EHESP

Master of Public Health

Master de Santé Publique

Machine Learning-Based Prediction of One Year Post-Transplant Outcomes in Allogeneic Hematopoietic Stem Cell Transplantation Patients with Malignant Hemopathies

---

< Fangchen XIA >

Class and year of the Master:

MPH 2

2023-2025

Location of the practicum:
Biomedicine de agency

Professional advisor:

Monia ZIDANE

Biomedicine de agency

Academic advisor:

Nolwenn LE MEUR

EHESP

Acknowledgements

I am grateful to the Agence de la biomédecine for providing the internship opportunity to make this study possible. I would like to sincerely thank Dr. Monia Zidane for her continuous support and guidance throughout both my internship and this thesis. Her contributions were essential to the project. Lastly, I would like to thank my school EHESP and my academic supervisor Prof. Nolwenn Le Meur for arranging and supporting this internship opportunity.

List of Acronyms

AI – Artificial Intelligence

ALL – Acute lymphoblastic leukemia

AML – Acute myeloid leukemia

ANC – Neutrophil recovery

Allo-HSCT – Allogeneic Hematopoietic Stem Cell Transplantation

AUC – Area Under the Curve

BART – Bayesian Additive Regression Trees

BMI – Body Mass Index

CMV – Cytomegalovirus

cGvHD – Chronic Graft-versus-Host Disease

DRI – Disease Risk Index

EBMT – European Society for Blood and Marrow Transplantation Registry

GBM – Gradient Boosting Machine

GvHD – Graft-versus-Host Disease

IMMISTepisode - Immunosuppression Treatment Episode

MAC – Myeloablative Conditioning

NRM – Non-Relapse Mortality

PB – Peripheral Blood

PLT – Platelet recovery

PRISMA – Preferred Reporting Items for Systematic Reviews and Meta-Analyses

SHAP – SHapley Additive exPlanations

SVM – Support Vector Machine

TBI – Total Body Irradiation

XGB – XGBoost

TABLE OF CONTENTS

Abstract 1
1. Introduction 2
2. Systematic Review 3
2.1 Methods 4
2.1.1 Eligibility Criteria 4
2.1.2 Search Strategy 4
2.1.3 Study Selection 4
2.4 Literature Review Results 5
2.4.1 Included Papers Characteristics 5
2.4.2 Extracted Data 5
3. Outcome Prediction 10
3.1 Methods 10
3.1.1 Data Source 10
3.1.2 Outcome Definition 10
3.1.3 Feature Used for The Modelling 10
3.1.4 Data Cleaning 11
3.1.5 Statistical Analysis 12
5. Results 13
5.1 Patient Characteristics 13
5.2 Outcomes 17
5.2.1 NRM 17
5.2.2 Rejection 19
5.2.3 Relapse 22
5.2.4 aGvHD 24
5.2.5 cGvHD 26
6. Discussion 29
7. Conclusion and Perspectives 31
8. References 33
9. Appendices 38
Abstract in French 49

Abstract

Background: Allogeneic hematopoietic stem cell transplantation (Allo-HSCT) is a curative treatment for hematological malignancies. However, patients remain at risk for serious complications, including non-relapse mortality (NRM), relapse, rejection, acute graft-versus-host disease (aGvHD), and chronic GVHD (cGvHD). Accurate early prediction of these outcomes can support clinical decision-making and improve long-term prognosis. This study uses pre-transplant data from 16,427 patients with malignancies recorded in the European Society for Blood and Marrow Transplantation registry (EBMT) between 2013 and 2023.

Objectives: This study aims to use machine learning algorithms to predict the probability of NRM, rejection, relapse, aGvHD and cGvHD in patients within one year after Allo-HSCT.

Methods: We developed and evaluated eight machine learning algorithms, including logistic regression, random forest, XGBoost (XGB), decision tree, elastic net, bayesian classifier, bayesian additive regression trees (BART) and a stacking ensemble model. Clinical data from patients receiving Allo-HSCT for malignant diseases were used. Model performance was assessed using Area under the curve (AUC) and accuracy. SHapley Additive exPlanations (SHAP) was applied to interpret the impact of individual features.

Results: The stacking model achieved the highest AUC across all outcomes, with the best performance in rejection (0.745), followed by relapse (0.735), NRM (0.732), aGVHD (0.700), and cGVHD (0.630). In terms of accuracy, stacking also ranked highest for cGVHD (0.615), aGVHD (0.613), and rejection (0.845), while the best-performing models for relapse and NRM were elastic net (0.704) and logistic regression (0.746), respectively. For rejection, the most influential features were treatment date, HLA matching, performance status, and conditioning regimen. For relapse, treatment date, DRI, conditioning regimen, and Thiotepa use were most important. For aGvHD, treatment date, donor type, and HLA matching ranked highest. For cGvHD, key features included treatment date, performance status, Thiotepa use, Cytomegalovirus (CMV) match, and HLA matching. For NRM, age at treatment, Thiotepa use, and treatment date were consistently top-ranking features.

Conclusion: The stacking model demonstrated the best overall performance. Among the five outcomes studied, treatment date consistently emerged as the most important feature. Machine learning shows strong potential as a supportive tool in clinical decision-making.

KEYWORDS: Allo-HSCT, Adult, Malignant Diseases, Machine Learning, Adverse Event Prediction

1. Introduction

Allogeneic hematopoietic stem cell transplantation (Allo-HSCT) is a pivotal treatment for high-risk hematologic malignancies, particularly refractory hematologic cancers such as Acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), high-risk myelodysplastic syndrome, and aggressive lymphomas [1]. This procedure involves transplanting healthy donor hematopoietic stem cells into the patient to replace the dysfunctional hematopoietic system. After intensive chemotherapy or other immunosuppressive conditioning regimens to eradicate malignant or abnormal clonal cells and interrupt disease pathogenesis in the recipient, either autologous or allogeneic hematopoietic stem cells are transplanted to rebuild normal hematopoiesis and immune function, thereby achieving therapeutic goals. Allo-HSCT was initially performed using bone marrow as the primary source of stem cells. It has since evolved to include peripheral blood stem cells and umbilical cord blood as alternative graft sources. [2].

But its clinical application still faces key challenges such as relapse, rejection, NRM and graft-versus-host disease (GvHD). Relapse after transplantation is mainly due to the immune escape or drug resistance of residual tumor cells [3], especially when the conditioning regimen fails to completely eliminate malignant cells or the graft-versus-tumor effect is insufficient [4]. NRM is the main cause of transplant-related death, with an incidence of 10-30% [5]. Although graft rejection is relatively rare, once it occurs, it will lead to transplant failure.The aGvHD represents one of the most prevalent and severe immune-mediated complications following HSCT. Typically manifesting within the first 100 days post-transplantation, aGvHD is primarily driven by donor T cells that recognize recipient Allo-antigens as foreign and subsequently trigger a cascade of inflammatory events [6]. The skin, liver, and gastrointestinal tract are the primary target organs [7], although involvement of critical sites such as the central nervous system can also occur [8]. Severity of clinical manifestations varies considerably. In severe cases, progression to multiple organ dysfunction syndrome markedly worsens prognosis and can be fatal [9]. The cGvHD is a prevalent and severe long-term complication following HSCT, typically manifesting beyond 100 days post-transplant. It involves multiple organs, commonly affecting the skin, oral mucosa, eyes, liver, lungs, and gastrointestinal tract [10]. cGvHD leads to impaired quality of life.

Outcome prediction for Allo-HSCT remains a highly complex and multidimensional task due to the wide range of interrelated factors involved. Patient-related characteristics such as age [11, 12, 61, 62], comorbidities [13, 14], performance status [15], disease type [16], and disease stage [17] significantly influence post-transplant outcomes. Donor-related variables, including

donor age [18], and the source of stem cells (e.g., peripheral blood, bone marrow, or cord blood) [19, 20] also play a critical role. Moreover, compatibility between donor and recipient—such as CMV serostatus matching [21, 22], HLA matching [23-25], gender matching [26, 27], and donor relationship [28]—further affects transplant success and complications. Treatment-related factors add another layer of complexity, including the use of total body irradiation (TBI) [29], the type of conditioning regimen [30], the specific agents used in the preparative regimen (e.g., Fludarabine or Thiotepa) [31-33], and prior immunosuppressive therapy [34]. The intricate interactions among these variables make outcome prediction especially challenging and underscore the need for robust, interpretable predictive models.

Machine learning, a pivotal branch of artificial intelligence, is capable of autonomously extracting meaningful patterns from complex datasets and generating accurate predictions. In hematological research and clinical practice, supervised machine learning has emerged as a powerful tool for early risk prediction of disease onset [35]. The most commonly employed supervised machine learning algorithms in medical prediction include classical approaches such as logistic regression, regularized linear regression, decision trees, and support vector machines (SVMs), as well as ensemble methods including boosting (e.g., XGB, LightGBM), bagging (e.g., random forest), and stacking algorithm. Recently, there is a growing adoption of Bayesian additive regression trees (BART) in public health. These methodologies are increasingly being integrated into hematological risk stratification systems to improve prognostic accuracy. To enhance the interpretability of machine learning results, SHAP (SHapley Additive exPlanations), a method grounded in cooperative game theory, was employed to quantify the contribution of each feature to the model's predictions [36]. SHAP has become a widely adopted tool in the machine learning research community for understanding complex model.

The objective of this study was to develop predictive models for five key one-year outcomes following first allogeneic HSCT: NRM, Relapse, Rejection, aGvHD, and cGvHD. Data for this study were obtained from the EBMT and analyzed on the basis of outcomes of systematic literature review.

2. Systematic Review

The systematic literature review was conducted across three major databases: PubMed, IEEE Xplore, and Scopus following the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) [37] guidelines to identify relevant literature on the application of machine learning and artificial intelligence (AI) in predicting relapse, rejection, or NRM, cGvHD and aGvHD following Allo-HSCT from January 1, 2020 and February 17, 2025.

2.1 Methods

2.1.1 Eligibility Criteria

Studies were eligible for inclusion if they met the following criteria:

1. Studies must be related with Allo-HSCT.
2. Studies that applied machine learning or AI techniques in predicting the following clinical outcomes: relapse, rejection, mortality, and death.
3. Studies must have involved human patients exclusively, excluding non-human or animal studies.
4. Studies published in English between January 1, 2020 and February 17, 2025

The following exclusion criteria were applied to ensure focus and consistency in the research theme:

1. Review articles were excluded.
2. Studies that lacked full-text content were excluded.
3. Studies that only used traditional statistical methods such as logistic regression for data analysis were excluded.
4. Studies related to autologous hematopoietic stem cell transplantation were excluded.
5. Studies focusing on complications unrelated to the five clinical outcomes of interest were excluded.

2.1.2 Search Strategy

Three search commands were employed using the terms listed in Appendix 1 to retrieve primary studies relevant to the research question in interest. The search terms included various synonyms and related concepts for Allo-HSCT, such as “GVHD”, “aGvHD”, “umbilical cord blood transplantation”, and “bone marrow transplant”. These were combined using Boolean operators with outcome-related terms including “relapse”, “rejection”, “GvHD”, and “death”, as well as methodology-related terms such as “machine learning” and “artificial intelligence”. Filters were applied where available to limit search results in journal articles published in English.

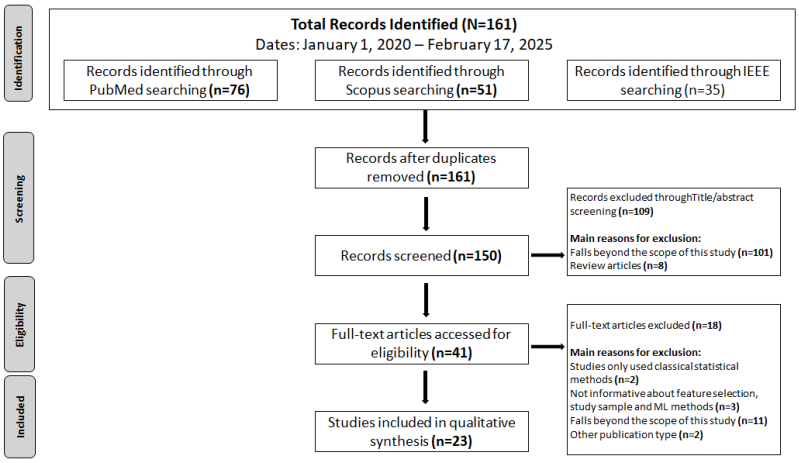
2.1.3 Study Selection

The PRISMA workflow illustrating the systematic identification, screening, eligibility assessment, and inclusion of scientific literature is presented in Figure 1.

We initially identified 76 relevant articles from PubMed, 51 from Scopus, and 35 from IEEE Xplore. After merging the search results and removing 10 duplicates, a total of 151 unique records were screened based on title and abstract. Of these, 109 articles were excluded. The

remaining 42 articles underwent full-text review. Based on predefined eligibility criteria, an additional 14 articles were excluded. Ultimately, 27 studies were included for final evaluation.

Figure 1: PRISMA Flow Diagram for Systematic Identification of Scientific Literature.



After screening, eligibility assessment, and removal of duplicates, a total of 23 studies were included and their key information, including the article's outcomes, applied algorithms, best machine learning methods, evaluation metrics, features, and sample details, was extracted and recorded from each study.

2.4 Literature Review Results

2.4.1 Included Papers Characteristics

Among 23 papers, 10 focused exclusively on paediatric populations, 6 investigated adult patients (aged ≥18 years), and 7 included both children and adults. In terms of outcomes, most of the papers focused on a single clinical outcome, though a few examined multiple outcomes simultaneously. Overall mortality was the most commonly studied outcome (18 studies). Four studies specifically targeted NRM, while one study examined mortality in patients who did not experience cGvHD, relapse, or graft rejection. Additionally, relapse was evaluated in five studies, and graft-versus-host disease (GvHD) in two (Appendix 2).

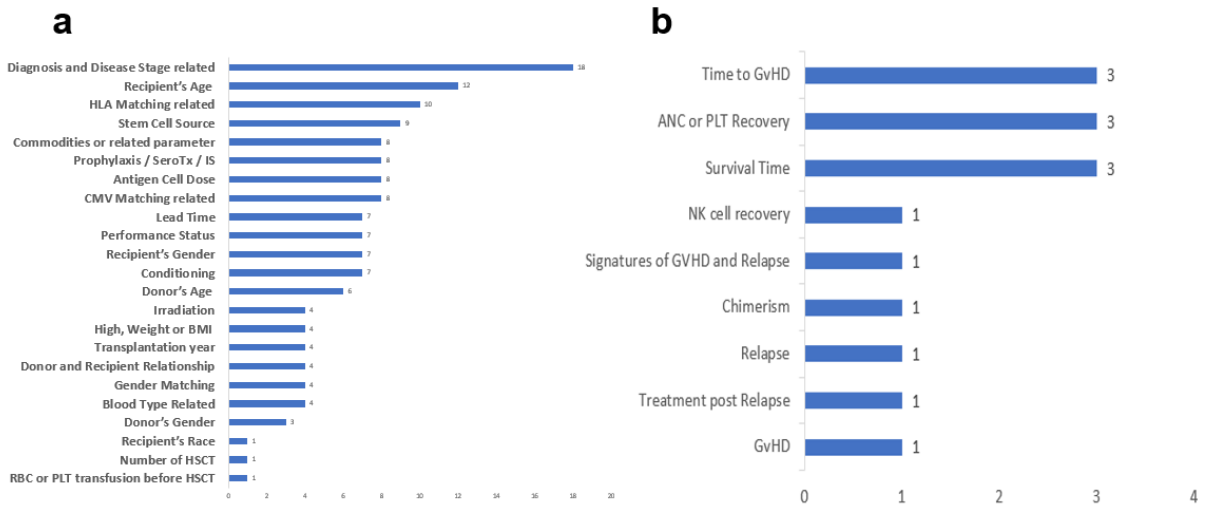
2.4.2 Extracted Data

Features

Among the 23 studies reviewed, one study (Gourisaria et al., 2023) selected 11 features from a broader set of variables but did not provide sufficient detail, therefore, this study was excluded from feature synthesis. Because of the inconsistencies in feature labelling across studies, we standardized the terminology and grouped the them into 34 unique features, which

can be classified as 25 pre-transplant and 9 post-transplant features based on the time of data collection. Notably, pre-transplant features predominated across the studies. As shown in the Figure 2, we can observe that the most frequently reported were diagnosis and disease stage (n = 18), HLA typing (n = 12), and CMV status (n = 10). More than half of the pre-transplant features were cited in over four studies. In contrast, post-transplant features were infrequently included, with survival time and ANC or platelet recovery being the most common—each appearing in only four studies.

