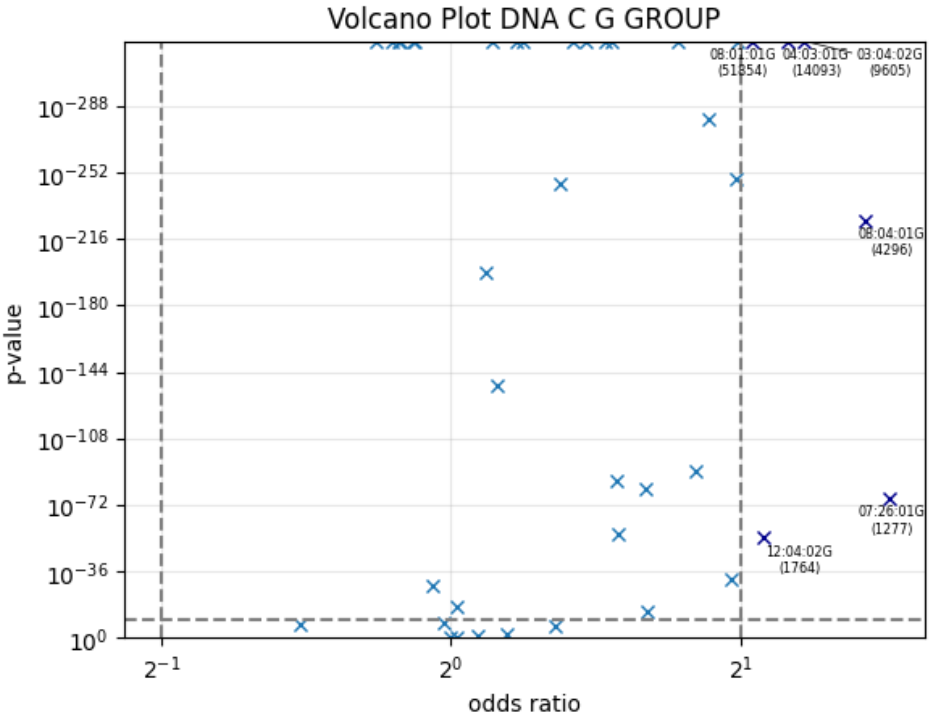


Figure A.13: Volcano plot for the different HLA G-groups for locus C. For each HLA group in the significant regions the name of the group is visible in the plot followed by the number of donors with the given HLA and a CMV serostatus.

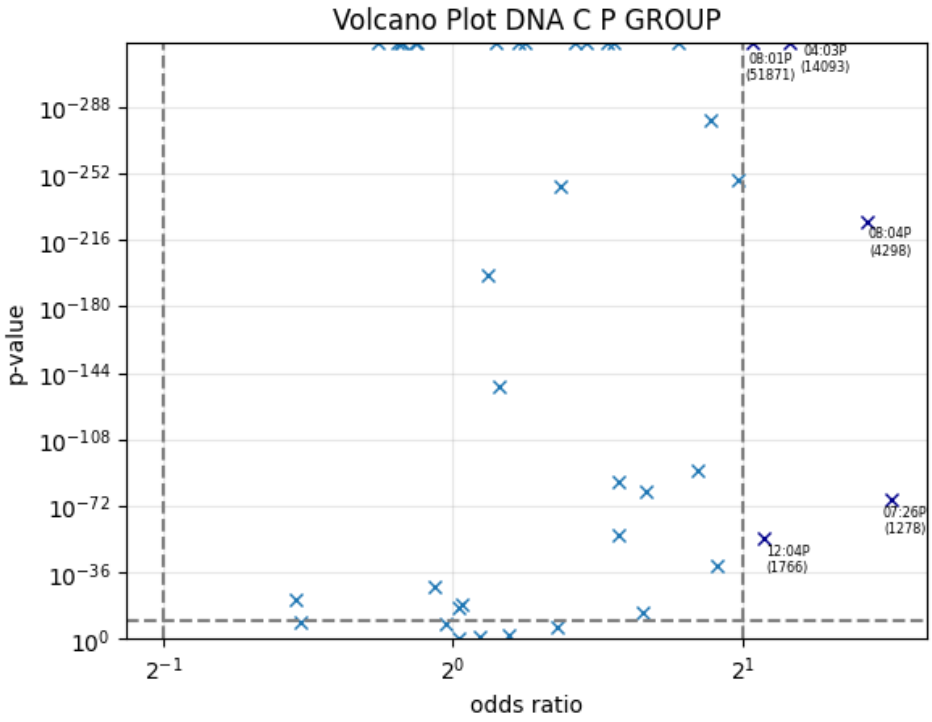


Figure A.14: Volcano plot for the different HLA P-groups for locus C. For each HLA group in the significant regions the name of the group is visible in the plot followed by the number of donors with the given HLA and a CMV serostatus.

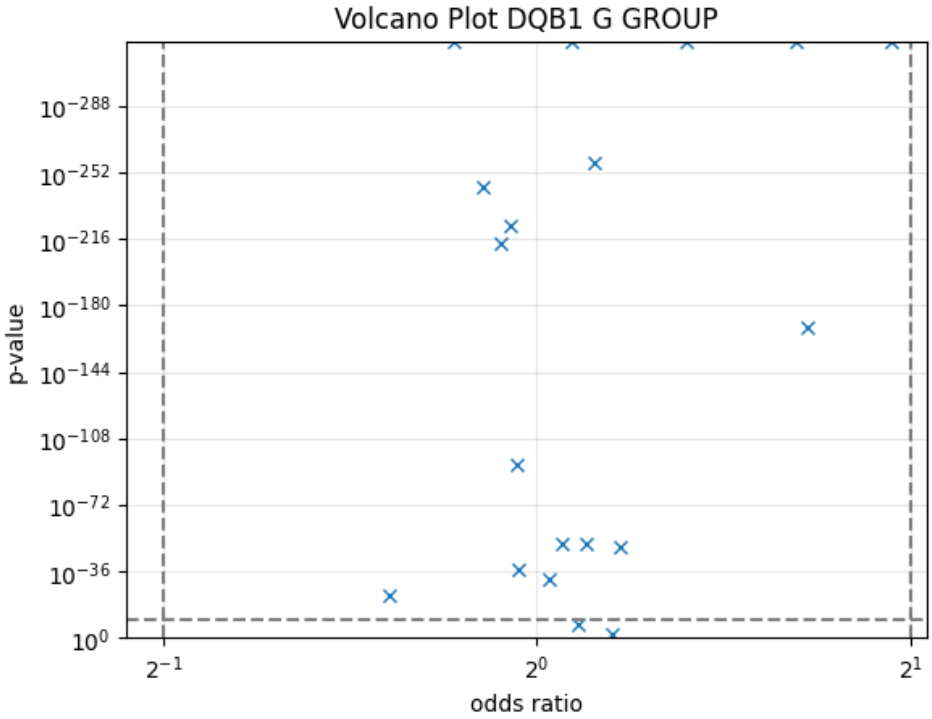


Figure A.15: Volcano plot for the different HLA G-groups for locus DQB1. For each HLA group in the significant regions the name of the group is visible in the plot followed by the number of donors with the given HLA and a CMV serostatus.

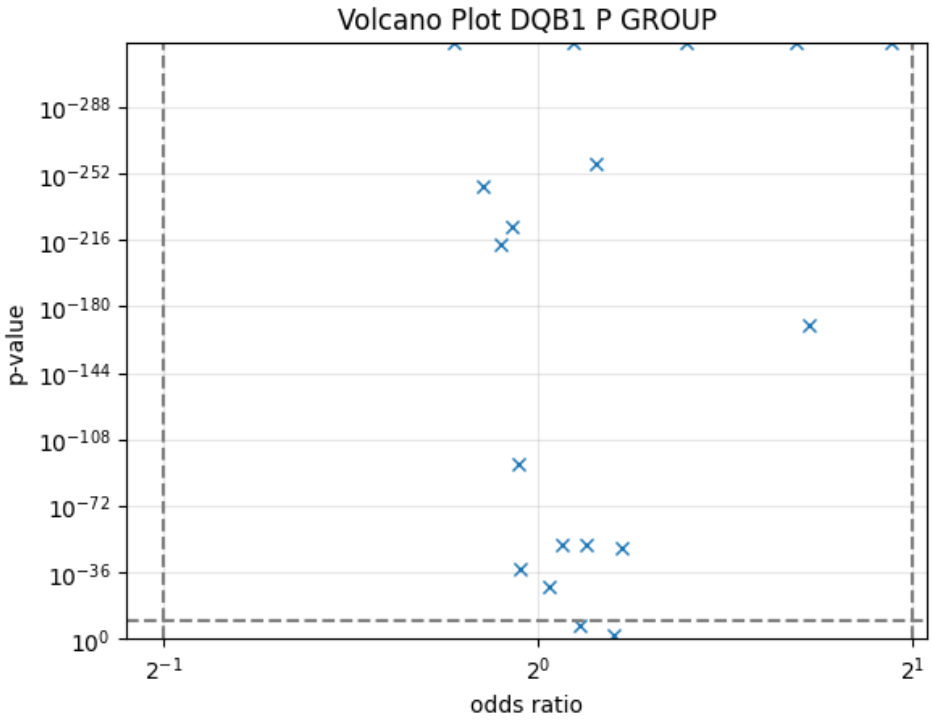


Figure A.16: Volcano plot for the different HLA P-groups for locus DQB1. For each HLA group in the significant regions the name of the group is visible in the plot followed by the number of donors with the given HLA and a CMV serostatus.