Figure 2: The Count of Features Listed in Systematic Reviews (a) Pre-transplant Features, (b) Post-transplant Features



Algorithms

All included studies employed supervised learning approaches. In total, 22 distinct machine learning algorithms were identified across the 23 studies. The corresponding predicted outcomes, algorithms used, and their reported performance metrics are summarized in Table 1. Among them, 4 algorithms were specifically designed for survival analysis (cox proportional hazards model, fine and gray competing risks model, nonparametric failure time bayesian additive regression trees, and random survival forest), while the remaining 18 algorithms were applied for regression or classification tasks.

Table 1. Summary of Reviewed Studies Information based on Machine Learning Techniques and Other Details.

Reference	Outcomes	Tested Machine Learning Techniques	Best Machine Learning Technique

von Asmuth et al., 2023 [38]	1-year, 2-years and 5-years overall survival	XGB, Random forest	XGB ([AUC] 0.728, 0.716 and 0.694)
Wang et al., 2023 [39]	Early mortality	Logistic regression, XGB, LASSO, Random forest	Logistic regression ([AUC] 0.739)
Afanaseva et al., 2023 [40]	Relapse	GBM, Logistic regression, SVM, Random forest	GBM ([AUC] 0.91)
Qu et al., 2025 [41]	Overall survival	XGB	No comparison
Shourabizadeh et al., 2024 [42]	100-days survival	RF, Logistic regression, XGB, Decision tree	Random forest ([AUC] 0.71)
Lee et al., 2022 [43]	VOD/SOS, 100-days survival	XGB, Logistic regression, Decision tree, Adaboost, Bayesian classifier	XGB ([AUC] VOD/SOS: 0.740; Early Death: 0.746)
Jo et al., 2024 [44]	Relapse	Random forest ([Accuracy] ALL: 0.85; AML: 0.81)	No comparison
Alawneh & Hasasneh, 2024 [45]	Mortality	RF, XGB, GBM, Decision tree, Adaboost, KNN	RF([AUC] 0.969)
Ratul et al., 2022 [46]	Overall survival	Logistic regression, XGB, GBM, Decision tree, RF, AdaBoost,	Logistic regression ([AUC] 0.947)
Chadaga et al., 2023 [47]	Overall survival	Stacking, Logistic regression, XGB, Decision tree, RF, Adaboost, Catboost,	Stacking ([AUC] 0.91)
Short et al., 2024 [48]	Relapse	XGB, Logistic regression, Decision tree, SVM, RF, Bayesian classifier, Neural Networks	XGB ([AUC] 0.902)
Marvin & Alam, 2022 [49]	Overall survival	XGB, Catboost, Light GBM	XGB ([AUC] 0.944)
Rifat et al., 2023 [50]	Overall survival	Decision tree, XGB, GBM, RF, Adaboost, KNN	Decision tree ([Accuracy] 0.968)
Hossain et al., 2022 [51]	Relapse, GVHD	RF, Decision tree, Extra Tree, KNN, Neural Networks	Random forest ([AUC] 0.973)
Gourisaria et al., 2023 [52]	Overall survival	Adaboost, GBM, Decision tree, RF, Catboost	Adaboost
Afanaseva et al., 2023 [53]	Overall survival	BCF	No comparison
Spellman et al., 2024 [54]	3-years overall survival and Event (cGvHD, Relapse and Rejection) - free survival	NFT BART	No comparison
Choi et al., 2022 [55]	Overall survival, Relapse, and	GBM, Logistic regression, RF, Adaboost, Neural Networks	GBM ([AUC] 0.788)

Okamura et al., 2021 [56]	Relapse-free survival 1-year overall survival, progression-free survival, relapse/progression, relapse-free survival	RSF	No comparison
Iwasaki et al., 2021 [57]	GVHD and Relapse-free survival	Stacking, Cox, Fine-Gray, RSF, XGB, GBM, Component-wise GBM, Neural Networks	Stacking ([C-Index] 0.023, 0.21, 0.044, 0.017 and 0.258)
Echecopar et al., 2024 [58]	1-year overall survival,	RSF	No comparison
McCurd et al., 2022 [59]	Overall survival	Decision tree	No comparison
Mussetti et al., 2023 [60]	Overall survival and Non-relapse survival	ElasticNet, Neural Networks, Logistic regression, SVM, RF, GBM	ElasticNet ([AUC] 0.64, 0.61)

Among the 23 studies included in our review, 16 compared the performance of multiple machine learning models, while the remaining seven employed only a single algorithm. For studies that evaluated several models, the AUC was used as the primary criterion for identifying the best-performing algorithm. In cases where AUC was not reported, accuracy and other metrics were considered. Based on these criteria, we recorded both the number of times each algorithm was tested and the number of times it was reported as the best-performing model in Table 2. In terms of frequency of application, the most commonly applied models were random forest (n = 14), XGB (n = 12), decision tree (n = 10), and logistic regression (n = 9). In terms of performance, XGB achieved the best in four studies, followed by random forest (n = 3), and logistic regression (n = 2). Although stacking was evaluated in only one study, and consequently identified as the best-performing algorithm only once, it substantially outperformed other machine learning algorithms in terms of AUC and accuracy in that study.

Table 2. Summary of Reviewed Studies Information based on machine learning count and performance.

Main Category	Machine Learning	Tested count (among 23 paper)	Number listed as best (among 16 papers)
Linear Logistic Model	Logistic regression	9	2
Linear Model (Regularization)	LASSO	2	0

	Elastic Net	1	1
Bagging Ensemble	RF	14	3
	Extra Tree	1	0
	BCF	1	0
Boosting Ensemble	XGB	12	4
	GBM	8	2
	Catboost	3	0
	Adaboost	7	1
	Light GBM	1	0
	Component-wise Gradient Boosting	1	0
Stacking Ensemble	Stacking	2	2
Single Tree Model	Decision tree	10	1
Probabilistic Model	Bayesian classifier	2	0
Neural Networks	Neural Networks	5	0
SVM	SVM	4	0
KNN	KNN	4	0

To enhance clarity and comparability, the 19 machine learning algorithms identified in the review were organized into broader categories reflecting shared methodological characteristics. Tree-based models were divided into single-tree model (decision tree), and ensemble tree models. Ensemble tree models were further classified into three types: bagging methods (e.g., random forest, extra trees, BCF), boosting methods (e.g., XGB, GBM, CatBoost, AdaBoost, LightGBM, Component-wise GB), and stacking method (Stacking algorithm).

Logistic regression was placed in its own category of linear logistic model. Regularized variants such as LASSO and Elastic net were grouped as regularized linear models. Bayesian classifier was classified as a probabilistic model. Neural networks, SVM and KNN were assigned to individual categories due to their distinct mechanisms.

Among the ten main algorithm groups, only three—Linear Models with Regularization, Bagging Ensembles, and Boosting Ensembles—incorporated more than one machine learning method. Within the linear regularization group, LASSO was used in two studies but was not identified as the best-performing model in any. In contrast, Elastic net was used in only one study yet was reported as the top performer. Therefore, Elastic net outperformed LASSO within this category.

In the Bagging Ensemble group, random forest is the best. Because it was the most frequently used and was identified as the best-performing algorithm in three studies. In the Boosting Ensemble group, XGB was used in 12 studies and ranked best in four, outperforming GBM

and CatBoost. Its strong performance and low computational cost support XGB as the best method in this category.

3. Outcome Prediction

3.1 Methods

3.1.1 Data Source

This study analyzed data from all patients (n = 16,427) who received their first Allo-HSCT at any center across France between January 1, 2013, and December 31, 2023. We included all adult patients (>=18 years old) with malignant diseases, including acute myeloid and lymphoid leukemia, chronic myeloid and lymphoid leukemia, lymphoma, myelodysplastic syndrome, myelodysplastic/myeloproliferative disease, myeloproliferative disorder and plasma cell neoplasm, from the EBMT registry. Patients' demographic, clinical, and genotypic information was extracted for analysis.

3.1.2 Outcome Definition

- NRM was defined as death within one year following transplantation in the absence of relapse during that period.
- Relapse was defined as the occurrence of disease recurrence within one year following transplantation.
- Rejection was defined as the absence of ANC or PLT recovery, absence of graft loss, and no second Allo-HSCT performed within one year after the initial transplant.
- aGvHD and cGvHD outcomes were defined as the occurrence of acute or chronic GVHD, respectively, within one year post-transplantation.

3.1.3 Feature Used for The Modelling

Based on the findings from the systematic review, all EBMT available pre-transplant features were included in model development. To avoid the confounding effect of competing risks, features that are only measurable during the post-transplant period were excluded.

We used the Disease Risk Index (DRI) to represent disease type and stage. Developed by Armand et al., the DRI is a well-validated comprehensive index that is independent of patient age, donor source, and comorbidities. By classifying patients into four distinct risk groups, the DRI is particularly well-suited for risk stratification for patients undergoing Allo-HSCT. In the systematic review, studies by Lee et al. (2022), Okamura et al. (2021), and Qu et al. (2025) employed the DRI to replace the disease diagnosis and disease stage. Based on the principal of the DRI, patients were categorized into four risk levels: low, intermediate, high, and very high. Multiple studies indicating that patients aged ≥55 years tend to experience poorer

outcomes following HSCT [61, 62]. Therefore, we chose age at 55 as one cut-off point and recipient age feature was grouped into three intervals: 18–30, 30–55, and ≥55 years. In terms of HLA matching, given that fully matched sibling donors are considered the gold standard [63] and are generally preferred over fully matched unrelated donors, we categorized them separately to distinguish fully matched siblings from fully matched non-sibling donors. Numerous studies have suggested that haploidentical transplantation is a viable option when conventional fully matched donor is not available, cause the haploidentical shares the similar outcomes with unrelated match [64, 65]. Passweg et al. analyzed data from the EBMT activity survey report and reported a significant increase in the use of haploidentical transplants since 2012 [66]. In light of this trend and to better capture the clinical nuances of donor compatibility, we categorized HLA matching into five groups: fully matched sibling donors, fully matched unrelated donors, single-allele mismatches (9/10), haploidentical mismatches, and other mismatches. Body Mass Index (BMI) was categorized based on the World Health Organization (WHO) definition [67], with values between 18.5 and 24.9 kg/m² considered normal, and all other values classified as abnormal. We categorized CMV matching into two groups: R−/D− and all other combinations. This classification is based on the widely accepted clinical practice that a CMV-seronegative donor for a CMV-seronegative recipient is associated with the best outcomes [18, 22, 68]. Donor type feature was categorized as related donor and unrelated donor. For gender matching, we categorized donor-recipient combinations into two groups: female-to-male and all other combinations. This classification was informed by findings from a 2021 internal survey conducted by the Pediatric Diseases Working Party of the EBMT, which revealed a prevalent clinical preference for selecting male donors, especially when the recipient is male [69].

We recorded the prophylaxis as a binary variable (yes or no). Pretransplant immunosuppressive treatment episodes were classified as never, once, or twice. Conditioning regimen was categorized as either myeloablative conditioning or reduced-intensity conditioning, and the presence or absence of TBI was also recorded. Performance status was categorized as ≤80, 90 or 100. Lead time, defined as the time from diagnosis to transplantation, was grouped into ≤6 months, 6–12 months, or >12 months. Due to compliance regulations, recipient race data are not recorded in the EBMT registry.

3.1.4 Data Cleaning

In this study, features with more than 20% missing values were excluded from the analysis. These included variables such as blood type, pre-transplant RBC and platelet transfusion, and antigen-specific cell dose metrics (e.g., $CD34^+$ and $CD3^+$ cell counts), which were unavailable for the majority of the study population. For features with less than 20% missing data,

imputation was performed to retain their predictive value in the model development. Appendix 3 summarizes the number and percentage of missing values for the features that met the inclusion threshold (i.e., missingness < 20%).

Overall, most variables had minimal missingness (less than 1%), such as recipient gender (0.01%), donor type (0.05%), and prophylaxis (0.07%). However, a few variables had relatively higher rates of missing data, including disease stage (4.14%) and performance status (7.69%). All data cleaning procedures were conducted using R version 4.4.3. Missing data were imputed using the 'mice' package. The dataset was randomly split into a training set (85%) and a validation set (15%).

3.1.5 Statistical Analysis

Logistic regression was performed using the stats package (0.1.0 version), while the elastic net model was fitted with the glmnet package (4.1-9 version). The bayesian classifier model was built using the e1071 package (1.7-16 version). Random Forest was implemented via the tuneRanger package (0.7 version). XGB was fitted using the xgboost package (1.7.11.1 version). The BART model was implemented with the dbarts package (0.9-32 version). XGB was implemented with 10-fold cross-validation. The key hyperparameters were set as follows: max_depth = 6, eta = 0.3, and the evaluation metric was AUC. Early stopping was applied when the performance did not improve after 10 consecutive rounds. The maximum number of boosting iterations was set to 100.

The random forest model consisted of 1,000 trees. Hyperparameter tuning was conducted in 70 iterations. The tuned parameters included mtry, min.node.size, and sample.fraction. For the elastic net model, the optimal penalty parameter was selected based on lambda.min. The decision tree model was trained with the rpart package. The hyperparameters minsplit (from 10 to 500, in steps of 10) and cp (0.001, 0.01, 0.1, 1) were tested. These parameters controlled the minimum number of samples for splitting and the cost-complexity of pruning, respectively. Class imbalance was adjusted by setting prior probabilities to 0.5 for each class. Model performance was evaluated by 10-fold cross-validation. For the BART model, the number of trees (ntree) and the prior parameter k were tuned. Three values were tested for each: 100, 200, 300 for ntree and 1.0, 2.0, 3.0 for k. The tuning process used 10-fold cross-validation.In this study, seven base learners—logistic regression, elastic net, XGB, random forest, bayesian classifier, decision tree, and BART—were trained using five-fold cross-validation with hyperparameter tuning. The meta-learner was constructed through logistic regression.

Eight algorithms were assessed using AUC and accuracy to evaluate the performance in predicting 1-year outcomes followed the Allo-HSCT. The optimal threshold for the confusion

matrix was selected based on the Youden index [70] instead of a fixed cutoff of 0.5. Sensitivity, and specificity were included as complementary measures to provide a more complete view of model behavior. To interpret model outputs, SHAP values were used to assess feature contributions. For Bayesian classifier, RF, and elastic net, SHAP values were computed using the iml package. For XGB, the SHAPforxgboost package was employed, and for BART, feature contributions were evaluated using the bartXViz package.

The summary of the selected hyperparameters for each algorithm is provided in Appendix 4. For the XGB models, the optimal AUC was achieved after 13 to 16 rounds of 10-fold cross-validation. For the random forests model, the mtry parameter generally ranged from 5 to 10, while the minimum node size (min.node.size) varied across outcomes—being smallest (24) for Rejection and largest (338) for Relapse. In the Elastic Net models, the optimal values of lambda.min were consistently low, with all values falling below 0.005. For the Decision tree model, the smallest minsplit value (10) was observed for Relapse, whereas the largest (210) was for aGvHD. In the case of BART model, the cGvHD model had the smallest number of trees (ntree = 100), while other models used 200 or 300 trees.

5. Results

5.1 Patient Characteristics

The summary of demographic characteristics of the included patients is shown in Table 3. Five distinct outcomes within the first year after allogeneic HSCT were analyzed. Death occurred in 25.0% of patients, while relapse, graft rejection, acute GVHD, and chronic GVHD were observed in 19.0%, 7.8%, 60.0%, and 38.9% of patients, respectively. Each outcome was assessed independently.

Over half of patients (52.1%) were older than 55 years at the time of treatment. Acute leukemia was the most common diagnosis (58.4%), followed by myelodysplastic syndrome (13.2%), lymphoma (11.8%), and myeloproliferative disorders (7.0%). Chronic leukemia, myelodysplastic/myeloproliferative disease and plasma cell neoplasms were each observed in less than 4% of patients. Over one third of patients received transplants within 6–12 months of diagnosis (35.9%), while 21.8% had lead times exceeding two years. Most patients had good performance status (62.3% with scores ≥90) and intermediate DRI assignment (51.6%), with only 3.8% classified as very high risk. Approximately 48.9% of patients had a normal BMI, while 32.8% were overweight and 14.5% obese. The HCT-comorbidity index was ≤2 in 72.9% of patients. Infectious complications occurred in 37.4%, whereas non-infectious complications were reported in 93.6%. Transplants were distributed across time periods, with 36.1% performed between 2013–2016, 35.8% between 2017–2020, and 28.1% between 2021–2023.

Reduced-intensity conditioning regimens were used in 65.6% of cases. Nearly all patients (99.5%) received prophylaxis, and total body irradiation was administered to 80.5%. The most commonly used conditioning agents were bendamustine (99.9%), fludarabine (85.1%), and busulfan (73.9%). Use of cyclophosphamide (29.4%), thiotepa (20.4%), melphalan (3.8%), and treosulfan (2.2%) was less frequent.

Table 3: Baseline Demographic Characteristics of the Study Cohort.

Variables	Number	Pourcentage
NRM		
Yes	2463	15.0%
No	13964	85.0%
Relapse		
Yes	3129	19.0%
No	13298	81.0%
Rejection		
Yes	1289	7.8%
No	15138	92.2%
aGvHD		
Yes	9833	60.0%
No	6594	40.0%
cGvHD		
Yes	6388	38.9%
No	10050	61.1%
Patient's Age at Treatment		
18~40	3303	20.1%
40~55	4559	27.8%
> 55	8565	52.1%

Main Diagnosis		
Acute leukemia	9589	58.4%
Chronic leukemia	589	3.6%
Lymphoma	1940	11.8%
Myelodysplastic syndrome	2159	13.2%
Myelodysplastic/myeloproliferative disease	583	3.6%
Myeloproliferative disorder	1155	7.0%
Plasma cell neoplasm	398	2.4%
Lead Time		
< 3 months	712	4.3%
3 to 6 months	5901	35.9%
7 to 12 months	3710	22.6%
12 months to 18 months	1566	9.53%
19 months to 24 months	950	5.78%
>2 years	3588	21.8%
Performance Status		
<= 80	6194	37.7%
>= 90	10234	62.3 %
DRI Assignment		
Low	854	5.2%
Intermediate	8470	51.6%
High	6477	39.4%
Very High	626	3.8%
BMI Category		
Normal Weight (18.5 to 24.9)	8037	48.9%
Underweight (below 18.5)	618	3.8%
Overweight (25.0 to 29.9)	5387	32.8%
Obesity (above 29.9)	2385	14.5%
HCT - Comorbidity Index		
0	2147	13.1%
1~2	9827	59.8%
3~5	3791	27.1%

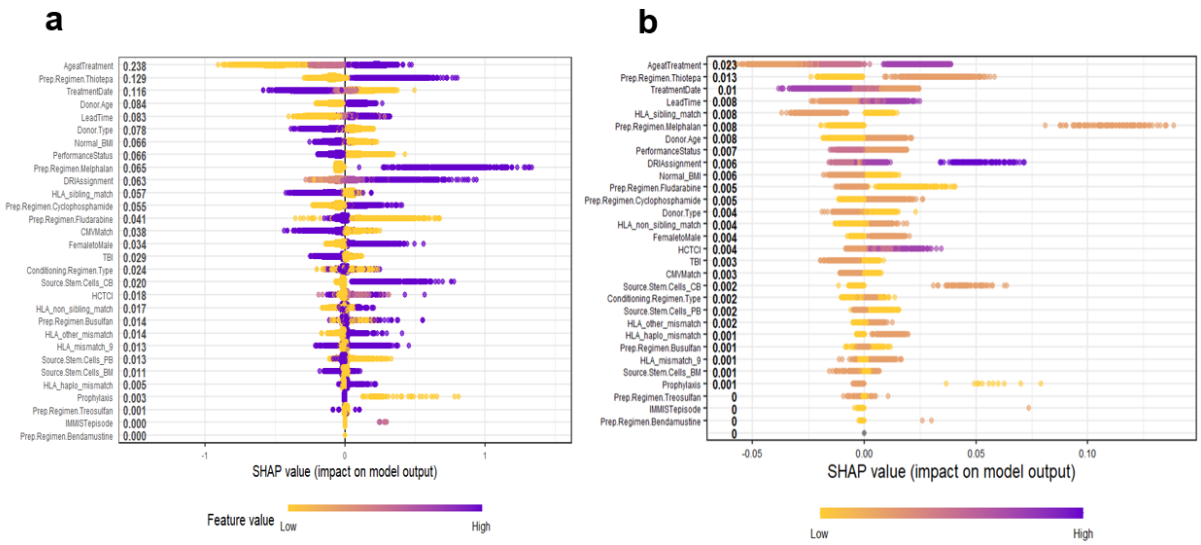
>5	662	4%
Treatment Period		
2013 - 2016	5935	36.1%
2017 - 2020	5882	35.8%
2021 - 2023	4610	28.1%
Source Stem Cell		
Peripheral blood	13905	84.6%
Cord blood	443	2.7%
Bone marrow	2079	12.7%
Conditioning Regimen		
Myeloablative conditioning regimen	5657	34.4%
Reduced intensity conditioning	10770	65.6%
Prophylaxis		
Yes	16353	99.5%
No	74	0.5%
Total Body Irradiation		
Yes	13218	80.5%
No	3209	19.5%
Preparation Regimen : Drug Used		
Bendamustine	16421	99.9%
Busulfan	12147	73.9%
Cyclophosphamide	4823	29.4%
Fludarabine	13973	85.1%
Melphalan	627	3.8%
Thiotepa	3354	20.4%
Treosulfan	354	2.2%

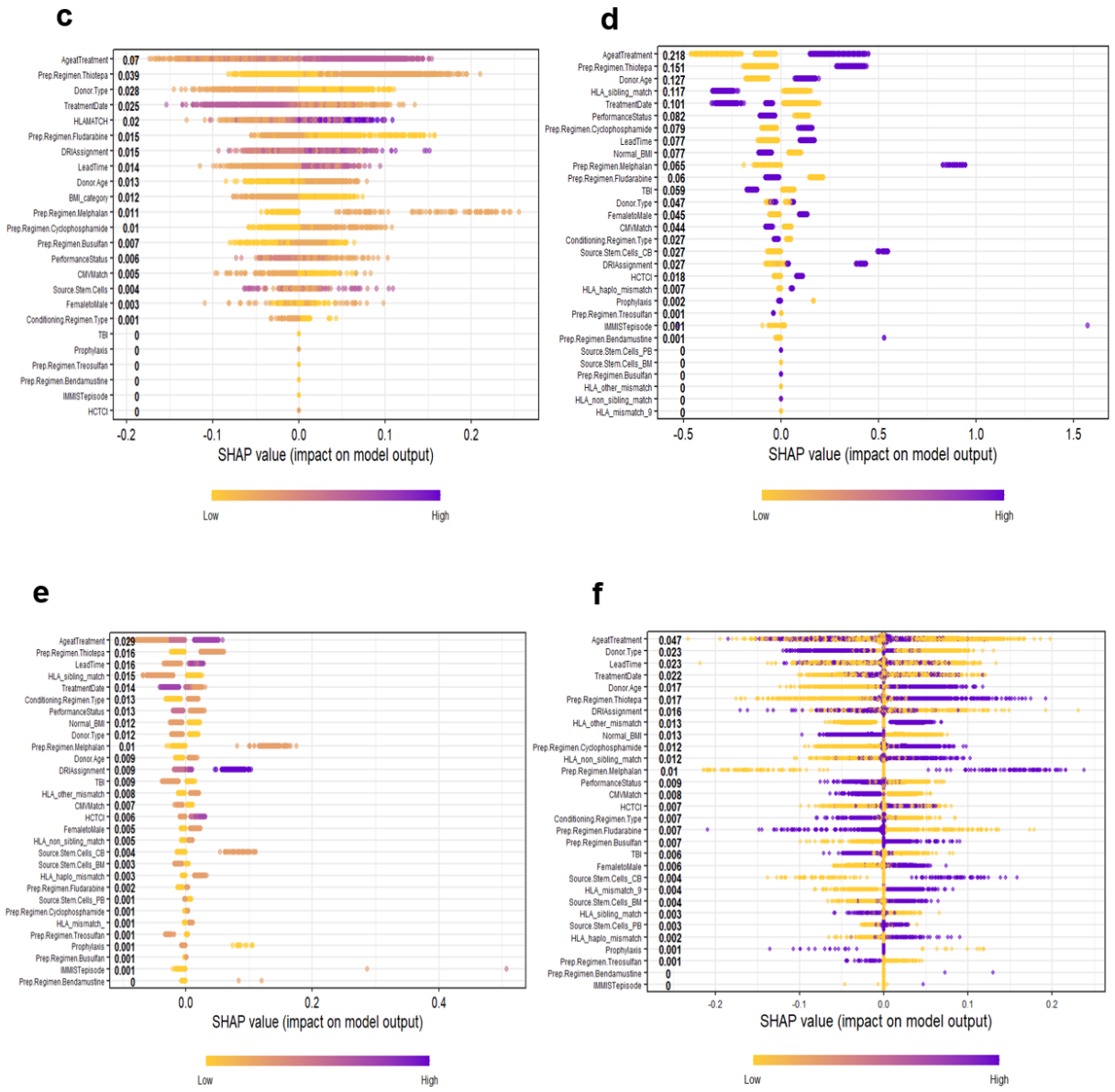
5.2 Outcomes

5.2.1 NRM

For NRM prediction, the stacking model achieved the highest AUC at 0.732, significantly outperforming all other models. BART ranked second with an AUC of 0.618, followed by the decision tree model with 0.610. The remaining models had similar AUC values, ranging narrowly between 0.59 and 0.60 (Appendix 5). In terms of sensitivity, stacking ranked first with a value of 0.742, while its specificity was relatively lower, ranking fourth at 0.614. Conversely, logistic regression achieved the highest specificity at 0.818, but this came at the cost of the lowest sensitivity, which was only 0.339. A similar trade-off was observed in the Decision tree model, which had the second-highest specificity (0.808) but the second-lowest sensitivity (0.375). Regarding accuracy, Logistic regression again outperformed all other models with a value of 0.746, notably higher than the rest. Elastic Net ranked second with an accuracy of 0.662, followed by stacking at 0.633.

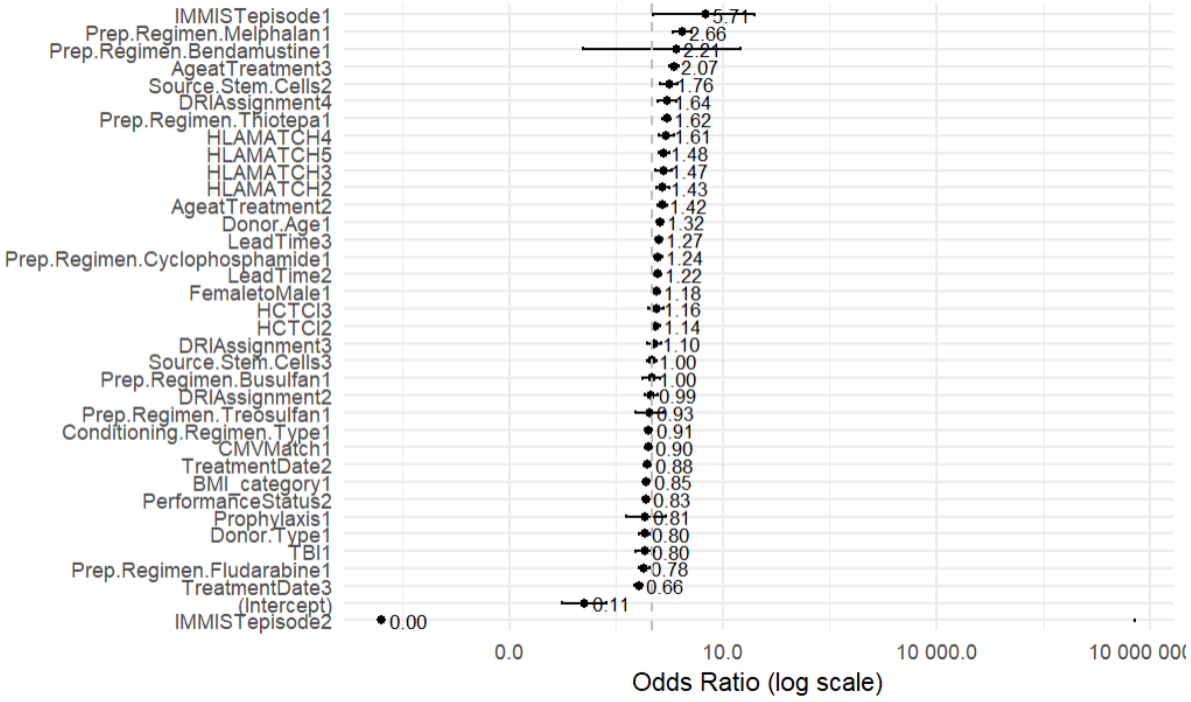
Figure 3: Summary SHAP value plots for NRM across six machine learning models: (a) XGBoost, (b) Random Forest, (c) Decision Tree, (d) Elastic Net, (e) Bayesian Classifier, and (f) BART. These plots illustrate the relative importance and direction of pre-transplant features across different models for predicting NRM.





Based on the results from the six SHAP-explainable models (Figure 3), age at treatment, Thiotepa use, and treatment date consistently ranked among the most important features. Notably, age at treatment was ranked as the most important feature across all six models. According to the beeswarm plots, older age was associated with a higher probability of NRM. Patients who received Thiotepa in the preparative regimen also had an increased risk of NRM. In contrast, more recent transplant dates were associated with a lower probability of NRM. Consistent with these findings, logistic regression analysis showed that, compared to patients aged 18–30, those aged 30–55 and over 55 had significantly higher odds of NRM (OR = 1.42 and 2.07, respectively; $p < 0.001$). The use of Thiotepa was also significantly associated with increased NRM risk (OR = 1.62, $p < 0.001$). With respect to transplant year, both the 2017–2020 (OR = 0.88, $p < 0.05$) and 2021–2023 (OR = 0.66, $p < 0.001$) periods were associated with significantly lower NRM compared to 2013–2016 (Figure 4).