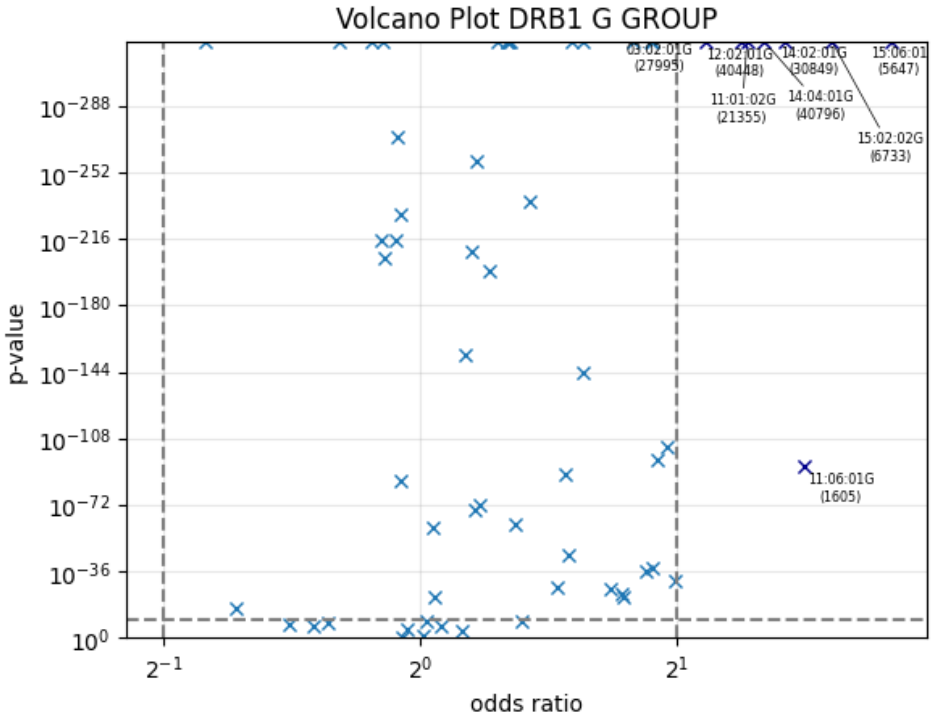


Figure A.17: Volcano plot for the different HLA G-groups for locus DRB1. For each HLA group in the significant regions the name of the group is visible in the plot followed by the number of donors with the given HLA and a CMV serostatus.

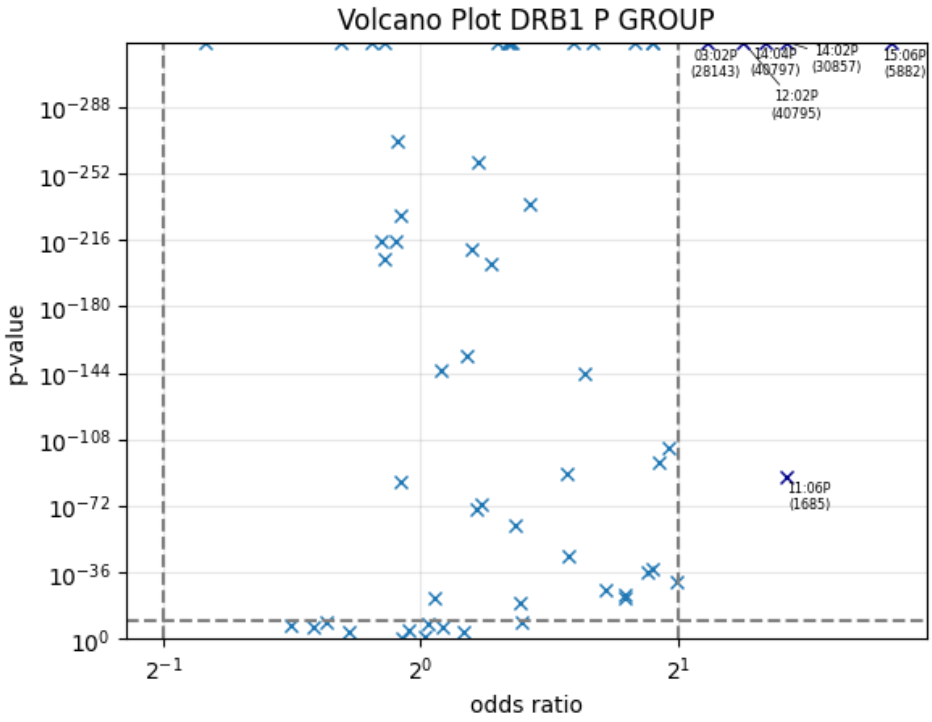


Figure A.18: Volcano plot for the different HLA P-groups for locus DRB1. For each HLA group in the significant regions the name of the group is visible in the plot followed by the number of donors with the given HLA and a CMV serostatus.

A.3. Meta Analysis Forest Plots

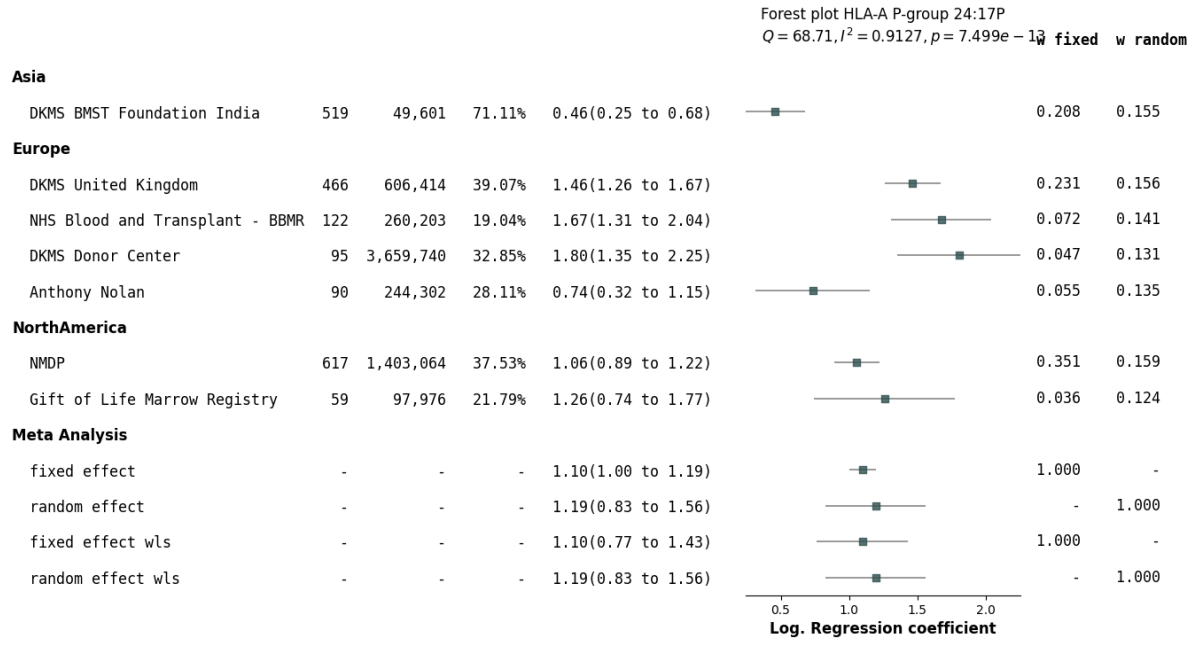


Figure A.19: Forest plot for the meta analysis of 24:17P on locus A with $OR = 3.50$, $p = 5.00 \times 10^{-171}$ and $N = 2\,086$.

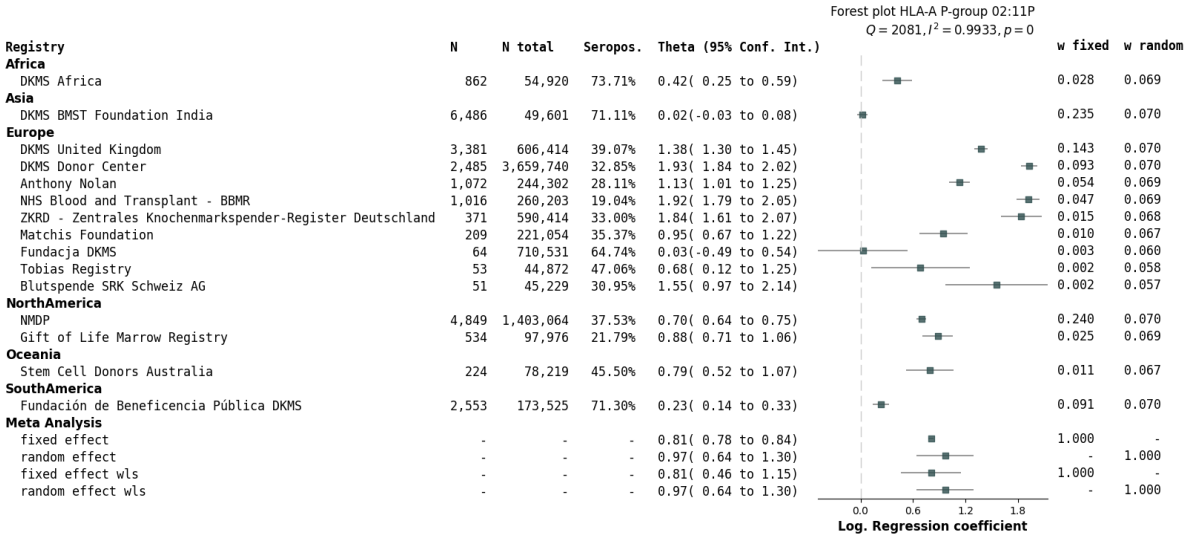


Figure A.20: Forest plot for the meta analysis of 02:11P on locus A with $OR = 3.29$, $p = 4.94 \times 10^{-324}$ and $N = 24\,544$.

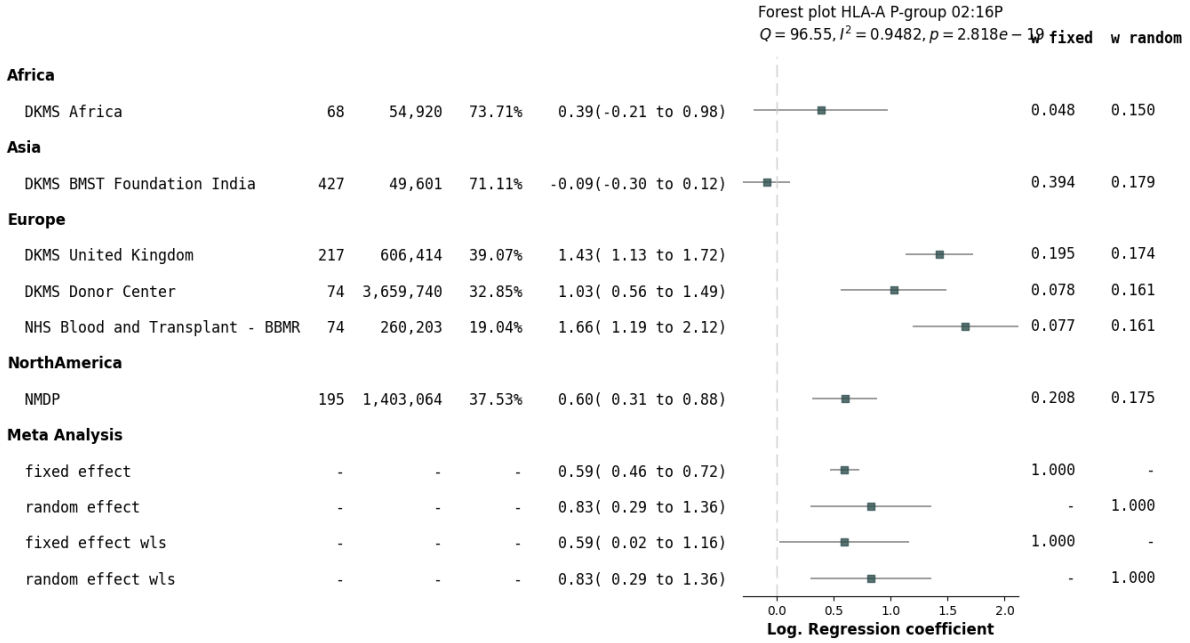


Figure A.21: Forest plot for the meta analysis of 02:16P on locus A with $OR = 2.85$, $p = 3.61 \times 10^{-70}$ and $N = 1\,174$.

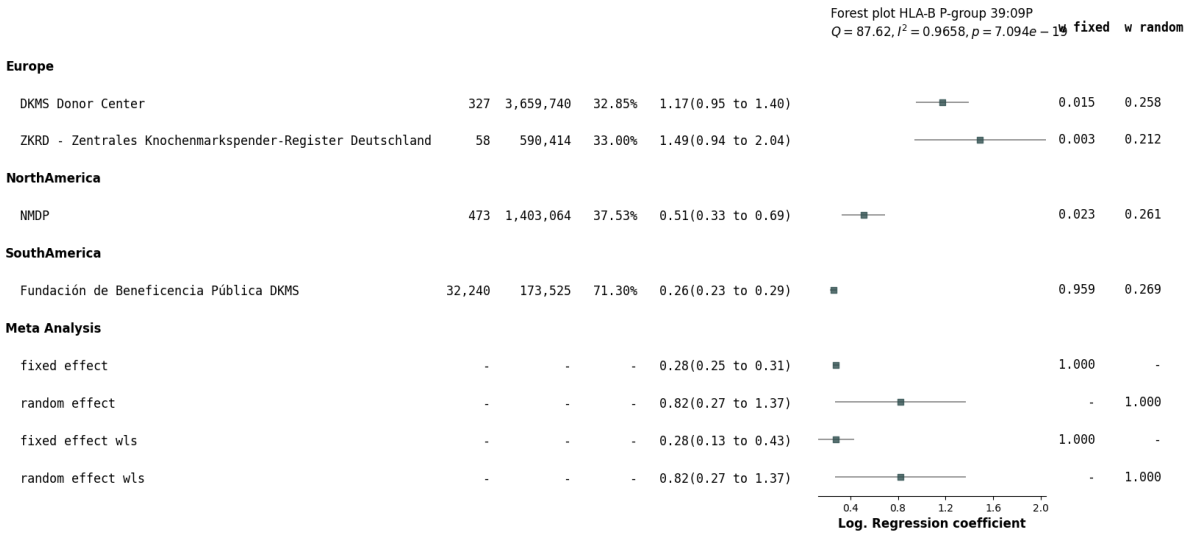


Figure A.22: Forest plot for the meta analysis of 39:09P on locus B with $OR = 4.76$, $p = 4.94 \times 10^{-324}$ and $N = 33\,382$.

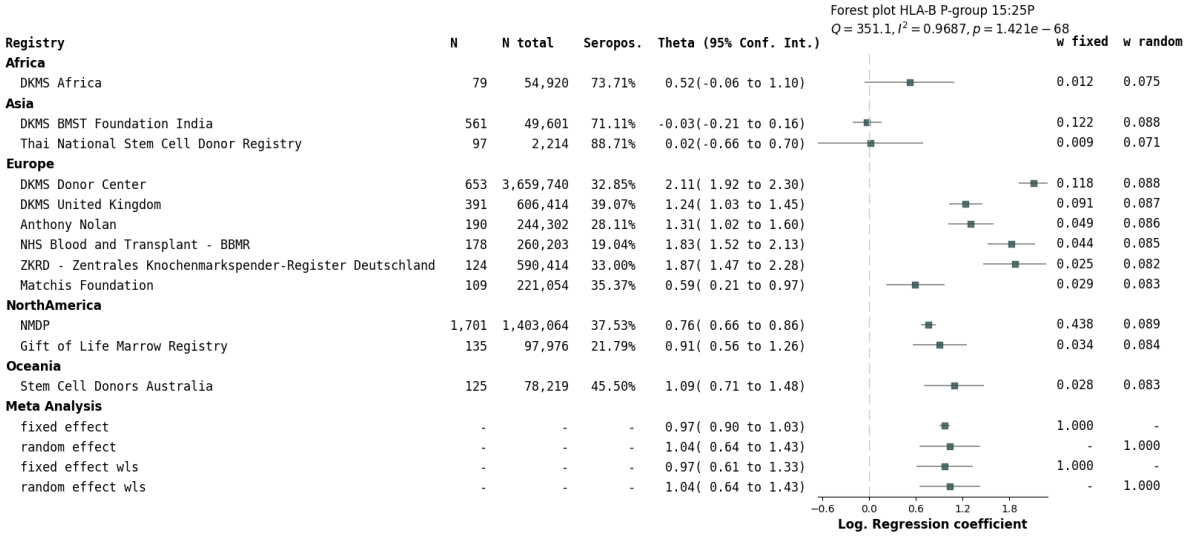


Figure A.23: Forest plot for the meta analysis of 15:25P on locus B with $OR = 2.83$, $p = 8.06 \times 10^{-258}$ and $N = 4\,474$..

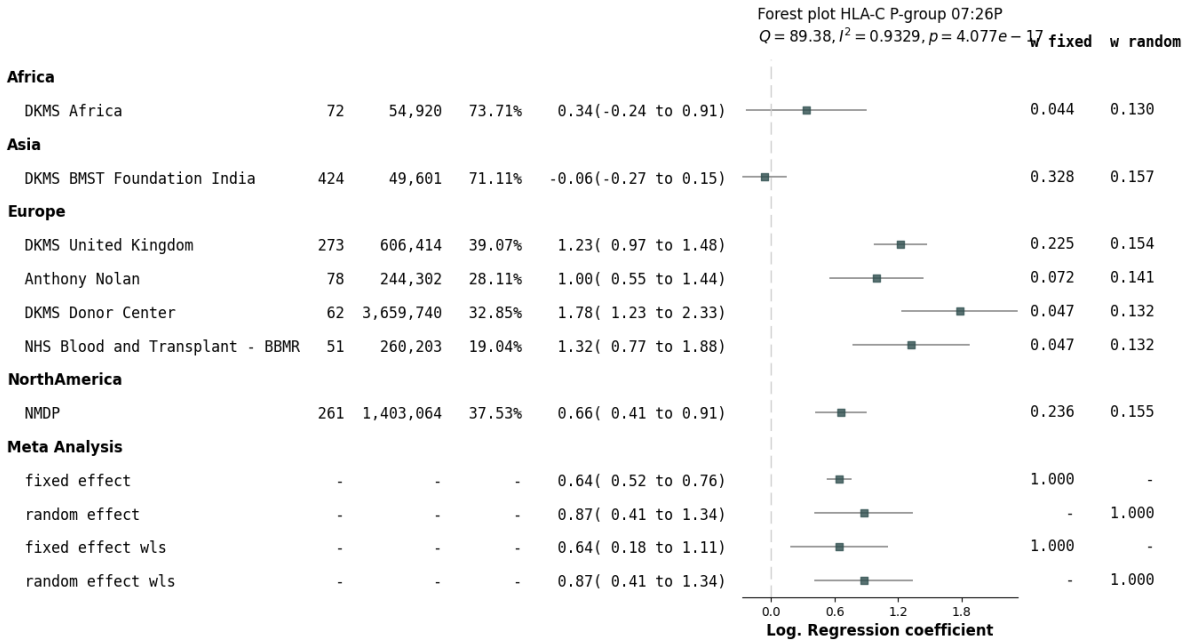


Figure A.24: Forest plot for the meta analysis of 07:26P on locus C with $OR = 2.85$, $p = 2.51 \times 10^{-76}$ and $N = 1\,278$.