Figure 4: Forest Plot for Predictors of NRM



5.2.2 Rejection

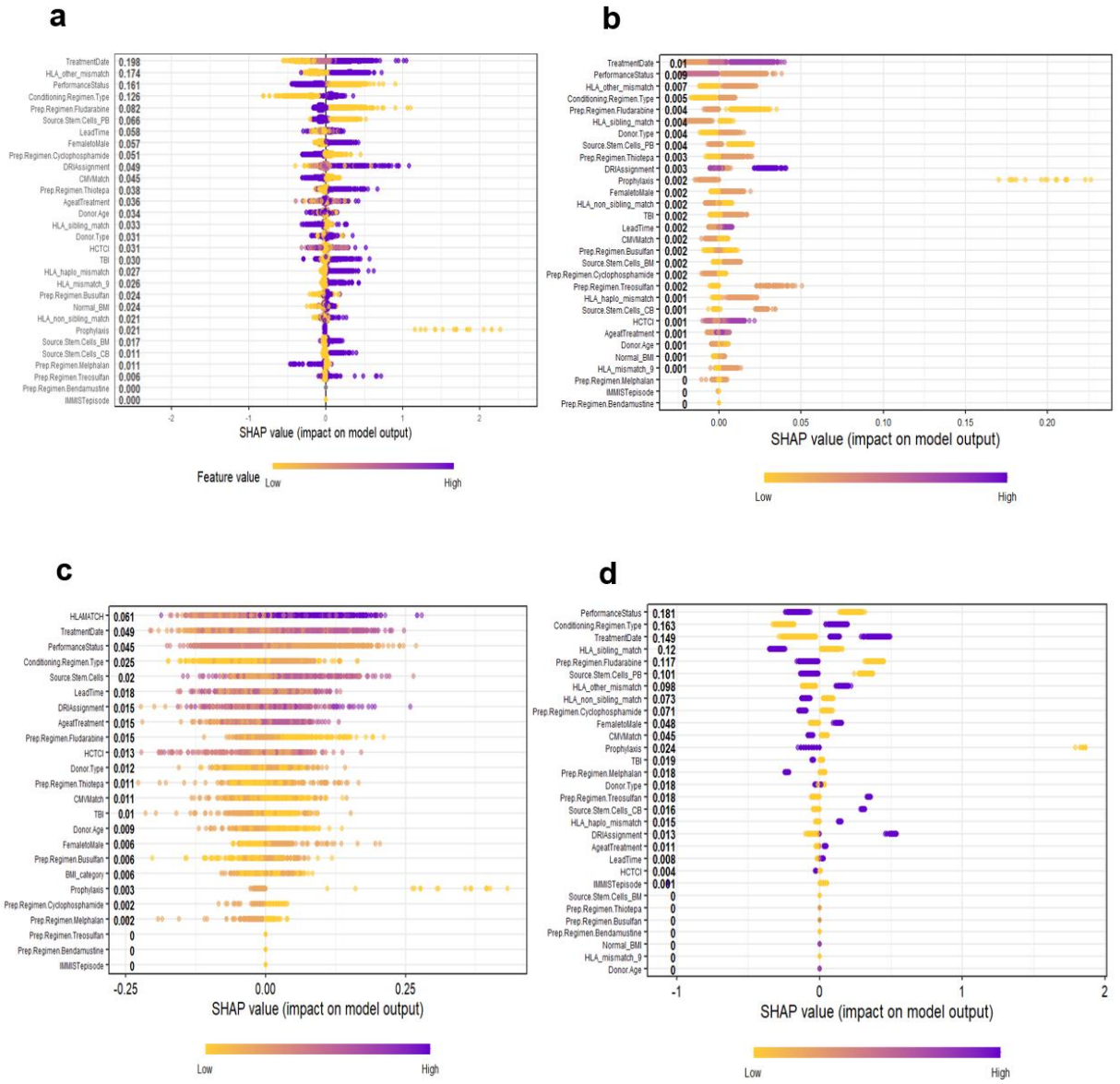
The performance of model prediction for rejection is shown in Appendix 6. The AUCs were similar across models, ranging narrowly from 0.658 (logistic regression) to 0.681 (BART). BART and random forest achieved the highest AUCs (0.681 and 0.680), followed by bagging classifier (0.671) and decision tree (0.662). XGB (0.660), elastic net (0.659), and logistic regression (0.658) showed slightly lower but comparable performance.

Regarding sensitivity and specificity, random forest demonstrated the highest sensitivity (0.80) but the lowest specificity (0.49). By contrast, BART showed the highest specificity (0.80) but the lowest sensitivity (0.478). XGB balanced a relatively high sensitivity (0.73) with moderate specificity (0.56). Other models presented more balanced sensitivity and specificity values. The stacking model outperformed all other models, with the highest AUC (0.745) and specificity (0.875), though its sensitivity remained moderate (0.494). The stacking model achieved the highest accuracy (0.845), while all other models yielded lower values. Only the elastic net reached an accuracy above 0.60 (0.620), but the rest remained below this threshold.

Based on the results from six SHAP-explainable models, we consistently observed that Treatment Date, HLA Matching, Performance Status, and Conditioning Regimen Type emerged as the most influential features in predicting rejection. Notably, Treatment Date was ranked as the most important feature in three out of the six models. As illustrated in the

beeswarm plots (Figure 5), more recent transplant dates were associated with a higher probability of rejection. Regarding HLA Matching, patients whose donor-recipient match was neither haploidentical nor 9/10 nor 10/10 exhibited a significantly higher risk of rejection. Higher Performance Status was linked to a lower likelihood of rejection. Additionally, patients receiving RIC regimens had a higher rejection risk compared to those who received myeloablative conditioning (MAC).

Figure 5: Summary SHAP value plots for Rejection across six machine learning models: (a) XGBoost, (b) Random Forest, (c) Decision Tree, (d) Elastic Net, (e) Bayesian Classifier, and (f) BART. These plots illustrate the relative importance and direction of pre-transplant features across different models for predicting Rejection.



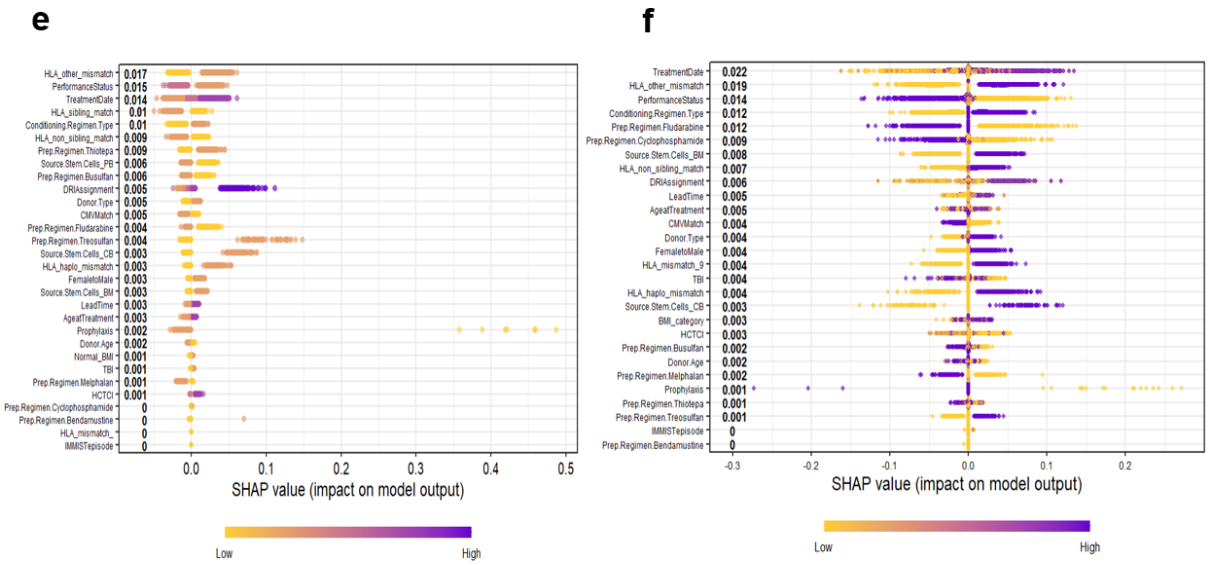
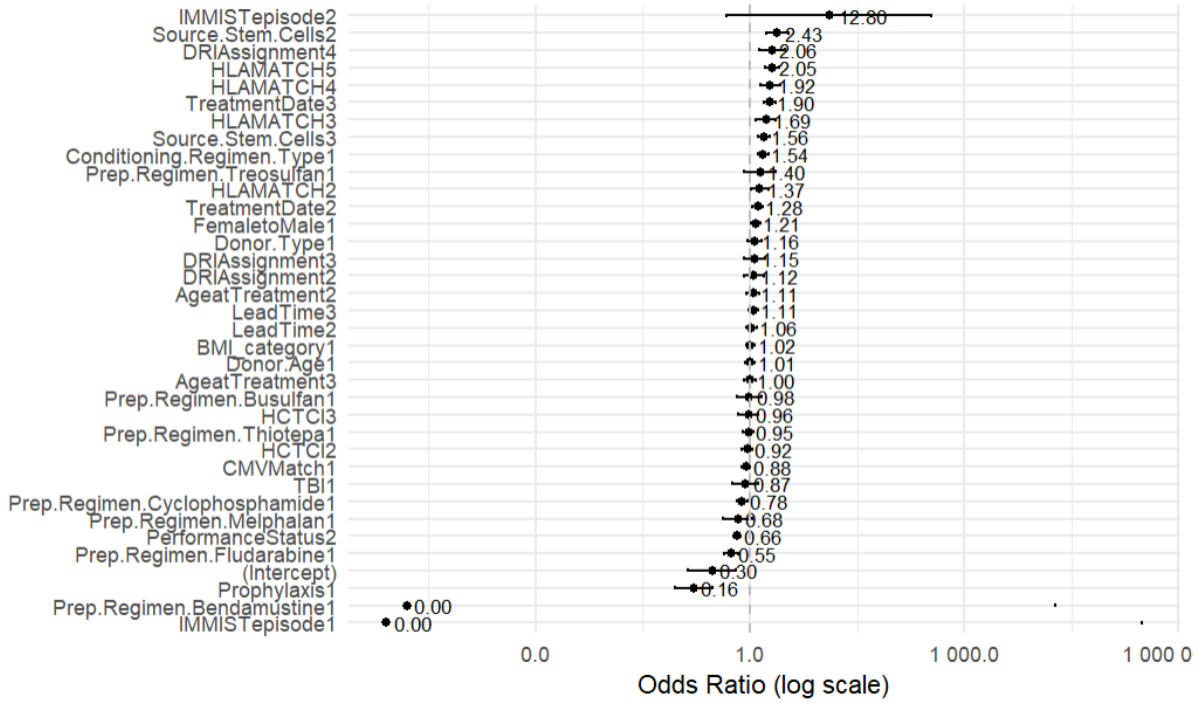


Figure 6: Forest Plot for Predictors of Rejection



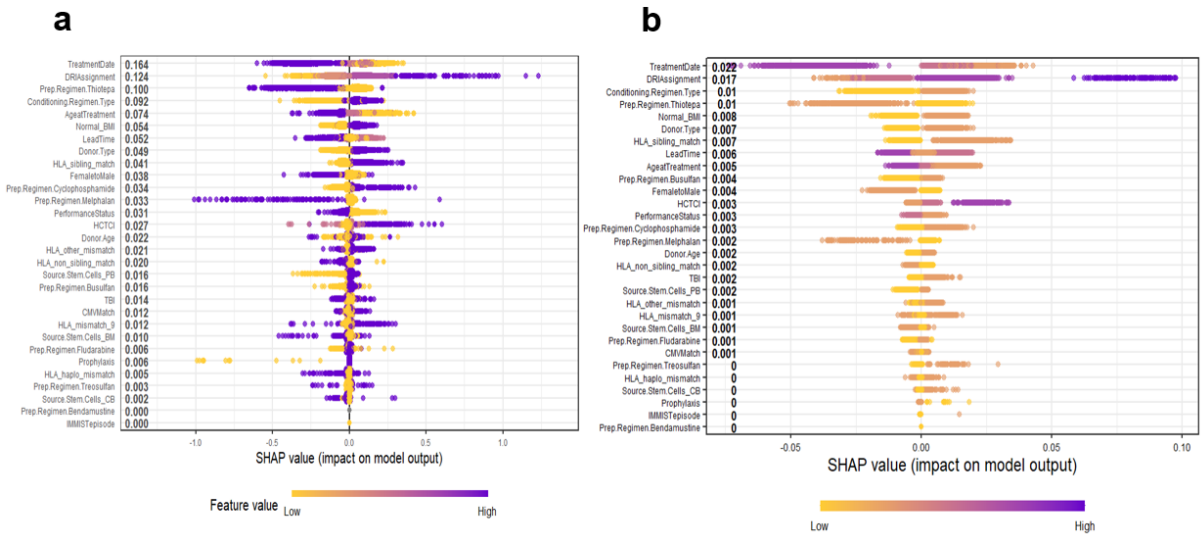
According to the logistic regression results (Figure 6), compared to transplants performed during 2013–2016, those conducted in 2017–2020 and 2021–2023 were associated with significantly higher odds of rejection, with ORs of 1.28 and 1.90, respectively (both $p < 0.001$). Similarly, for HLA matching, all non-sibling categories showed significantly higher odds of rejection compared to sibling matches, with all ORs greater than 1. Regarding conditioning regimen, patients receiving RIC had significantly higher odds of rejection compared to those receiving myeloablative conditioning MAC, with an OR of 1.54 ($p < 0.001$).

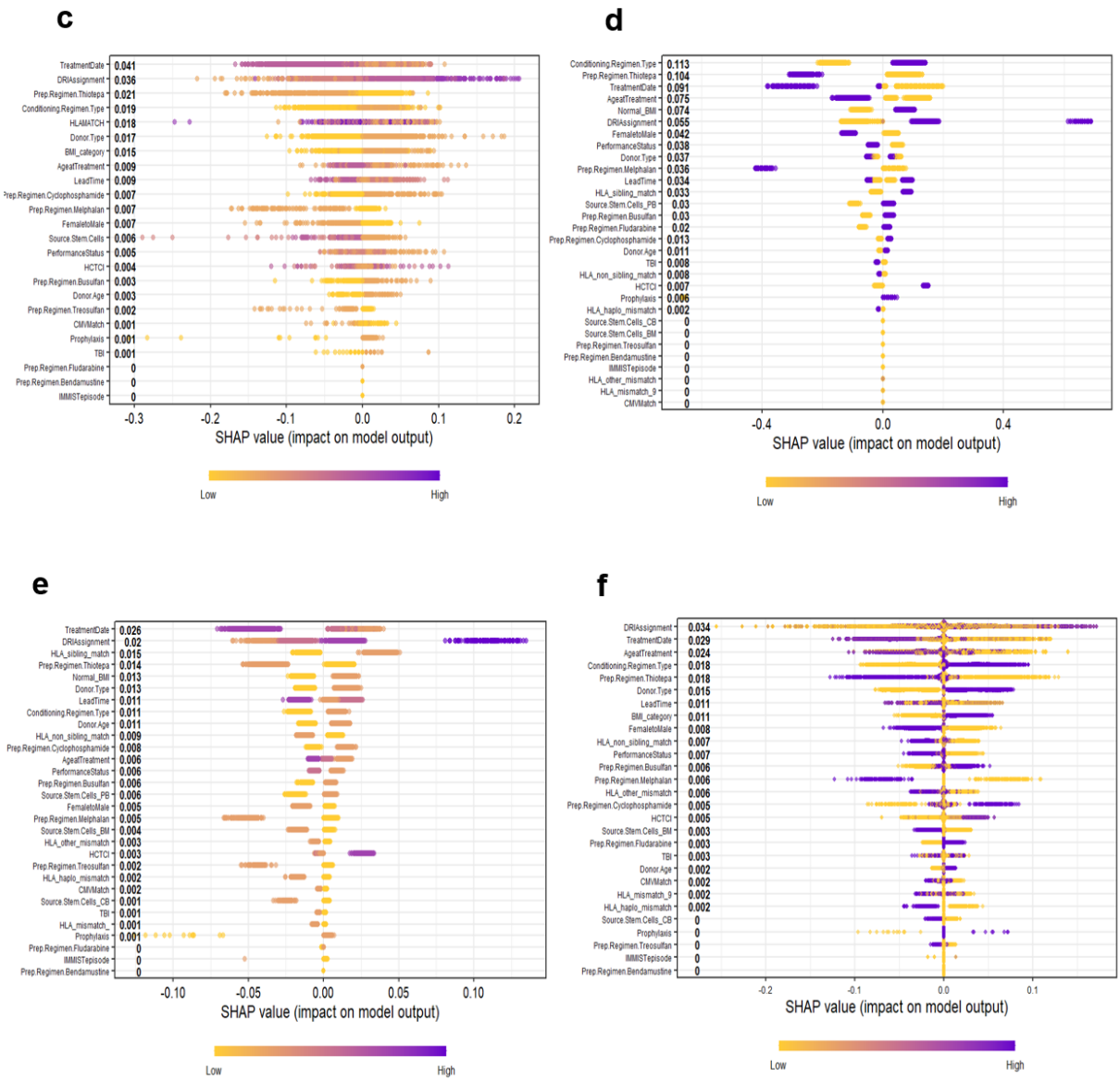
5.2.3 Relapse

The performance of the model performances is shown in Appendix 7. The stacking model achieved the highest AUC (0.735). It also had the highest sensitivity (0.734) and the third-highest specificity (0.615). XGB ranked second in AUC (0.660), with a sensitivity of 0.730 and a specificity of 0.559. Logistic regression, bagging classifier, elastic net, random forest and BART all had AUCs around 0.62, with limited variation. Logistic regression had the highest specificity (0.808), but its sensitivity was low (0.371). The decision tree recorded the lowest AUC (0.605), but maintained a relatively balanced sensitivity (0.623) and specificity (0.531). In contrast, accuracy rankings differed. Logistic regression and elastic net achieved the highest accuracies (0.725 and 0.704, respectively), while stacking ranked third (0.690). The bagging classifier showed the lowest accuracy (0.526).

Based on the results from six SHAP-explainable models (Figure 7), Treatment Date, DRI, Conditioning Regimen Type, and Thiotepa Use were consistently identified as high-importance features in predicting relapse. Notably, Treatment Date ranked among the top three features in all six models and was considered the most important variable in four of them.

Figure 7: Summary SHAP value plots for Relapse across six machine learning models: (a) XGB, (b) Random Forest, (c) Decision Tree, (d) Elastic Net, (e) Bayesian Classifier, and (f) BART. These plots illustrate the relative importance and direction of pre-transplant features across different models for predicting Relapse.

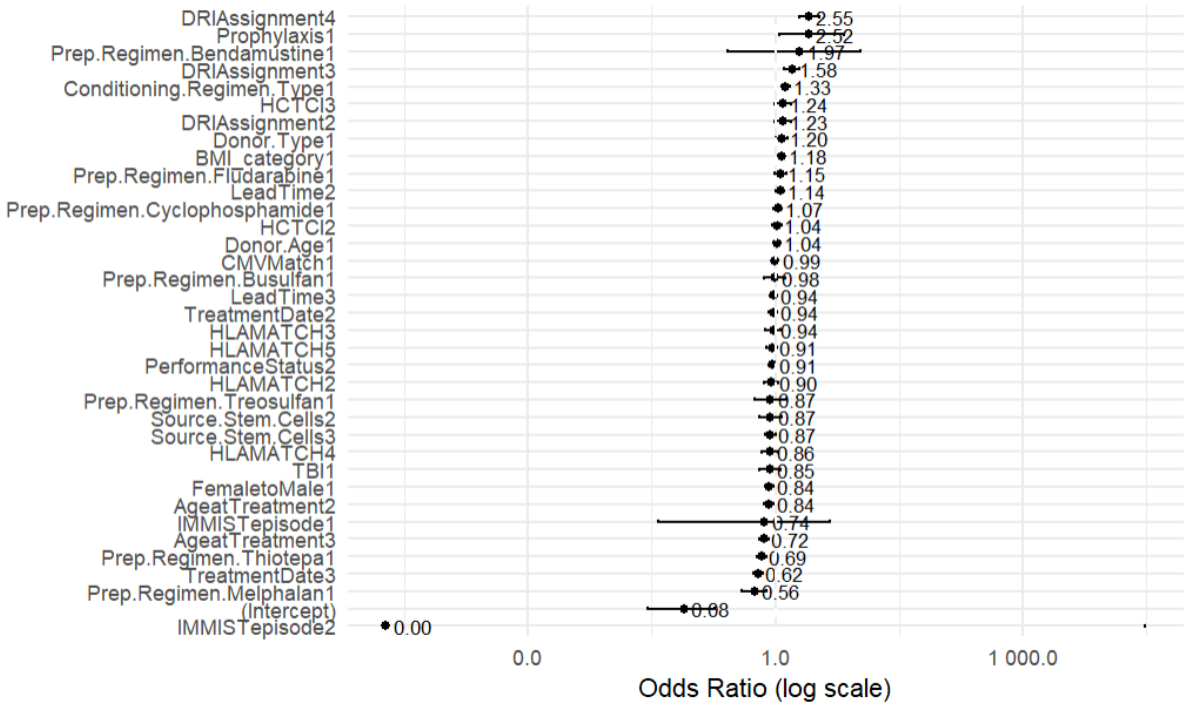




More recent transplant dates were associated with a lower risk of relapse. A higher DRI score was linked to an increased likelihood of relapse. Patients who received RIC had a higher probability of relapse compared to those who received MAC. Additionally, the use of Thiotepa in the preparative regimen was associated with a lower risk of relapse. These observations were in line with the logistic regression analysis results (Figure 8). Compared to transplants performed during 2013–2016 (reference group), those conducted in 2021–2023 were significantly associated with a reduced relapse risk ($p < 0.001$, $OR < 1$), while the 2017–2020 group also showed an $OR < 1$, although this result was not statistically significant at the 5% level.

All DRI categories had odds ratios greater than 1. However, only high-risk and very high-risk categories reached statistical significance at the $p < 0.001$ level. The intermediate-risk category did not show a significant association ($p > 0.05$).

Figure 8: Forest Plot for Predictors of Relapse



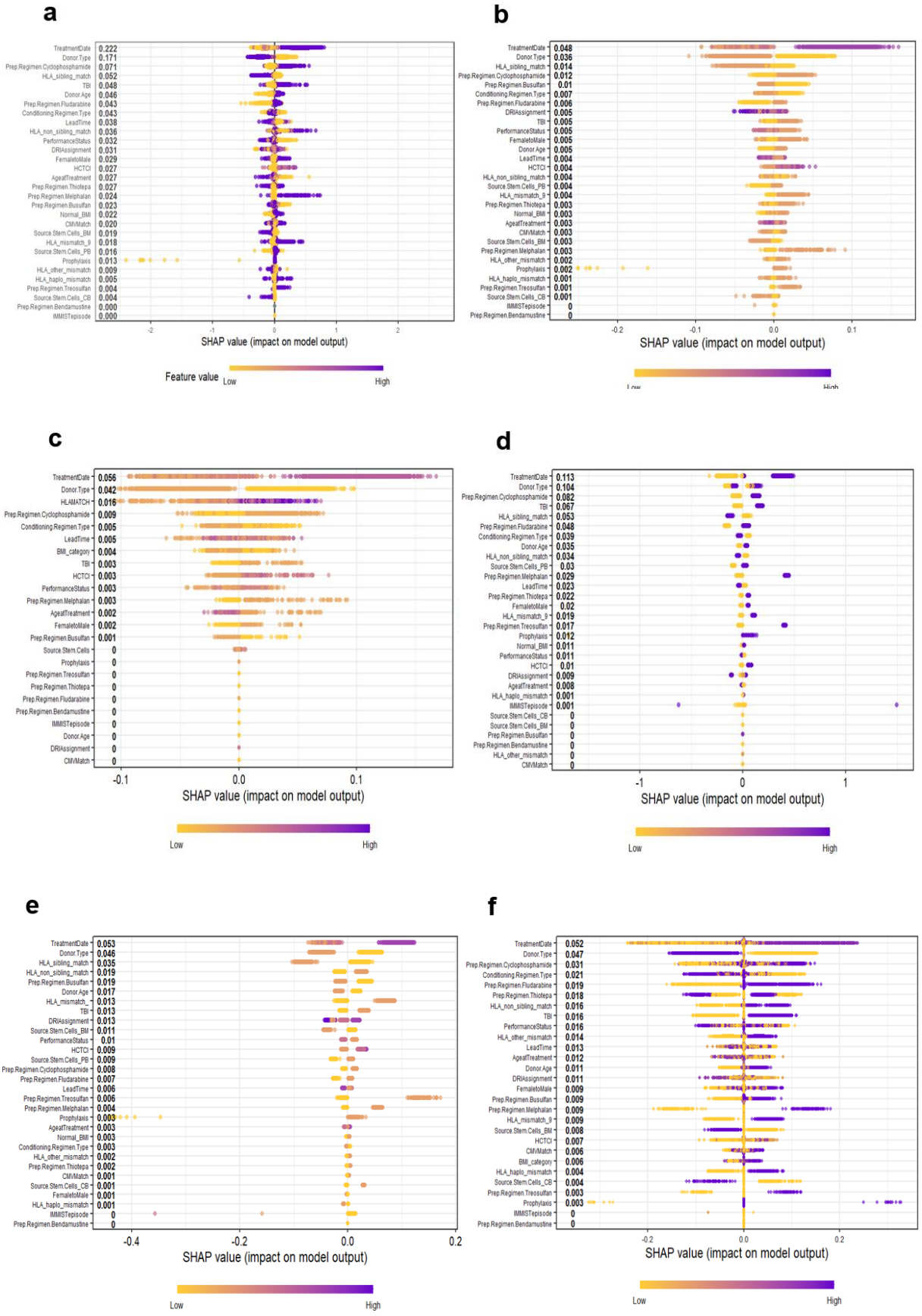
Regarding conditioning intensity, RIC was significantly associated with increased relapse risk compared to MAC (OR = 1.33, $p < 0.001$). The use of Thiotepa was significantly associated with reduced relapse risk (OR = 0.69, $p < 0.001$).

5.2.4 aGvHD

For aGvHD prediction, stacking yielded the highest AUC (0.686), though the difference compared to other models was minimal. All models achieved AUCs above 0.60 and were closely clustered. Stacking was the only model to surpass 0.60 in accuracy (0.613), whereas XGB had the lowest performance, with an accuracy of 0.558.

Stacking had the highest specificity (0.774), but its sensitivity was relatively low (0.504). XGB also showed high specificity (0.717), yet with lower sensitivity (0.451). BART displayed an imbalanced performance, with sensitivity of 0.700 and specificity of 0.508. In contrast, the remaining models exhibited more balanced sensitivity and specificity values (Appendix 8).

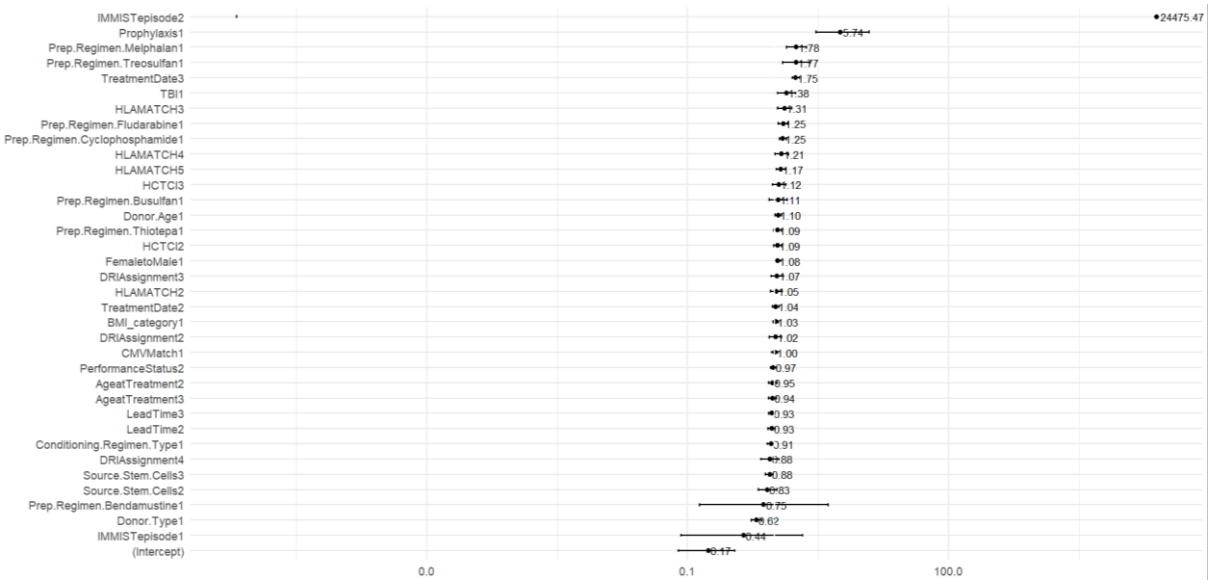
Figure 9: Summary SHAP value plots for aGvHD across six machine learning models: (a) XGB, (b) Random Forest, (c) Decision Tree, (d) Elastic Net, (e) Bayesian Classifier, and (f) BART. These plots illustrate the relative importance and direction of pre-transplant features across different models for predicting aGvHD.