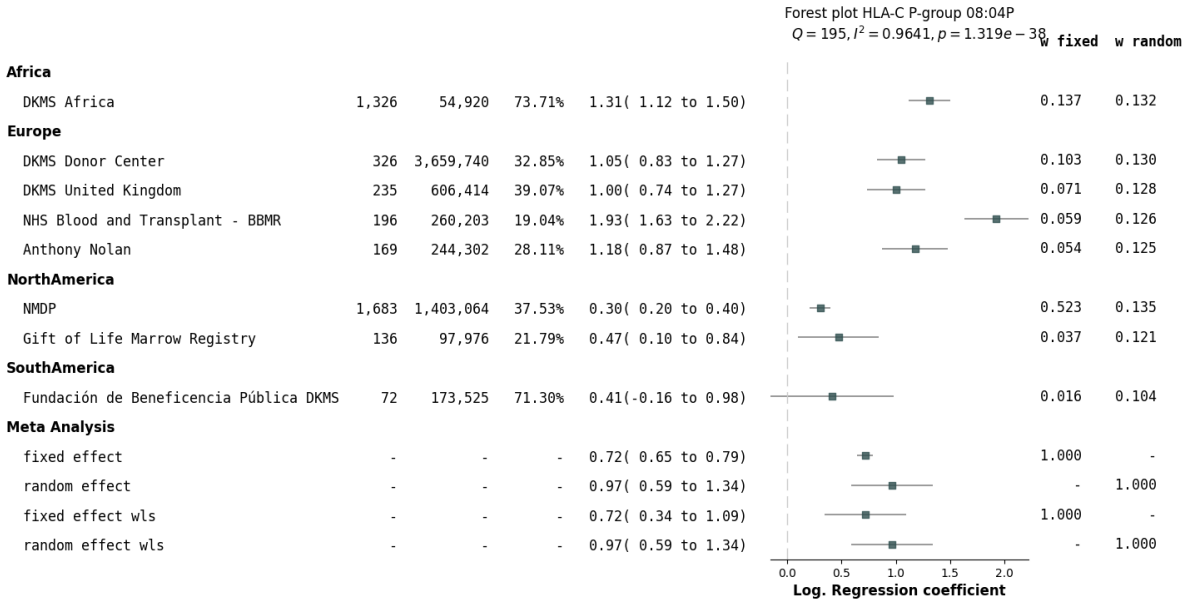
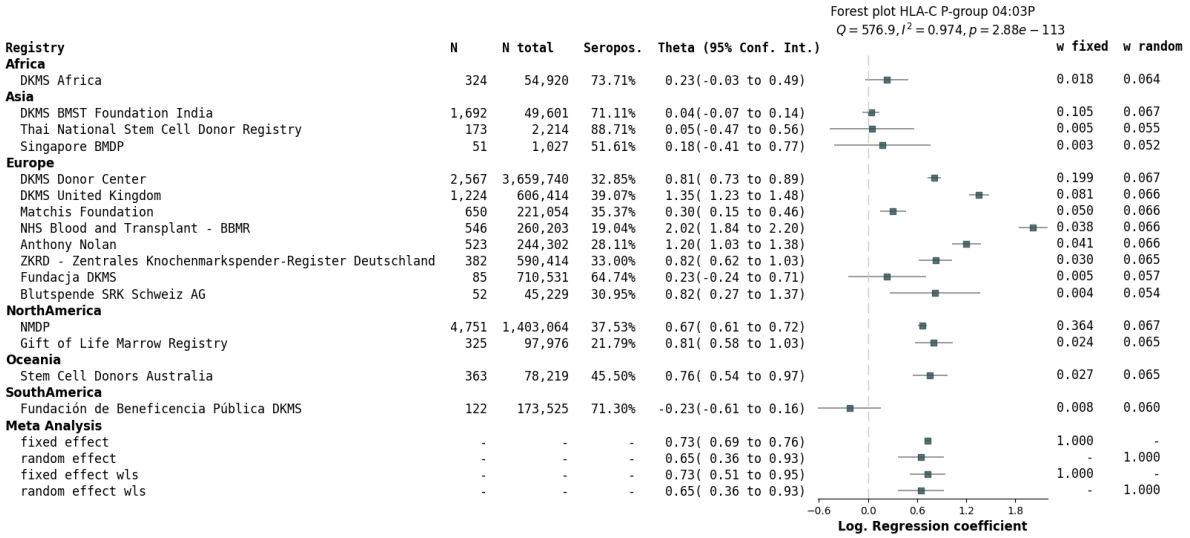
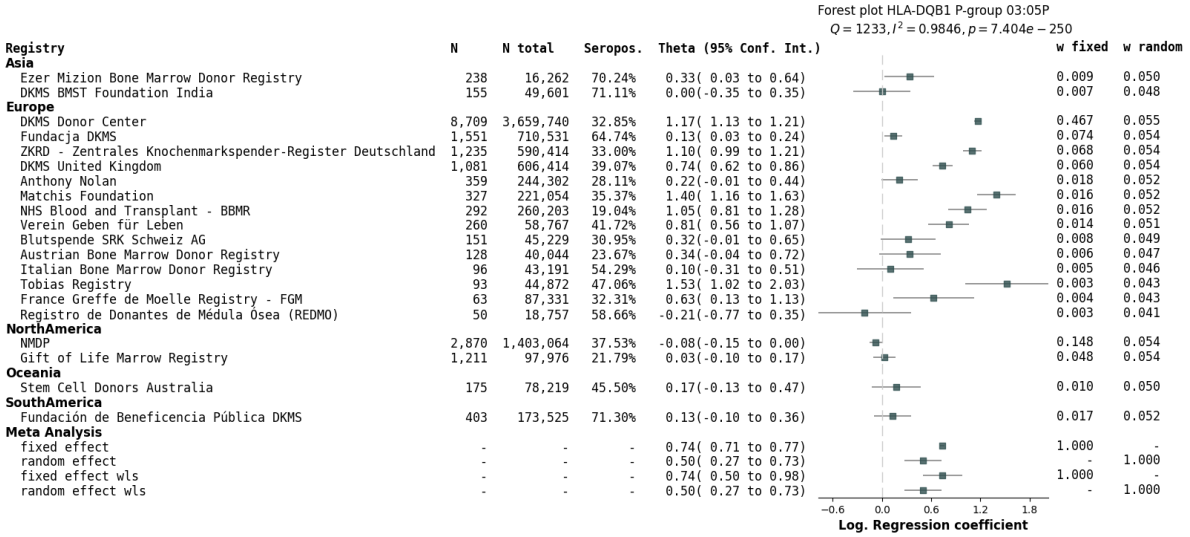


Figure A.25: Forest plot for the meta analysis of 08:04P on locus C with $OR = 2.69$, $p = 1.32 \times 10^{-226}$ and $N = 4\,298$.



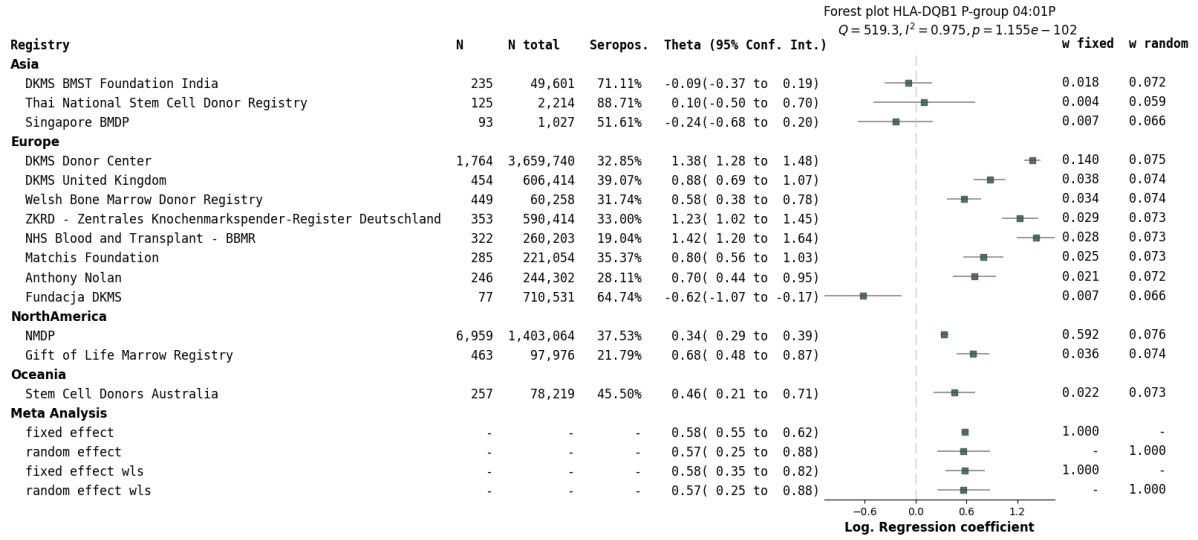
Registry	N	N total	Seropos.	Theta (95% Conf. Int.)	w fixed	w random
Africa						
DKMS Africa	324	54,920	73.71%	0.23(-0.03 to 0.49)	0.018	0.064
Asia						
DKMS BMST Foundation India	1,692	49,601	71.11%	0.04(-0.07 to 0.14)	0.105	0.067
Thai National Stem Cell Donor Registry	173	2,214	88.71%	0.05(-0.47 to 0.56)	0.005	0.055
Singapore BMDP	51	1,027	51.61%	0.18(-0.41 to 0.77)	0.003	0.052
Europe						
DKMS Donor Center	2,567	3,659,740	32.85%	0.81(0.73 to 0.89)	0.199	0.067
DKMS United Kingdom	1,224	606,414	39.07%	1.35(1.23 to 1.48)	0.081	0.066
Matchis Foundation	650	221,054	35.37%	0.30(0.15 to 0.46)	0.050	0.066
NHS Blood and Transplant - BBMR	546	260,203	19.04%	2.02(1.84 to 2.20)	0.038	0.066
Anthony Nolan	523	244,302	28.11%	1.20(1.03 to 1.38)	0.041	0.066
ZKRD - Zentrales Knochenmarkspender-Register Deutschland	382	590,414	33.00%	0.82(0.62 to 1.03)	0.030	0.065
Fundacja DKMS	85	710,531	64.74%	0.23(-0.24 to 0.71)	0.005	0.057
Blutspende SRK Schweiz AG	52	45,229	30.95%	0.82(0.27 to 1.37)	0.004	0.054
NorthAmerica						
NMDP	4,751	1,403,064	37.53%	0.67(0.61 to 0.72)	0.364	0.067
Gift of Life Marrow Registry	325	97,976	21.79%	0.81(0.58 to 1.03)	0.024	0.065
Oceania						
Stem Cell Donors Australia	363	78,219	45.50%	0.76(0.54 to 0.97)	0.027	0.065
SouthAmerica						
Fundación de Beneficencia Pública DKMS	122	173,525	71.30%	-0.23(-0.61 to 0.16)	0.008	0.060
Meta Analysis						
fixed effect	-	-	-	0.73(0.69 to 0.76)	1.000	-
random effect	-	-	-	0.65(0.36 to 0.93)	-	1.000
fixed effect wls	-	-	-	0.73(0.51 to 0.95)	1.000	-
random effect wls	-	-	-	0.65(0.36 to 0.93)	-	1.000