a
TreatmentDate 0.222
Donor.Type 0.171
Prep.Regimen.Cyclophosphamide 0.071
HLA_sibling_match 0.052
TBI 0.048
Donor.Age 0.046
Prep.Regimen.Fludarabine 0.043
Conditioning.Regimen.Type 0.043
LeadTime 0.038
HLA_non_sibling_match 0.036
PerformanceStatus 0.032
DRIAssignment 0.031
FemaletoMale 0.029
HCTCI 0.027
AgeatTreatment 0.027
Prep.Regimen.Thiotepa 0.027
Prep.Regimen.Melphalan 0.024
Prep.Regimen.Busulfan 0.023
Normal_BMI 0.022
CMVMatch 0.020
Source.Stem.Cells_BM 0.019
HLA_mismatch_9 0.018
Source.Stem.Cells_PB 0.016
Prophylaxis 0.013
HLA_other_mismatch 0.009
HLA_haplo_mismatch 0.005
Prep.Regimen.Treosulfan 0.004
Source.Stem.Cells_CB 0.004
Prep.Regimen.Bendamustine 0.000
IMMSTepisode 0.000
SHAP value (impact on model output)
Feature value Low High
b
TreatmentDate 0.048
Donor.Type 0.036
HLA_sibling_match 0.014
Prep.Regimen.Cyclophosphamide 0.012
Prep.Regimen.Busulfan 0.01
Conditioning.Regimen.Type 0.007
Prep.Regimen.Fludarabine 0.006
DRIAssignment 0.005
TBI 0.005
PerformanceStatus 0.005
FemaletoMale 0.005
Donor.Age 0.005
LeadTime 0.004
HCTCI 0.004
HLA_non_sibling_match 0.004
Source.Stem.Cells_PB 0.004
HLA_mismatch_9 0.004
Prep.Regimen.Thiotepa 0.003
Normal_BMI 0.003
AgeatTreatment 0.003
CMVMatch 0.003
Source.Stem.Cells_BM 0.003
Prep.Regimen.Melphalan 0.003
HLA_other_mismatch 0.002
Prophylaxis 0.002
HLA_haplo_mismatch 0.001
Prep.Regimen.Treosulfan 0.001
Source.Stem.Cells_CB 0.001
IMMSTepisode 0
Prep.Regimen.Bendamustine 0
SHAP value (impact on model output)
c
TreatmentDate 0.056
Donor.Type 0.042
HLAMATCH 0.016
Prep.Regimen.Cyclophosphamide 0.009
Conditioning.Regimen.Type 0.005
LeadTime 0.005
BMI_category 0.004
TBI 0.003
HCTCI 0.003
PerformanceStatus 0.003
Prep.Regimen.Melphalan 0.003
AgeatTreatment 0.002
FemaletoMale 0.002
Prep.Regimen.Busulfan 0.001
Source.Stem.Cells 0
Prophylaxis 0
Prep.Regimen.Treosulfan 0
Prep.Regimen.Thiotepa 0
Prep.Regimen.Fludarabine 0
Prep.Regimen.Bendamustine 0
IMMSTepisode 0
Donor.Age 0
DRIAssignment 0
CMVMatch 0
-0.1 0.0 0.1
SHAP value (impact on model output)
Low High
d
TreatmentDate 0.113
Donor.Type 0.104
Prep.Regimen.Cyclophosphamide 0.082
TBI 0.067
HLA_sibling_match 0.053
Prep.Regimen.Fludarabine 0.048
Conditioning.Regimen.Type 0.039
Donor.Age 0.035
HLA_non_sibling_match 0.034
Source.Stem.Cells_PB 0.03
Prep.Regimen.Melphalan 0.029
LeadTime 0.023
Prep.Regimen.Thiotepa 0.022
FemaletoMale 0.02
HLA_mismatch_9 0.019
Prep.Regimen.Treosulfan 0.017
Prophylaxis 0.012
Normal_BMI 0.011
PerformanceStatus 0.011
HCTCI 0.01
DRIAssignment 0.009
AgeatTreatment 0.008
HLA_haplo_mismatch 0.001
IMMSTepisode 0.001
Source.Stem.Cells_CB 0
Source.Stem.Cells_BM 0
Prep.Regimen.Busulfan 0
Prep.Regimen.Bendamustine 0
HLA_other_mismatch 0
CMVMatch 0
-1 0 1
SHAP value (impact on model output)
Low High
e
TreatmentDate 0.053
Donor.Type 0.046
HLA_sibling_match 0.035
HLA_non_sibling_match 0.019
Prep.Regimen.Busulfan 0.019
Donor.Age 0.017
HLA_mismatch_ 0.013
TBI 0.013
DRIAssignment 0.013
Source.Stem.Cells_BM 0.011
PerformanceStatus 0.01
HCTCI 0.009
Source.Stem.Cells_PB 0.009
Prep.Regimen.Cyclophosphamide 0.008
Prep.Regimen.Fludarabine 0.007
LeadTime 0.006
Prep.Regimen.Treosulfan 0.006
Prep.Regimen.Melphalan 0.004
Prophylaxis 0.003
AgeatTreatment 0.003
Normal_BMI 0.003
Conditioning.Regimen.Type 0.003
HLA_other_mismatch 0.002
Prep.Regimen.Thiotepa 0.002
CMVMatch 0.001
Source.Stem.Cells_CB 0.001
FemaletoMale 0.001
HLA_haplo_mismatch 0.001
IMMSTepisode 0
Prep.Regimen.Bendamustine 0
-0.4 -0.2 0.0 0.2
SHAP value (impact on model output)
Low High
f
TreatmentDate 0.052
Donor.Type 0.047
Prep.Regimen.Cyclophosphamide 0.031
Conditioning.Regimen.Type 0.021
Prep.Regimen.Fludarabine 0.019
Prep.Regimen.Thiotepa 0.018
HLA_non_sibling_match 0.016
TBI 0.016
PerformanceStatus 0.016
HLA_other_mismatch 0.014
LeadTime 0.013
AgeatTreatment 0.012
Donor.Age 0.011
DRIAssignment 0.011
FemaletoMale 0.009
Prep.Regimen.Busulfan 0.009
Prep.Regimen.Melphalan 0.009
HLA_mismatch_9 0.009
Source.Stem.Cells_BM 0.008
HCTCI 0.007
CMVMatch 0.006
BMI_category 0.006
HLA_haplo_mismatch 0.004
Source.Stem.Cells_CB 0.004
Prep.Regimen.Treosulfan 0.003
Prophylaxis 0.003
IMMSTepisode 0
Prep.Regimen.Bendamustine 0
-0.2 0.0 0.2
SHAP value (impact on model output)
Low High

Based on the results from six SHAP-explainable models (Figure 9), Treatment Date, Donor Type, and HLA Matching consistently ranked among the most important features in predicting aGvHD. Notably, Treatment Date was ranked as the most important feature across all six models. According to the SHAP beeswarm plots, more recent transplant dates were associated with a higher risk of developing aGvHD. In terms of HLA Matching, patients who received transplants from fully matched sibling donors had a lower probability of developing aGvHD. Similarly, transplants from related donors were associated with a reduced risk of aGvHD compared to unrelated donors.

According to the logistic regression results (Figure 10), compared to transplants performed during 2013–2016, only those conducted in 2021–2023 were statistically significant, with an odds ratio (OR) of 1.75 ($p < 0.001$). Although transplants from 2017–2020 also had an OR greater than 1, the association was not statistically significant at the 5% level. Compared to unrelated donors, transplants from related donors were associated with a significantly lower risk of aGVHD (OR = 0.62, $p < 0.001$). In terms of HLA matching, all donor types other than sibling matching had ORs greater than 1. However, while most of these associations were statistically significant, non-sibling matched donors were not significant at the 5% level.

Figure 10: Forest Plot for Predictors of aGvHD

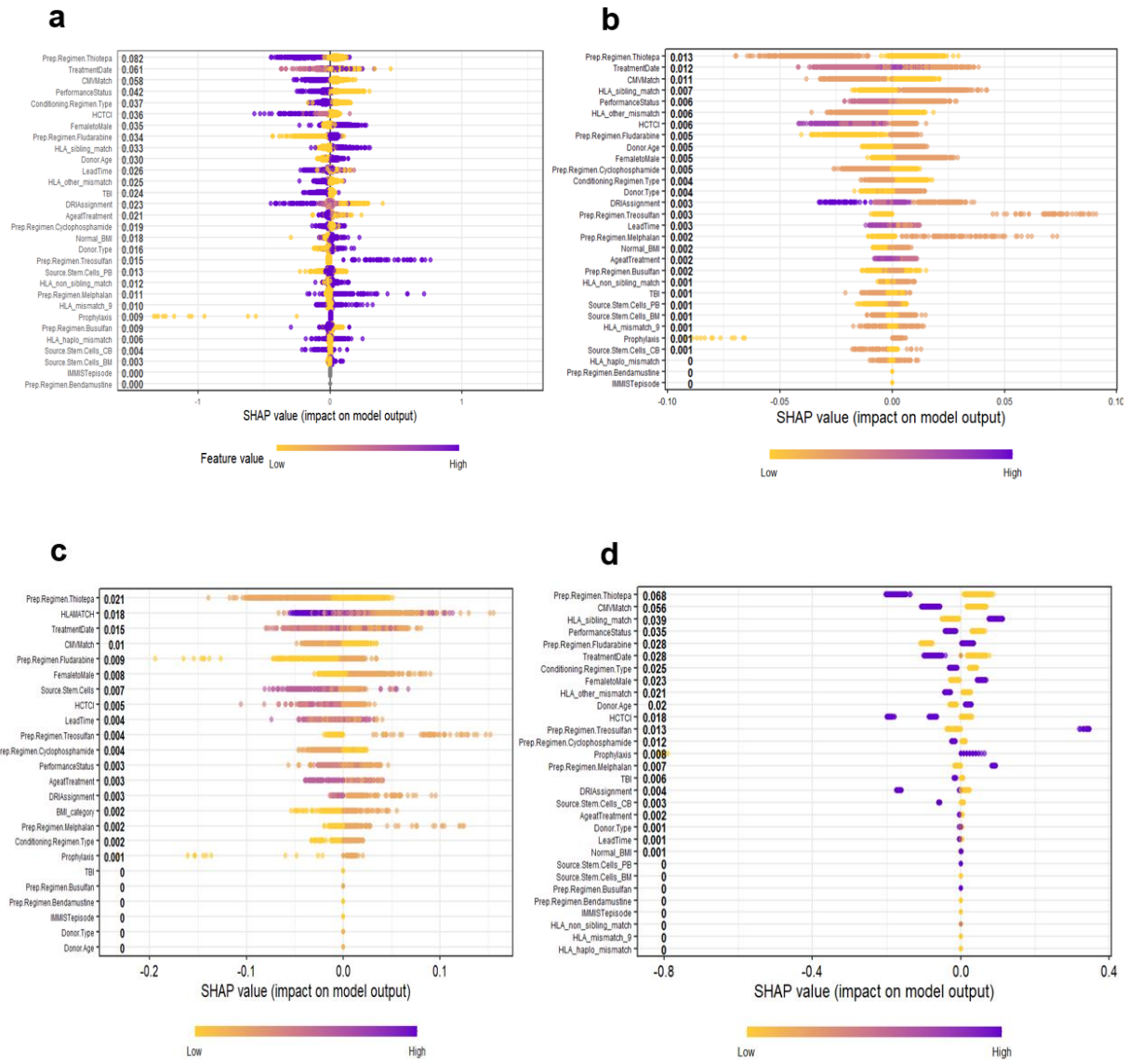


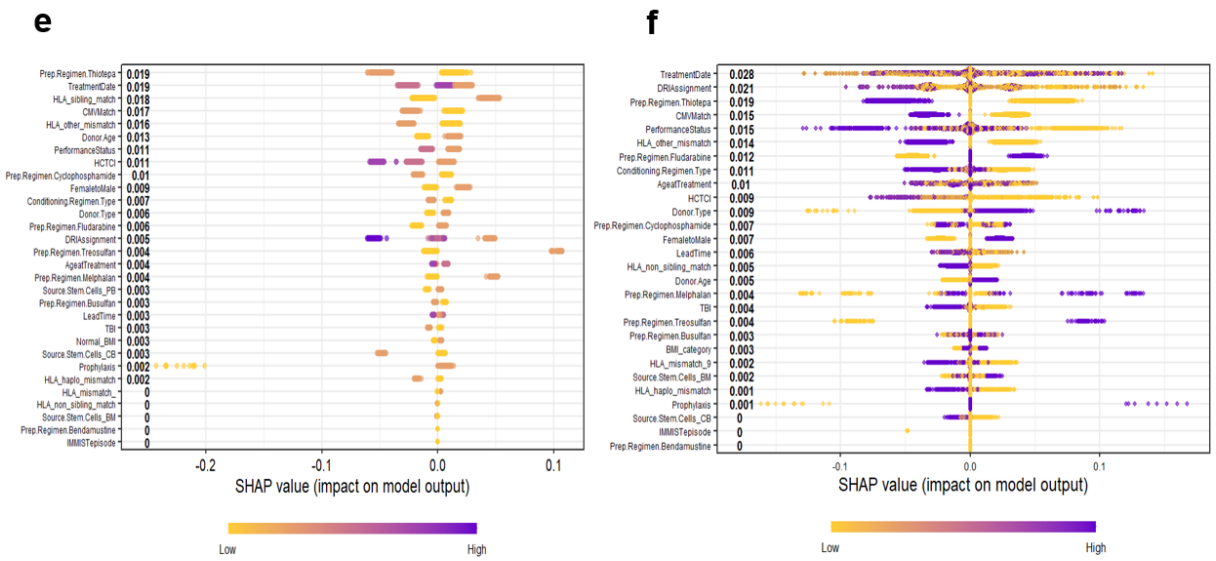
5.2.5 cGvHD

The predictive performance for cGvHD was generally unsatisfactory. Even the stacking model, which achieved the highest AUC and accuracy, only reached 0.563 and 0.549, respectively. For the other models, AUC values were clustered within a narrow range of 0.55 to 0.56, and

accuracy values ranged from 0.54 to 0.57. Sensitivity and specificity were similarly close across models, but all remained low, with none exceeding 0.7 (Appendix 9).

Figure 11: Summary SHAP value plots for cGvHD across six machine learning models: (a) XGB, (b) Random Forest, (c) Decision Tree, (d) Elastic Net, (e) Bayesian Classifier, and (f) BART. These plots illustrate the relative importance and direction of pre-transplant features across different models for predicting cGvHD.

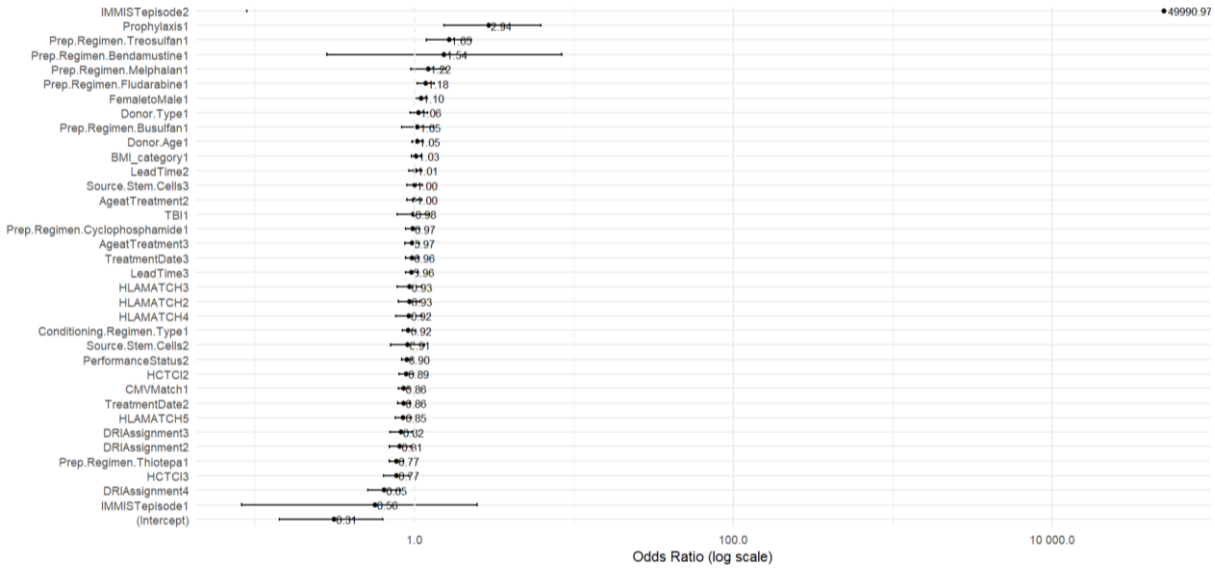




Based on the results from the six SHAP-explainable models (Figure 11), Thiotepa Used, Treatment Date, CMV Match, Performance Status, and HLA Match consistently ranked among the most important features. Notably, Thiotepa Use was ranked among the top three features across all models and was identified as the most important feature in four out of six models. The use of Thiotepa was associated with a lower probability of developing cGvHD. Similarly, patients with higher Performance Status scores were also less likely to develop cGvHD. Regarding CMV matching, patients with any mismatch (i.e., other than negative-to-negative) showed a higher likelihood of cGvHD. Nortably, the impact of Treatment Date showed a non-linear trend: patients who underwent transplantation during 2013–2016 had the highest probability of developing cGvHD, those transplanted in 2017–2020 had the lowest, while those in the 2021–2023 period showed an intermediate probability. Compared with sibling match, other HLA matching categories are associated with lower probability of developing cGvHD.

According to the logistic regression results (Figure 12), with respect to transplant year, both the 2017–2020 and 2021–2023 periods had ORs less than 1 when compared to 2013–2016. However, only 2017–2020 reached statistical significance (OR = 0.04, $p < 0.001$), while 2021–2023 was not significant at the 5% level. Patients with a Performance Status of 90 or 100 had a significantly lower risk of developing cGvHD compared to those with a score of 80 or below (OR = 0.04, $p < 0.01$). The use of Thiotepa was associated with a substantially reduced risk of cGvHD (OR = 0.05, $p < 0.001$). Regarding CMV matching, mismatched pairs showed a significantly increased risk compared to matched negative-to-negative pairs (OR = 0.05, $p < 0.001$). Among the HLA mismatch categories, only the "other mismatch" group showed a statistically significant association (OR = 0.85, $p < 0.001$).

Figure 12: Forest Plot for Predictors of cGvHD



6. Discussion

In the present study, we developed seven machine learning models—including one stacking model using the others as base learners—to predict the one-year probabilities of NRM, relapse, rejection, aGvHD, and cGvHD following Allo-HSCT. SHAP was employed to investigate the impact of individual features on these outcomes.

We identified the important features shared by the six SHAP-explainable algorithms and compared them with the findings from logistic regression. All of the most commonly ranked features by SHAP were statistically significant in logistic regression at least at the 5% level. It is noteworthy that for patients who had received at least two pre-transplant immunosuppressive treatments, the logistic regression model produced high odds ratios. However, the SHAP values for this feature remained very low in all models except the Elastic Net. This discrepancy may result from the fact that logistic regression tends to yield extreme OR estimates for rare but highly associated variables. In contrast, tree-based models employ built-in regularization, which often prevents rare features from being selected for splits [71, 72]. The Bayesian classifier implemented the Laplace smoothing, which is more tolerant with the outliers [73].

For the rejection outcome, a more recent transplant period and the use of RIC were identified as risk factors across both SHAP and logistic regression models. In contrast, sibling donors with 10/10 HLA matching were found to be protective. Notably, HLA matching showed an ordinal pattern: compared to sibling match, the odds of rejection increased progressively for non-sibling match, 9/10 mismatch, haploidentical mismatch, and other mismatches, with all

comparisons being statistically significant. Regarding relapse, higher DRI and the use of RIC were associated with increased relapse risk, while more recent transplant dates and the use of Thiotepa in the preparative regimen were also identified as risk factors. These findings were consistent across SHAP interpretations and regression estimates. For aGvHD, more recent transplant dates were a prominent risk factor, while having a related donor and a sibling HLA match appeared to be protective. In the case of chronic cGvHD, transplant in more recent years, use of Thiotepa, and higher performance status were protective factors. Conversely, CMV matching that was not negative-to-negative increased the risk of cGvHD. For NRM, the use of Thiotepa and older age at transplant were associated with higher risk, while more recent transplant periods were protective. Both transplant year and age at treatment demonstrated clear ordinal patterns.

In the analysis of different features across the five outcomes, Treatment Date consistently emerged as a key predictor. However, its associations with outcomes varied: more recent transplant date were associated with higher risks of rejection and aGvHD, but lower risks of relapse, cGvHD, and NRM. The use of Thiotepa in the preparative regimen was associated with lower probabilities of relapse and cGvHD, yet a higher risk of NRM. Patients with higher performance status were less likely to experience rejection and cGvHD. Notably, when comparing predictors for aGvHD and cGvHD, we observed that the same feature could have opposing effects on these two outcomes. For instance, more recent transplant years were associated with increased risk of aGvHD, but decreased risk of cGvHD. Similarly, sibling HLA matching appeared protective against aGvHD, but was paradoxically associated with a higher risk of cGvHD.

In evaluating model performance, this study primarily relied on AUC and accuracy as key metrics. Notably, accuracy was calculated after dichotomizing predicted probabilities using the Youden index, rather than applying a default cutoff of 0.5. This approach was taken to avoid inflated accuracy values that merely reflect the No Information Rate, a common issue when outcomes are imbalanced [74]. For outcomes including relapse, rejection, and NRM, the event was absent in the majority of samples.

In terms of accuracy, the stacking model also performed well, ranking highest for cGvHD, aGvHD, and especially for rejection, where its accuracy (0.845) significantly outperformed the second-best algorithm (0.620). However, for NRM, logistic regression achieved the highest accuracy, and for relapse, the best accuracy was obtained by the elastic net model.

For NRM, the stacking model demonstrated relatively balanced performance, with a specificity of 0.614 and sensitivity of 0.742. Similarly, for relapse, the model achieved a balanced

specificity (0.615) and sensitivity (0.734). In contrast, for aGvHD, stacking showed high specificity (0.774) but low sensitivity (0.504), suggesting the model is more effective at identifying patients unlikely to develop aGvHD, but may miss nearly half of true cases. A similar pattern was observed for rejection, where stacking achieved a high specificity of 0.875 but low sensitivity of 0.494. This indicates the model is well-suited for identifying "low-risk" patients, but may fail to detect a substantial proportion of those truly at high risk. Therefore, the model may be more appropriate as a rule-out tool. If the model predicts no rejection, the result is likely reliable. However, due to the low sensitivity, supplementary methods with higher sensitivity may be necessary to avoid under-diagnosing high-risk patients. For cGvHD, the stacking model demonstrated low sensitivity (0.513) and only moderate specificity (0.680). Given the relatively low AUC of 0.630 for cGvHD prediction, the model's overall discriminative ability is limited. Therefore, it is not recommended to use this model for clinical decision-making in this context.

7. Conclusion and Perspectives

In this study, eight machine learning algorithms were employed to evaluate the predictive performance for five major post-transplant outcomes—NRM, relapse, rejection, aGvHD, and cGvHD—within the first year after Allo-HSCT in patients with malignant diseases. SHAP was applied to enhance the interpretability of the models by quantifying the impact of individual clinical features. The stacking model consistently achieved robust performance.

However, to enhance the clinical applicability of the models, it is important to consider several limitations and potential directions for future research. Overall, the models demonstrated promising predictive potential across most outcomes, with the exception of cGvHD. One possible explanation is that cGvHD typically develops 3 to 6 months after transplantation. Therefore, relying solely on pre-transplant information may not be sufficient to accurately predict the occurrence of cGvHD. Future research should consider incorporating variables collected within the first 100 days post-transplant, such as chimerism status, infection, and post-transplant immunosuppressive treatment details. Analytical approaches such as the Cox proportional hazards model is appropriate. Alternatively, we could restrict the study population to patients who survive beyond 100 days and examine their risk of adverse outcomes over the subsequent nine months.

Additionally, this study did not include information about the treatment center. However, differences in treatment practices across centers may influence outcomes. For example, some centers may prioritize reducing mortality, while others may focus on minimizing the risk of

cGvHD. Patient selection criteria may also vary, potentially introducing bias if certain centers preferentially transplant patients with less severe disease.

Furthermore, this study included all patients with malignant hematologic diseases and used the DRI as a proxy for specific diagnoses. While this approach increases the generalizability of the findings, it may limit disease-specific insights, as DRI is a relatively broad indicator. Future work could perform subgroup analyses to identify risk factors specific to different diagnostic categories.

8. References

1. Lv, M., Gorin, N. C., & Huang, X. J. (2022). A vision for the future of allogeneic hematopoietic stem cell transplantation in the next decade. *Sci Bull*, *67*(19), 1921-1924.
2. Takami A. Hematopoietic stem cell transplantation for acute myeloid leukemia. *Int J Hematol.* 2018;107(5):513–8.
3. Zeiser R, Vago L. Mechanisms of immune escape after allogeneic hematopoietic cell transplantation. *Blood.* 2019;133(12):1290–7.
4. Wolff SN. Second hematopoietic stem cell transplantation for the treatment of graft failure, graft rejection or relapse after allogeneic transplantation. *Bone Marrow Transplant.* 2002;29(7):545–52.
5. Penack O, Peczynski C, Mohty M, Yakoub-Agha I, Styczynski J, Montoto S, et al. How much has allogeneic stem cell transplant–related mortality improved since the 1980s? A retrospective analysis from the EBMT. *Blood Adv.* 2020;4(24):6283–90.
6. Ball LM, Egeler RM. Acute GvHD: pathogenesis and classification. *Bone Marrow Transplant.* 2008;41(Suppl 2):S58–64.
7. Olivieri A, Mancini G. Current approaches for the prevention and treatment of acute and chronic GVHD. *Cells.* 2024;13(18):1524.
8. Kamble RT, Chang CC, Sanchez S, Carrum G. Central nervous system graft-versus-host disease: report of two cases and literature review. *Bone Marrow Transplant*. 2007;39(1):49–52.
9. Upperman JS, Lacroix J, Curley MA, Checchia PA, Lee DW, Cooke KR, et al. Specific etiologies associated with the multiple organ dysfunction syndrome in children: part 1. *Pediatr Crit Care Med*. 2017;18(3):S50–S57.
10. Olivieri A, Mancini G. Current approaches for the prevention and treatment of acute and chronic GVHD. *Cells.* 2024;13(18):1524.
11. Kollman C, Spellman SR, Zhang MJ, Hassebroek A, Anasetti C, Antin JH, et al. The effect of donor characteristics on survival after unrelated donor transplantation for hematologic malignancy. *Blood*. 2016;127(2):260–7.
12. De Haan G, Lazare SS. Aging of hematopoietic stem cells. *Blood*. 2018;131(5):479–87.
13. Bastos-Oreiro M, Gutierrez A, Reguera JL, Iacoboni G, López-Corral L, Terol MJ, et al. Best treatment option for patients with refractory aggressive B-cell lymphoma in the CAR-T cell era: real-world evidence from GELTAMO/GETH Spanish groups. *Front Immunol*. 2022;13:855730.
14. Sorror ML, Sandmaier BM, Storer BE, Maris MB, Baron F, Maloney DG, et al. Comorbidity and disease status–based risk stratification of outcomes among patients with acute myeloid leukemia or myelodysplasia receiving allogeneic hematopoietic cell transplantation. *J Clin Oncol*. 2007;25(27):4246–54.
15. Saraceni F, Labopin M, Forcade E, Kroger N, Socie G, Niittyvuopio R, et al. Allogeneic stem cell transplant in patients with acute myeloid leukemia and Karnofsky performance status score less than or equal to 80%: a study from the acute leukemia working party of the European Society for Blood and Marrow Transplantation (EBMT). *Cancer Med*. 2021;10(1):23–33.
16. Salas MQ, Prem S, Atenafu EG, Datt Law A, Lam W, Al-Shaibani Z, et al. Dual T-cell depletion with ATG and PTCy for peripheral blood reduced intensity conditioning allo-HSCT results in very low rates of GVHD. *Bone Marrow Transplant*. 2020;55(9):1773–83.
17. Salehnasab C, Hajifathali A, Asadi F, Roshandel E, Kazemi A, Roshanpoor A. Machine learning classification algorithms to predict aGvHD following allo-HSCT: a systematic review. *Methods Inf Med*. 2019;58(6):205–12.
18. Shaw BE, Logan BR, Spellman SR, et al. Development of an unrelated donor selection score predictive of survival after HCT: donor age matters most. *Biol Blood Marrow Transplant*. 2018;24(6):1049–56.

19. Gratwohl A, Pasquini MC, Aljurf M, Worldwide Network for Blood and Marrow Transplantation (WBMT), et al. One million haemopoietic stem-cell transplants: a retrospective observational study. *Lancet Haematol*. 2015;2:e91–100.
20. Gratwohl A, Sureda A, Baldomero H, Joint Accreditation Committee (JACIE) of the International Society for Cellular Therapy (ISCT), the European Society for Blood and Marrow Transplantation (EBMT), the European Leukemia Net (ELN), et al. Economics and outcome after hematopoietic stem cell transplantation: a retrospective cohort study. *EBioMedicine*. 2015;2:2101–9.
21. Ljungman P. The role of cytomegalovirus serostatus on outcome of hematopoietic stem cell transplantation. *Curr Opin Hematol*. 2014;21(6):466–9.
22. Kalra A, Williamson T, Daly A, et al. Impact of donor and recipient cytomegalovirus serostatus on outcomes of antithymocyte globulin-conditioned hematopoietic cell transplantation. *Biol Blood Marrow Transplant*. 2016;22(9):1654–63.
23. Eapen M, Wang T, Veys PA, et al. Allele-level HLA matching for umbilical cord blood transplantation for non-malignant diseases in children: a retrospective analysis. *Lancet Haematol*. 2017;4(5):325–33.
24. Little AM, Akbarzad-Yousef A, Anand A, et al. BSHI guideline: HLA matching and donor selection for haematopoietic progenitor cell transplantation. *Int J Immunogenet*. 2021;48(1):75–109.
25. Mayor NP, Hayhurst JD, Turner TR, et al. Better HLA matching as revealed only by next generation sequencing technology results in superior overall survival post-allogeneic haematopoietic cell transplantation with unrelated donors. *Biol Blood Marrow Transplant*. 2018;24(1):63–4.
26. Gratwohl A, Stern M, Brand R, et al. European Group for Blood and Marrow Transplantation and the European leukemia net. Risk score for outcome after allogeneic hematopoietic stem cell transplantation: a retrospective analysis. *Cancer*. 2009;115(22):4715–26.
27. Gratwohl A, Sureda A, Cornelissen J, et al. Alloreactivity: the Janus-face of hematopoietic stem cell transplantation. *Leukemia*. 2017;31(8):1752–9.
28. Riezzo I, Pascale N, La Russa R, Liso A, Salerno M, Turillazzi E. Donor selection for allogenic hemopoietic stem cell transplantation: clinical and ethical considerations. *Stem Cells Int*. 2017;2017:5250790.
29. Nagler A, Rocha V, Labopin M, et al. Allogeneic hematopoietic stem-cell transplantation for acute myeloid leukemia in remission: comparison of intravenous busulfan plus cyclophosphamide (Cy) versus total-body irradiation plus Cy as conditioning regimen—a report from the acute leukemia working party of the European group for blood and marrow transplantation. *J Clin Oncol*. 2013;31(27):3549–56.
30. Abelsson J, Merup M, Birgegård G, WeisBjerrum O, Brinch L, Brune M, et al. The outcome of allo-HSCT for 92 patients with myelofibrosis in the Nordic countries. *Bone Marrow Transplant*. 2012;47(3):380–6.
31. Beynarovich A, Lepik K, Mikhailova N, Borzenkova E, Volkov N, Moiseev I, et al. Favorable outcomes of allogeneic hematopoietic stem cell transplantation with fludarabine–bendamustine conditioning and posttransplantation cyclophosphamide in classical Hodgkin lymphoma. *Int J Hematol*. 2022;116(3):401–10.
32. Liu H, Zhai X, Song Z, Sun J, Xiao Y, Nie D, et al. Busulfan plus fludarabine as a myeloablative conditioning regimen compared with busulfan plus cyclophosphamide for acute myeloid leukemia in first complete remission undergoing allogeneic hematopoietic stem cell transplantation: a prospective and multicenter study. *J Hematol Oncol*. 2013;6:34.
33. Eder S, Beohou E, Labopin M, Sanz J, Finke J, Arcese W, et al. Thiotepa-based conditioning for allogeneic stem cell transplantation in acute lymphoblastic leukemia—A survey from the Acute Leukemia Working Party of the European Society for Blood and Marrow Transplantation. *Am J Hematol*. 2017;92(1):18–22.