Figure A.26: Forest plot for the meta analysis of 04:03P on locus C with $OR = 2.24$, $p = 4.94 \times 10^{-324}$ and $N = 14\,093$.



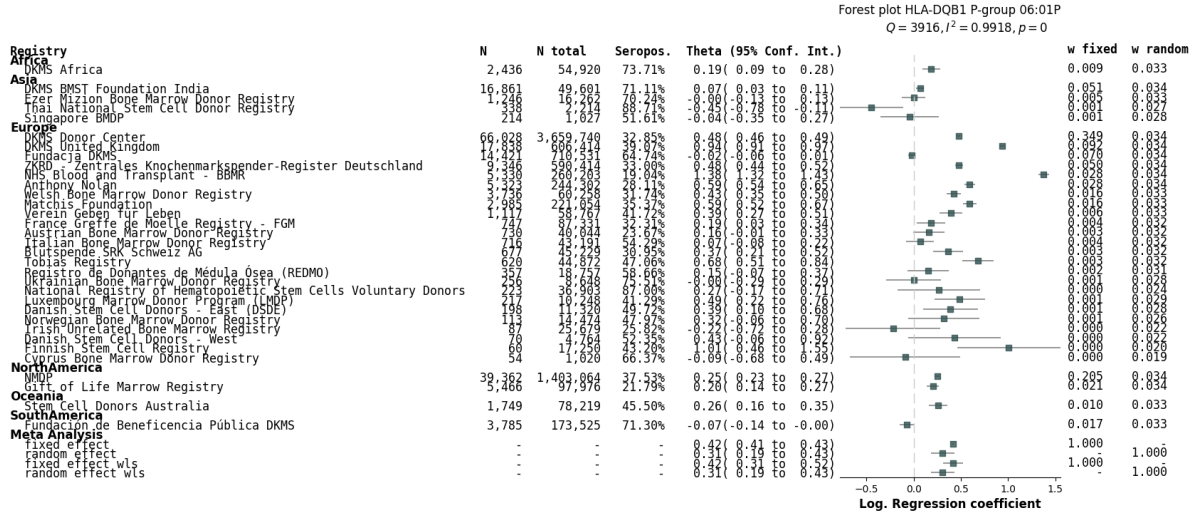
Registry	N	N total	Seropos.	Theta (95% Conf. Int.)	w fixed	w random
Asia						
Ezer Mizion Bone Marrow Donor Registry	238	16,262	70.24%	0.33(0.03 to 0.64)	0.009	0.050
DKMS BMST Foundation India	155	49,601	71.11%	0.00(-0.35 to 0.35)	0.007	0.048
Europe						
DKMS Donor Center	8,709	3,659,740	32.85%	1.17(1.13 to 1.21)	0.467	0.055
Fundacja DKMS	1,551	710,531	64.74%	0.13(0.03 to 0.24)	0.074	0.054
ZKRD - Zentrales Knochenmarkspender-Register Deutschland	1,235	590,414	33.00%	1.10(0.99 to 1.21)	0.068	0.054
DKMS United Kingdom	1,081	606,414	39.07%	0.74(0.62 to 0.86)	0.060	0.054
Anthony Nolan	359	244,302	28.11%	0.22(-0.01 to 0.44)	0.018	0.052
Matchis Foundation	327	221,054	35.37%	1.40(1.16 to 1.63)	0.016	0.052
NHS Blood and Transplant - BBMR	292	260,203	19.04%	1.05(0.81 to 1.28)	0.016	0.052
Verein Geben für Leben	260	58,767	41.72%	0.81(0.56 to 1.07)	0.014	0.051
Blutspende SRK Schweiz AG	151	45,229	30.95%	0.32(-0.01 to 0.65)	0.008	0.049
Austrian Bone Marrow Donor Registry	128	40,044	23.67%	0.34(-0.04 to 0.72)	0.006	0.047
Italian Bone Marrow Donor Registry	96	43,191	54.29%	0.10(-0.31 to 0.51)	0.005	0.046
Tobias Registry	93	44,872	47.06%	1.53(1.02 to 2.03)	0.003	0.043
France Greffe de Moelle Registry - FGM	63	87,331	32.31%	0.63(0.13 to 1.13)	0.004	0.043
Registro de Donantes de Médula Ósea (REDMO)	50	18,757	58.66%	-0.21(-0.77 to 0.35)	0.003	0.041
NorthAmerica						
NMDP	2,870	1,403,064	37.53%	-0.08(-0.15 to 0.00)	0.148	0.054
Gift of Life Marrow Registry	1,211	97,976	21.79%	0.03(-0.10 to 0.17)	0.048	0.054
Oceania						
Stem Cell Donors Australia	175	78,219	45.50%	0.17(-0.13 to 0.47)	0.010	0.050
SouthAmerica						
Fundación de Beneficencia Pública DKMS	403	173,525	71.30%	0.13(-0.10 to 0.36)	0.017	0.052
Meta Analysis						
fixed effect	-	-	-	0.74(0.71 to 0.77)	1.000	-
random effect	-	-	-	0.50(0.27 to 0.73)	-	1.000
fixed effect wls	-	-	-	0.74(0.50 to 0.98)	1.000	-
random effect wls	-	-	-	0.50(0.27 to 0.73)	-	1.000

Figure A.27: Forest plot for the meta analysis of 03:05P on locus DQB1 with $OR = 1.92$, $p = 4.94 \times 10^{-324}$ and $N = 19\,744$.



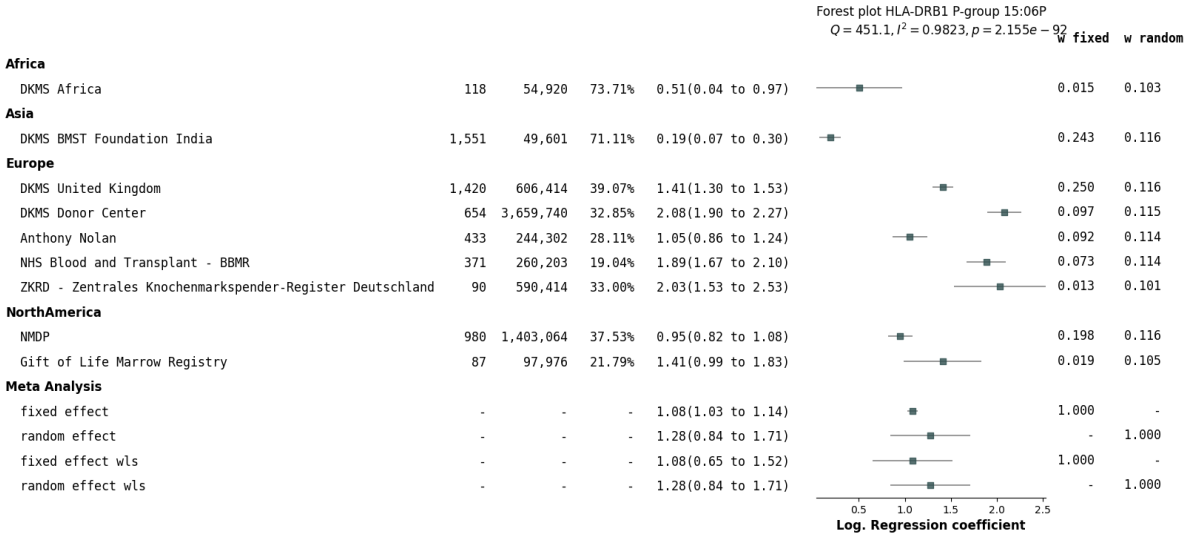
Registry	N	N total	Seropos.	Theta (95% Conf. Int.)	w fixed	w random
Asia						
DKMS BMST Foundation India	235	49,601	71.11%	-0.09(-0.37 to 0.19)	0.018	0.072
Thai National Stem Cell Donor Registry	125	2,214	88.71%	0.10(-0.50 to 0.70)	0.004	0.059
Singapore BMDP	93	1,027	51.61%	-0.24(-0.68 to 0.20)	0.007	0.066
Europe						
DKMS Donor Center	1,764	3,659,740	32.85%	1.38(1.28 to 1.48)	0.140	0.075
DKMS United Kingdom	454	606,414	39.07%	0.88(0.69 to 1.07)	0.038	0.074
Welsh Bone Marrow Donor Registry	449	60,258	31.74%	0.58(0.38 to 0.78)	0.034	0.074
ZKRD - Zentrales Knochenmarkspender-Register Deutschland	353	590,414	33.00%	1.23(1.02 to 1.45)	0.029	0.073
NHS Blood and Transplant - BBMR	322	260,203	19.04%	1.42(1.20 to 1.64)	0.028	0.073
Matchis Foundation	285	221,054	35.37%	0.80(0.56 to 1.03)	0.025	0.073
Anthony Nolan	246	244,302	28.11%	0.70(0.44 to 0.95)	0.021	0.072
Fundacja DKMS	77	710,531	64.74%	-0.62(-1.07 to -0.17)	0.007	0.066
NorthAmerica						
NMDP	6,959	1,403,064	37.53%	0.34(0.29 to 0.39)	0.592	0.076
Gift of Life Marrow Registry	463	97,976	21.79%	0.68(0.48 to 0.87)	0.036	0.074
Oceania						
Stem Cell Donors Australia	257	78,219	45.50%	0.46(0.21 to 0.71)	0.022	0.073
Meta Analysis						
fixed effect	-	-	-	0.58(0.55 to 0.62)	1.000	-
random effect	-	-	-	0.57(0.25 to 0.88)	-	1.000
fixed effect wls	-	-	-	0.58(0.35 to 0.82)	1.000	-
random effect wls	-	-	-	0.57(0.25 to 0.88)	-	1.000

Figure A.28: Forest plot for the meta analysis of 04:01P on locus DQB1 with $OR = 1.65$, $p = 2.77 \times 10^{-170}$ and $N = 12\,429$.



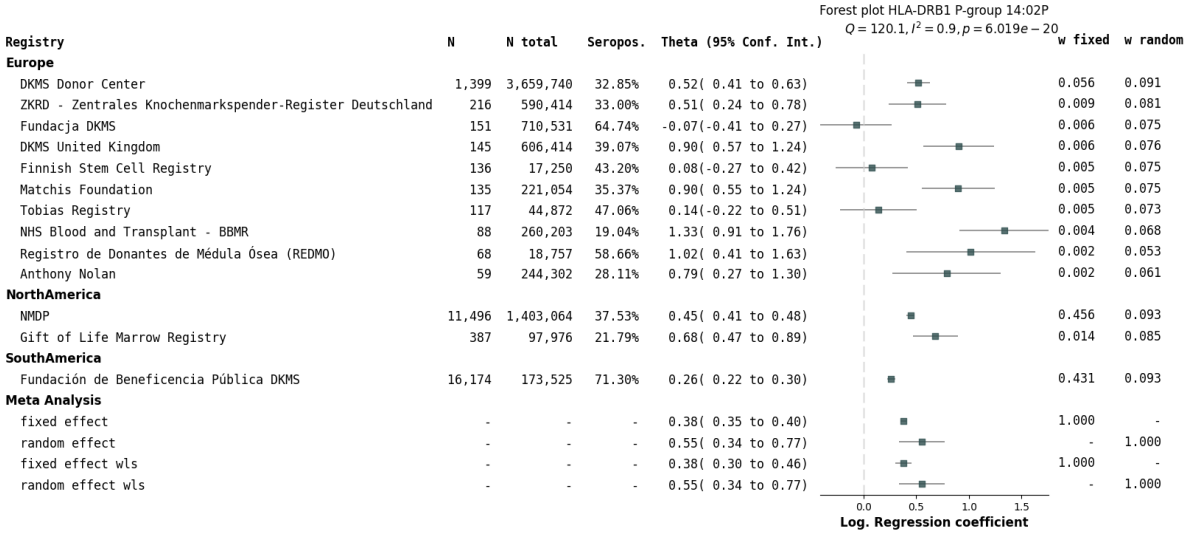
Registry	N	N total	Seropos.	Theta (95% Conf. Int.)	w fixed	w random
Africa						
DKMS Africa	2,436	54,920	73.71%	0.19(0.09 to 0.28)	0.009	0.033
Asia						
DKMS BMST Foundation India	16,861	49,601	71.11%	0.07(0.03 to 0.11)	0.051	0.034
Ezer Mizion Bone Marrow Donor Registry	1,246	16,262	70.24%	-0.00(-0.13 to 0.13)	0.005	0.033
Thai National Stem Cell Donor Registry	338	2,214	88.71%	-0.45(-0.78 to -0.11)	0.001	0.027
Singapore BMDP	214	1,027	51.61%	-0.04(-0.35 to 0.27)	0.001	0.028
Europe						
DKMS Donor Center	66,028	3,659,740	32.85%	0.48(0.46 to 0.49)	0.349	0.034
DKMS United Kingdom	17,838	606,414	39.07%	0.94(0.91 to 0.97)	0.092	0.034
Fundacja DKMS	14,421	710,531	64.74%	-0.02(-0.06 to 0.01)	0.070	0.034
ZKRD - Zentrales Knochenmarkspender-Register Deutschland	9,346	590,414	33.00%	0.48(0.44 to 0.52)	0.060	0.034
NHS Blood and Transplant - BBMR	5,330	260,203	19.04%	1.38(1.32 to 1.43)	0.028	0.034
Anthony Nolan	5,323	244,302	28.11%	0.59(0.54 to 0.65)	0.028	0.034
Welsh Bone Marrow Donor Registry	3,745	60,258	31.74%	0.33(0.25 to 0.41)	0.010	0.033
Matchis Foundation	3,736	221,054	35.37%	0.43(0.35 to 0.50)	0.016	0.033
Verein Geben für Leben	2,985	58,707	41.72%	0.59(0.52 to 0.67)	0.016	0.033
France Greffe de Moelle Registry - FGM	1,117	87,331	32.31%	0.39(0.27 to 0.51)	0.006	0.033
Austrian Bone Marrow Donor Registry	747	40,044	23.67%	0.19(0.03 to 0.34)	0.004	0.032
Italian Bone Marrow Donor Registry	736	43,191	54.29%	0.16(-0.01 to 0.33)	0.003	0.032
Blutspende SRK Schweiz AG	716	42,774	54.62%	0.07(-0.08 to 0.22)	0.004	0.032
Tobias Registry	677	45,229	30.95%	0.37(0.21 to 0.52)	0.003	0.032
Registro de Donantes de Médula Ósea (REDMO)	620	44,872	47.06%	0.68(0.51 to 0.84)	0.003	0.032
Ukrainian Bone Marrow Donor Registry	357	18,737	58.66%	0.15(-0.07 to 0.37)	0.002	0.031
National Registry of Hematopoietic Stem Cells Voluntary Donors	256	8,646	75.51%	-0.00(-0.29 to 0.29)	0.001	0.028
Luxembourg Marrow Donor Program (LMDP)	253	36,903	67.00%	0.27(-0.17 to 0.71)	0.000	0.024
Danish Stem Cell Donors - East (DSDE)	217	10,248	41.29%	0.49(0.22 to 0.76)	0.001	0.029
Norwegian Bone Marrow Donor Registry	198	11,320	49.72%	0.39(0.10 to 0.68)	0.001	0.028
Irish Unrelated Bone Marrow Registry	113	14,474	47.97%	0.32(-0.06 to 0.70)	0.001	0.026
Danish Stem Cell Donors - West	87	25,679	25.82%	-0.22(-0.72 to 0.28)	0.000	0.022
Cyprus Bone Marrow Donor Registry	70	4,764	52.35%	0.43(-0.06 to 0.92)	0.000	0.022
Finnish Stem Cell Registry	60	17,256	43.26%	1.01(0.46 to 1.55)	0.000	0.020
	54	1,020	66.37%	-0.09(-0.68 to 0.49)	0.000	0.019
NorthAmerica						
NMDP	39,362	1,403,064	37.53%	0.25(0.23 to 0.27)	0.205	0.034
Gift of Life Marrow Registry	5,466	97,976	21.79%	0.20(0.14 to 0.27)	0.021	0.034
Oceania						
Stem Cell Donors Australia	1,749	78,219	45.50%	0.26(0.16 to 0.35)	0.010	0.033
SouthAmerica						
Fundacion de Beneficencia Pública DKMS	3,785	173,525	71.30%	-0.07(-0.14 to -0.00)	0.017	0.033
Meta Analysis						
fixed effect	-	-	-	0.42(0.41 to 0.43)	1.000	-
random effect	-	-	-	0.31(0.19 to 0.43)	-	1.000
fixed effect wls	-	-	-	0.42(0.31 to 0.52)	1.000	-
random effect wls	-	-	-	0.31(0.19 to 0.43)	-	1.000

Figure A.29: Forest plot for the meta analysis of 06:01P on locus DQB1 with $OR = 1.62$, $p = 4.94 \times 10^{-324}$ and $N = 202\,919$.



					w fixed	w random
Africa						
DKMS Africa	118	54,920	73.71%	0.51(0.04 to 0.97)	0.015	0.103
Asia						
DKMS BMST Foundation India	1,551	49,601	71.11%	0.19(0.07 to 0.30)	0.243	0.116
Europe						
DKMS United Kingdom	1,420	606,414	39.07%	1.41(1.30 to 1.53)	0.250	0.116
DKMS Donor Center	654	3,659,740	32.85%	2.08(1.90 to 2.27)	0.097	0.115
Anthony Nolan	433	244,302	28.11%	1.05(0.86 to 1.24)	0.092	0.114
NHS Blood and Transplant - BBMR	371	260,203	19.04%	1.89(1.67 to 2.10)	0.073	0.114
ZKRD - Zentrales Knochenmarkspender-Register Deutschland	90	590,414	33.00%	2.03(1.53 to 2.53)	0.013	0.101
NorthAmerica						
NMDP	980	1,403,064	37.53%	0.95(0.82 to 1.08)	0.198	0.116
Gift of Life Marrow Registry	87	97,976	21.79%	1.41(0.99 to 1.83)	0.019	0.105
Meta Analysis						
fixed effect	-	-	-	1.08(1.03 to 1.14)	1.000	-
random effect	-	-	-	1.28(0.84 to 1.71)	-	1.000
fixed effect wls	-	-	-	1.08(0.65 to 1.52)	1.000	-
random effect wls	-	-	-	1.28(0.84 to 1.71)	-	1.000

Figure A.30: Forest plot for the meta analysis of 15:06P on locus DRB1 with $OR = 3.54$, $p = 4.94 \times 10^{-324}$ and $N = 5\,882$.



Registry	N	N total	Seropos.	Theta (95% Conf. Int.)	w fixed	w random
Europe						
DKMS Donor Center	1,399	3,659,740	32.85%	0.52(0.41 to 0.63)	0.056	0.091
ZKRD - Zentrales Knochenmarkspender-Register Deutschland	216	590,414	33.00%	0.51(0.24 to 0.78)	0.009	0.081
Fundacja DKMS	151	710,531	64.74%	-0.07(-0.41 to 0.27)	0.006	0.075
DKMS United Kingdom	145	606,414	39.07%	0.90(0.57 to 1.24)	0.006	0.076
Finnish Stem Cell Registry	136	17,250	43.20%	0.08(-0.27 to 0.42)	0.005	0.075
Matchis Foundation	135	221,054	35.37%	0.90(0.55 to 1.24)	0.005	0.075
Tobias Registry	117	44,872	47.06%	0.14(-0.22 to 0.51)	0.005	0.073
NHS Blood and Transplant - BBMR	88	260,203	19.04%	1.33(0.91 to 1.76)	0.004	0.068
Registro de Donantes de Médula Ósea (REDMO)	68	18,757	58.66%	1.02(0.41 to 1.63)	0.002	0.053
Anthony Nolan	59	244,302	28.11%	0.79(0.27 to 1.30)	0.002	0.061
NorthAmerica						
NMDP	11,496	1,403,064	37.53%	0.45(0.41 to 0.48)	0.456	0.093
Gift of Life Marrow Registry	387	97,976	21.79%	0.68(0.47 to 0.89)	0.014	0.085
SouthAmerica						
Fundación de Beneficencia Pública DKMS	16,174	173,525	71.30%	0.26(0.22 to 0.30)	0.431	0.093
Meta Analysis						
fixed effect	-	-	-	0.38(0.35 to 0.40)	1.000	-
random effect	-	-	-	0.55(0.34 to 0.77)	-	1.000
fixed effect wls	-	-	-	0.38(0.30 to 0.46)	1.000	-
random effect wls	-	-	-	0.55(0.34 to 0.77)	-	1.000

Figure A.31: Forest plot for the meta analysis of 14:02P on locus DRB1 with $OR = 2.68$, $p = 4.94 \times 10^{-324}$ and $N = 30\,857$.

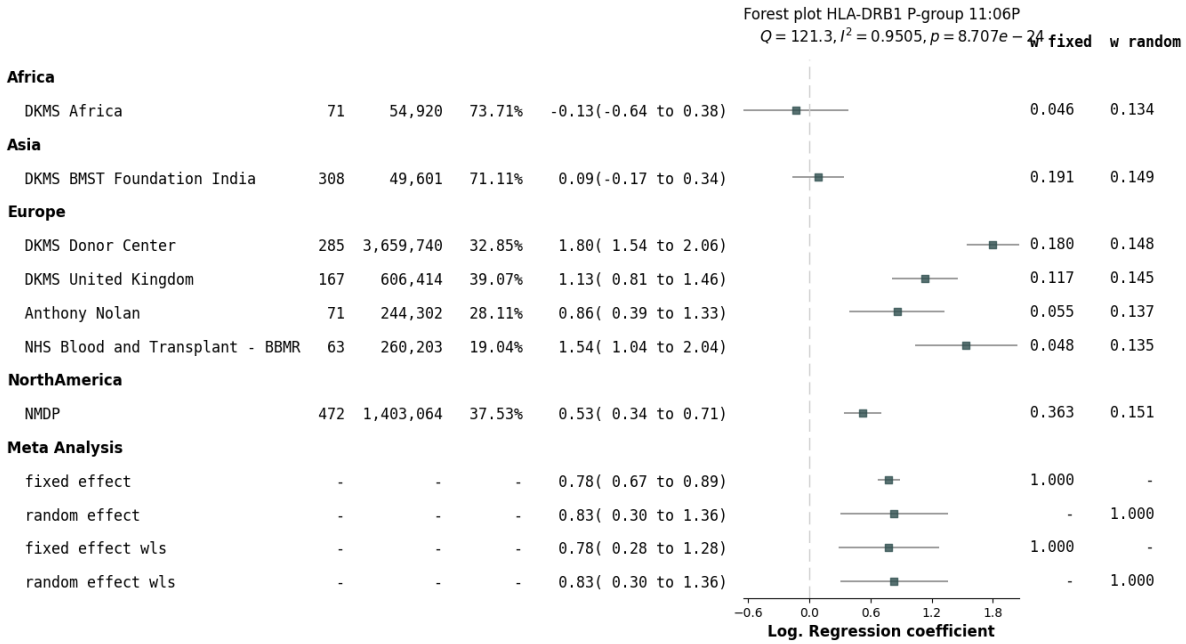


Figure A.32: Forest plot for the meta analysis of 11:06P on locus DRB1 with $OR = 2.67$, $p = 1.86 \times 10^{-88}$ and $N = 1\,685$.

A.4. Calibration Plots per Classifier

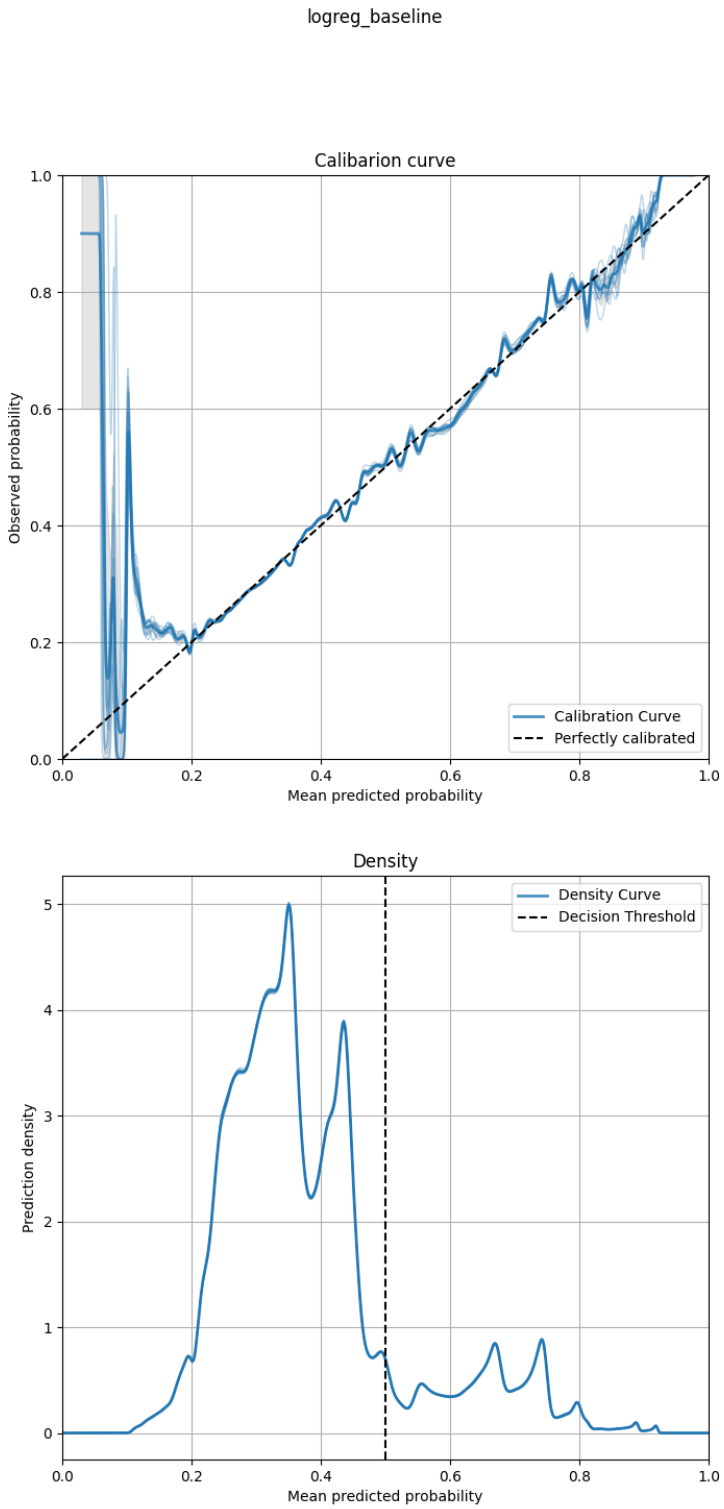


Figure A.33: Calibration and density plot for the Logistic Regression Baseline.

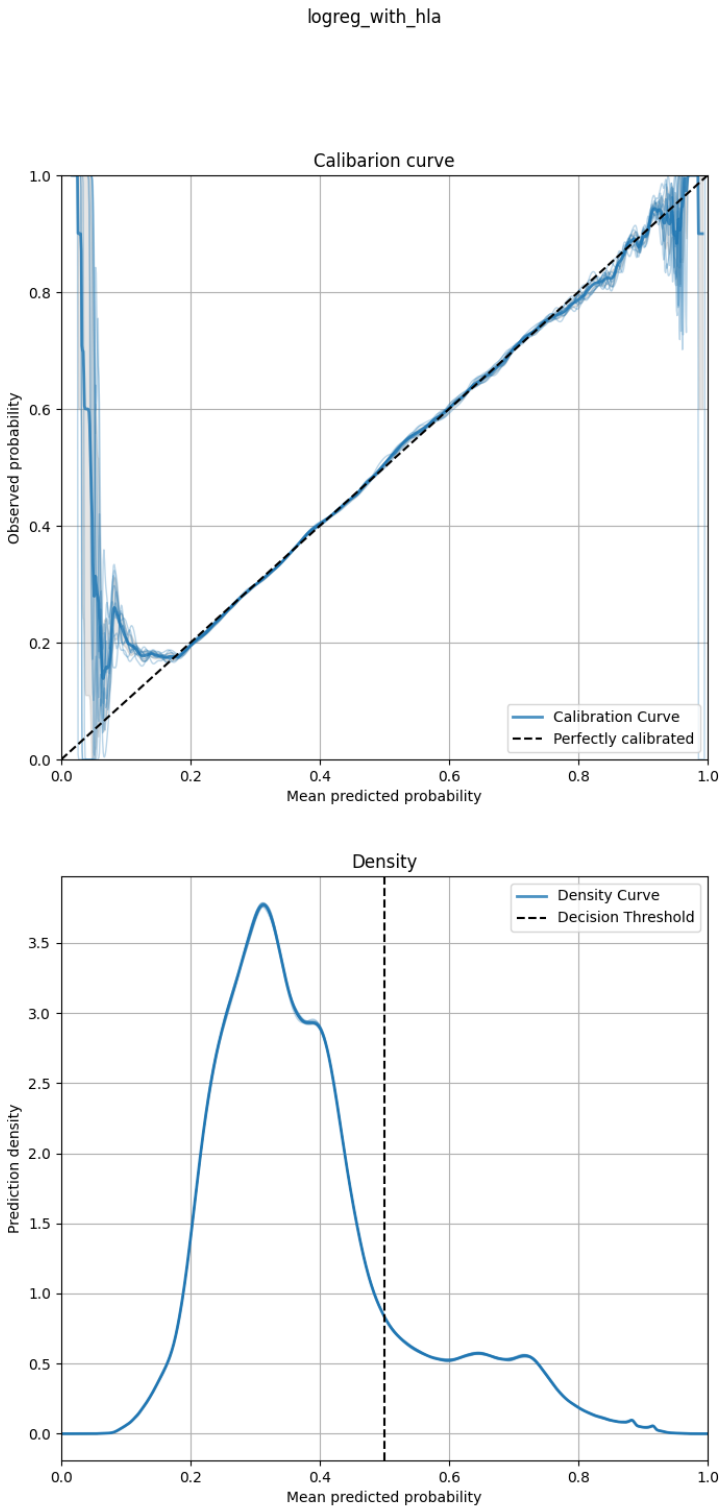


Figure A.34: Calibration and density plot for the Logistic Regression with HLA features.

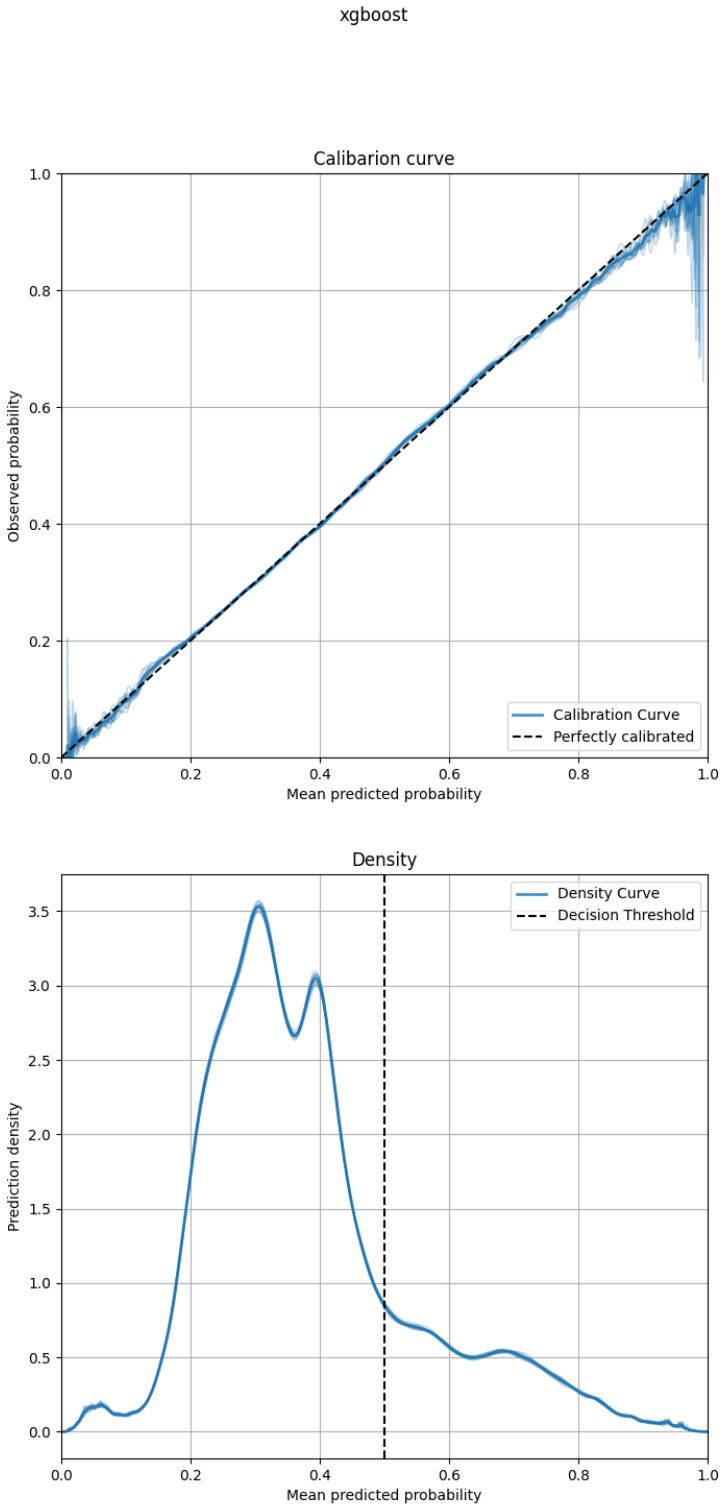


Figure A.35: Calibration and density plot for the XGBoost with default hyper parameters.

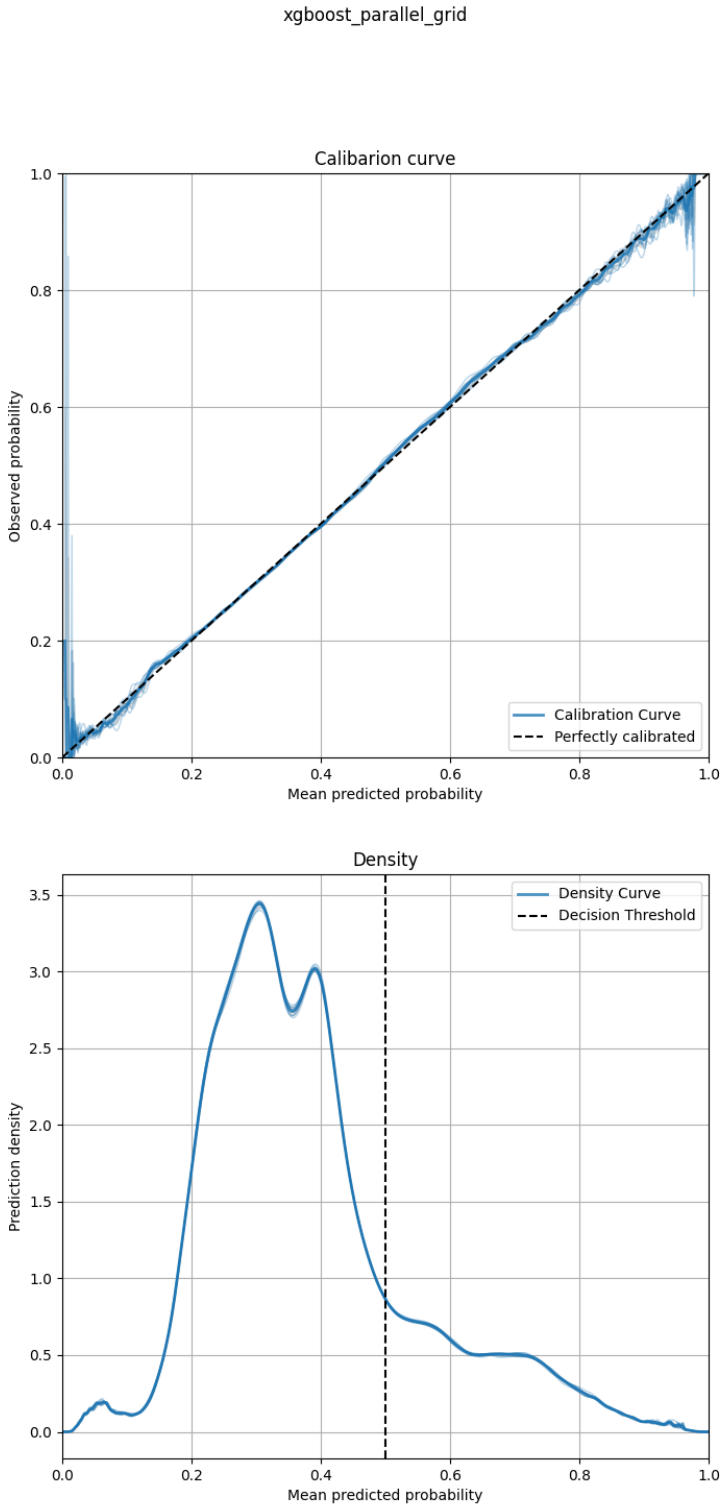


Figure A.36: Calibration and density plot for the XGBoost with grid trained hyper parameters.