34. Michniacki TF, Choi SW, Peltier DC. Immune suppression in allogeneic hematopoietic stem cell transplantation. In: *Pharmacology of Immunosuppression*. Cham: Springer International Publishing; 2021. p. 209–43.
35. Radakovich N, Nagy M, Nazha A. Machine learning in haematological malignancies. *Lancet Haematol.* 2020;7(7):e541–50. doi:10.1016/S2352-3026(20)30121-6.
36. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Comput Methods Programs Biomed.* 2022;214:106584.
37. Tugwell P, Tovey D. PRISMA 2020. *J Clin Epidemiol*. 2021;134:A5–A6.
38. von Asmuth EG, Neven B, Albert MH, Mohseny AB, Schilham MW, Binder H, et al. Predicting patient death after allogeneic stem cell transplantation for inborn errors using machine learning (PREPAD): a European Society for Blood and Marrow Transplantation Inborn Errors Working Party Study. *Transplant Cell Ther.* 2023;29(12):775.e1.
39. Wang P, Liu C, Wei Z, Jiang W, Sun H, Wang Y, et al. Nomogram for predicting early mortality after umbilical cord blood transplantation in children with inborn errors of immunity. *J Clin Immunol.* 2023;43(6):1379–92.
40. Afanaseva KS, Bakin EA, Smirnova AG, Barkhatov IM, Gindina TL, Moiseev IS, Bondarenko SN. A pilot study of implication of machine learning for relapse prediction after allogeneic stem cell transplantation in adults with Ph-positive acute lymphoblastic leukemia. *Sci Rep.* 2023;13(1):16790.
41. Qu Y, Shourabizadeh H, Subramanian A, Aleman DM, Rousseau LM, Law AD, et al. Differential impact of CD34+ cell dose for different age groups in allogeneic hematopoietic cell transplantation for acute leukemia: a machine learning–based discovery. *Exp Hematol.* 2025;141:104684.
42. Shourabizadeh H, Aleman DM, Rousseau LM, Law AD, Viswabandya A, Michelis FV. Machine learning for the prediction of survival post-allogeneic hematopoietic cell transplantation: A single-center experience. *Acta Haematol.* 2024;147(3):280–91.
43. Lee S, Lee E, Park SS, Park MS, Jung J, Min GJ, et al. Prediction and recommendation by machine learning through repetitive internal validation for hepatic veno-occlusive disease/sinusoidal obstruction syndrome and early death after allogeneic hematopoietic cell transplantation. *Bone Marrow Transplant.* 2022;57(4):538–46.
44. Jo T, Inoue K, Ueda T, Iwasaki M, Akahoshi Y, Nishiwaki S, et al. Machine learning evaluation of intensified conditioning on haematopoietic stem cell transplantation in adult acute lymphoblastic leukemia patients. *Communications Medicine.* 2024;4(1):247.
45. Alawneh H, Hasasneh A. Survival prediction of children after bone marrow transplant using machine learning algorithms. *Int Arab J Inf Technol.* 2024;21(3):394–407.
46. Ratul IJ, Wani UH, Nishat MM, Al-Monsur A, Ar-Rafi AM, Faisal F, Kabir MR. Survival prediction of children undergoing hematopoietic stem cell transplantation using different machine learning classifiers by performing chi-square test and hyperparameter optimization: a retrospective analysis. *Comput Math Methods Med.* 2022;2022:9391136.
47. Chadaga K, Prabhu S, Sampathila N, Chadaga R. A machine learning and explainable artificial intelligence approach for predicting the efficacy of hematopoietic stem cell transplant in pediatric patients. *Healthc Anal.* 2023;3:100170.
48. Short SM, Perez MD, Morse AE, Jennings RD, Howard DS, Foureau D, et al. High-dimensional immune profiles and machine learning may predict acute myeloid leukemia relapse early following transplant. *J Immunol.* 2024;213(10):1441–51.
49. Marvin G, Alam MGR. Explainable computational pathology for survival prediction in hematologic pediatric patients. In: *2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*; 2022 Dec; [place if known]. IEEE; 2022. p. 1–6.
50. Rifat AI, Hossain M, Nahid N, Akter S, Islam A. Children hematopoietic stem cell transplant survival status prediction using machine learning. In: *2023 Annual International Conference on*

Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS); 2023 Nov; [地点]. IEEE; 2023. p. 1–6.

51. Hossain MJ, Ferdaus J, Rahman S, Santa SA. Multiclass classification of GVHD and relapse: A comparative analysis among machine learning and ANN-based algorithms. In: *2022 25th International Conference on Computer and Information Technology (ICCIT)*; 2022 Dec; Cox's Bazar, Bangladesh. *IEEE*; 2022. p. 489–94.
52. Gourisaria MK, Patel AV, Chatterjee R, Sahoo B. Predicting the survival status of patient after bone marrow transplant using fuzzy discernibility matrix. In: *2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OTCON)*; 2023 Feb; Raigarh, India. *IEEE*; 2023. p. 1–6.
53. Afanaseva KS, Bakin EA, Smirnova AG, Barkhatov IM, Gindina TL, Moiseev IS, Bondarenko SN. A pilot study of implication of machine learning for relapse prediction after allogeneic stem cell transplantation in adults with Ph-positive acute lymphoblastic leukemia. *Scientific Reports*. 2023;13(1):16790.
54. Spellman SR, Sparapani R, Maiers M, Shaw BE, Laud P, Bupp C, *et al.* Novel machine learning technique further clarifies unrelated donor selection to optimize transplantation outcomes. *Blood Adv*. 2024;8(23):6082–7.
55. Choi EJ, Jun TJ, Park HS, Lee JH, Lee KH, Kim YH, *et al.* Predicting long-term survival after allogeneic hematopoietic cell transplantation in patients with hematologic malignancies: machine learning–based model development and validation. *JMIR Med Inform*. 2022;10(3):e32313.
56. Okamura H, Nakamae M, Koh S, Nanno S, Nakashima Y, Koh H, *et al.* Interactive web application for plotting personalized prognosis prediction curves in allogeneic hematopoietic cell transplantation using machine learning. *Transplantation*. 2021;105(5):1090–6.
57. Iwasaki M, Kanda J, Arai Y, Kondo T, Ishikawa T, Ueda Y, *et al.* Establishment of a predictive model for GVHD-free, relapse-free survival after allogeneic HSCT using ensemble learning. *Blood Adv*. 2022;6(8):2618–27.
58. Echecopar C, Abad I, Galán-Gómez V, Mozo del Castillo Y, Sisinni L, Bueno D, et al. An artificial intelligence-driven predictive model for pediatric allogeneic hematopoietic stem cell transplantation using clinical variables. *Eur J Haematol*. 2024;112(6):910-6.
59. McCurdy SR, Radojcic V, Tsai HL, Vulic A, Thompson E, Ivcevic S, *et al.* Signatures of GVHD and relapse after posttransplant cyclophosphamide revealed by immune profiling and machine learning. *Blood*. 2022;139(4):608–23.
60. Mussetti A, Rius-Sansalvador B, Moreno V, Peczynski C, Polge E, Galimard JE, *et al.* Artificial intelligence methods to estimate overall mortality and non-relapse mortality following allogeneic HCT in the modern era: an EBMT-TCWP study. *Bone Marrow Transplant*. 2024;59(2):232–8.
61. Saengboon S, Ciurea S, Popat U, Ramdial J, Bashir Q, Alousi A, et al. Long-term outcomes after haploidentical stem cell transplantation for hematologic malignancies. *Blood Adv*. 2024;8(12):3237–45.
62. Weller JF, Lengerke C, Finke J, Schetelig J, Platzbecker U, Einsele H, et al. Allogeneic hematopoietic stem cell transplantation in patients aged 60–79 years in Germany (1998–2018): a registry study. *Haematologica*. 2023;109(2):431.
63. Nowak J. Role of HLA in hematopoietic SCT. *Bone Marrow Transplant*. 2008;42(2):S71–S76.
64. Martínez C, Gayoso J, Canals C, Finel H, Peggs K, Dominietto A, et al. Post-transplantation cyclophosphamide-based haploidentical transplantation as alternative to matched sibling or unrelated donor transplantation for Hodgkin lymphoma: a registry study of the Lymphoma Working Party of the European Society for Blood and Marrow Transplantation. *J Clin Oncol*. 2017;35(30):3425–32.
65. Reisner Y, Hagin D, Martelli MF. Haploidentical hematopoietic transplantation: current status and future perspectives. *Blood*. 2011;118(23):6006–17.

66. Passweg JR, Baldomero H, Chabannon C, Basak GW, de La Cámara R, Corbacioglu S, et al.; European Society for Blood and Marrow Transplantation (EBMT). Hematopoietic cell transplantation and cellular therapy survey of the EBMT: monitoring of activities and trends over 30 years. *Bone Marrow Transplant*. 2021;56(7):1651–64.
67. WHO Expert Consultation. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet*. 2004;363(9403):157–63.
68. Jeljeli M, Guérin-El Khourouj V, Porcher R, Fahd M, Leveillé S, Yakouben K, et al. Relationship between cytomegalovirus (CMV) reactivation, CMV-driven immunity, overall immune recovery and graft-versus-leukaemia effect in children. *Br J Haematol*. 2014;166(2):229–39.
69. Balduzzi A, Bönig H, Jarisch A, Nava T, Ansari M, Cattoni A, et al.; Pediatric Diseases Working Party EBMT. ABO incompatibile graft management in pediatric transplantation. *Bone Marrow Transplant*. 2021;56(1):84–90. doi:10.1038/s41409-020-0981-7. Epub 2020 Jun 27. Erratum in: *Bone Marrow Transplant*. 2020 Aug 4.
70. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biometr J*. 2005;47(4):458–72.
71. Muchlinski D, Siroky D, He J, Kocher M. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Polit Anal*. 2016;24(1):87-103.
72. Speiser JL, Durkalski VL, Lee WM. Random forest classification of etiologies for an orphan disease. *Stat Med*. 2015;34(5):887-99.
73. Schonlau M. The naive Bayes classifier. In: *Applied statistical learning: With case studies in Stata*. Cham: Springer International Publishing; 2023. p. 143-60.
74. Thabtah F, Hammoud S, Kamalov F, Gonsalves A. Data imbalance in classification: Experimental evaluation. *Inf Sci (N Y)*. 2020;513:429–41.

9. Appendices

Appendix 1: Search Commands for the Studies Retrieval

Database	Search Commands
PUBMED	(allogeneic hematopoietic stem cell transplantation OR GvHD OR acute GvHD OR aGvHD OR allogeneic HCT OR Hematopoietic Cell Transplantation OR umbilical cord blood transplantation OR allogeneic Bone marrow transplant OR allogeneic Hematopoietic cell transplant OR allogeneic Hematopoietic stem cell transplantation OR Graft-versus-host disease) AND (Relapse OR rejection OR Mortality) AND (Machine learning OR Artificial Intelligence)
SCOPUS	(allogeneic AND hsct OR GvHD OR acute AND GvHD OR aGvHD OR allogeneic AND hct OR hematopoietic AND cell AND transplantation OR allogeneic AND bone AND marrow AND transplant OR allogeneic AND hematopoietic AND cell AND transplant OR allogeneic AND hematopoietic AND stem AND cell AND transplantation OR graft-versus-host AND disease) AND (relapse OR rejection OR survival OR death OR mortality) AND (machine AND learning OR artificial AND intelligence) AND (LIMIT-TO (EXACTKEYWORD , "Human") OR LIMIT-TO (EXACTKEYWORD , "Humans") OR LIMIT-TO (EXACTKEYWORD , "Artificial Intelligence") OR LIMIT-TO (EXACTKEYWORD , "Hematopoietic Stem Cell Transplantation") OR LIMIT-TO (EXACTKEYWORD , "Allogeneic Hematopoietic Stem Cell Transplantation") OR LIMIT-TO (EXACTKEYWORD , " Bone Marrow Transplant")) AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (LANGUAGE , "English"))
IEEE Xplore	("All Metadata":allogeneic HSCT) OR ("All Metadata":GvHD) OR ("All Metadata":acute GvHD) OR ("All Metadata": aGvHD) OR ("All Metadata":allogeneic HCT) OR ("All Metadata": Hematopoietic Cell Transplantation) OR ("All Metadata":allogeneic Bone marrow transplant) OR ("All Metadata":allogeneic Hematopoietic cell transplant) OR ("All Metadata":allogeneic Hematopoietic stem cell transplantation) OR ("All Metadata":Graft-versus-host disease) AND ("All Metadata":Machine learning)

Appendix 2. Summary of Reviewed Studies Information

Reference	Outcomes	Participants	Age	Features
von Asmuth et al., 2023 [38]	1-year, 2-years and 5-years overall survival	10888	(01-10)	Diagnosis, Donor Type, Performance Score, Age Patient, CMV Matching, Conditioning Agent, Stem Cell Source, HLA Mismatches, Donor Sex, Serotherapy, Patient Sex, Myeloablative Conditioning, Age Donor, Irradiation
Wang et al., 2023 [39]	Early mortality	230	(01-14)	Demographics, Sex, Female, Male, Weight, Height, BMI, Disease, CGD, SCID, VEO-IBD, Other IEI, Age at onset, Age at diagnosis, Age at UCBT, Onset to diagnosis, Diagnosis to UCBT, Medical history, Sepsis, Pneumonia, Pulmonary fungal infection, Severe pneumonia, Intestinal infection, Urinary tract infection, CNS infection, SSTI, CMV infection, EBV infection, BCG disease, Liver dysfunction, Laboratory tests, Albumin, ALT, Total bilirubin, IgA, IgG, IgM, IgE, CD19 count, CD19 ratio, CD3 count, CD3 ratio, CD4 count, CD4 ratio, CD8 count, CD8 ratio, CD56 count, CD56 ratio, elevated CRP, elevated PCT, IL-6
Afanaseva et al., 2023 [40]	Relapse	74	(18-44)	Highest Preceding BCR-ABL1, BCR::ABL1 Expression Level at Prediction Moment, Current BCR-ABL1, Chronic GVHD Presence, TKL2, TKL1
Qu et al., 2025 [41]	Overall survival	1153	(18-76)	Age at BMT, Mismatch, Donor Age, ATG Dose, Diagnosis ALL, KPS, Cytogenetics Risk, -5, -7, -17 Cytogenetics Abnormality, CMV D-R+, Days Dx to Tx, HCT CI, ABO Compatibility, RIC MAC, CD34 Dose, DRI
Shourabizadeh et al., 2024 [42]	100-days survival	2697	(12-74)	Donor relationship, TBI dose, Recipient age, CP1, Baseline HGB, FEV1%, Time between diagnosis to transplant, Total graft MNC count, Donor type - mismatch, GVHD prophylaxis - CSA, CR1, ALL, Other lymphoma, Graft-PBSC, Baseline WBC, Total graft MNC

				concentration, DLCO corrected for Hb, Baseline AST
Lee et al., 2022 [43]	VOD/SOS, 100-days survival	2572	(18-74)	TBI dose, Busulfan dose, Conditioning intensity, VOD prophylaxis, Immunosuppressants, Stem cell source
Jo et al., 2024 [44]	Relapse	109	(01-22)	Recipient age, Sex, Graft, HLA match, GVHD prophylaxis, Whole bone marrow chimerism, Whole peripheral blood chimerism, Remission status prior to transplant, Peripheral blood CD3 chimerism, TBI, Bone marrow CD15 chimerism, Bone marrow CD3 chimerism, Peripheral blood CD15 chimerism, Bone marrow CD34 chimerism, Peripheral blood CD34 chimerism
Alawneh & Hasasneh, 2024 [45]	Mortality	187	Children	Survival time, PLT recovery, Relapse, CD3 dose, CD34 dose, ANC recovery, Recipient age, CD3/CD34, Donor age, Disease lymphoma, Time to acute GvHD III IV, Donor ABO AB, Recipient CMV, Treatment post relapse, CMV status, Recipient RH plus, Stem cell source, Risk group, CMV status, Chronic, Extensive chronic GvHD, HLA group
Ratul et al., 2022 [46]	Overall survival	187	Children	Recipients Age, Recipient Body Mass, CD34 dose, CD3 dose, CD3/CD34, ANC Recovery, PLT Recovery, Time To Acute GvHD III IV, Survival Time, Disease Lymphoma, Relapse, Survival Status
Chadaga et al., 2023 [47]	Overall survival	187	Children	Recipient Gender, Stem cell source, Donor age, Donorage35, AGVHD IIIV, Gender Match, Donor ABO, Recipent ABO, Recipient Rh, ABO match, CMV Status, Donor CMV, Recipient CMV, Disease, Risk group, Txpostrelapse, Disease group, HLA match, Antigen, Allel, HLA, Recipient age, Recipient age below 10, Recipient age, Relapse, aGvHDIIIIV, extcGvHD, CD34 dose, CD3/CD34, CD3 dose, Rbodymass, ANCrecovery, PLTrecovery, Time to aGvHD III IV, Survival time, Survival status
Short et al., 2024 [48]	Relapse	60	(21-72)	All kinds of antibodies (e.g., CD4, CD19, etc.)

Marvin & Alam, 2022 [49]	Overall survival	187	Children	Survival time, Relapse, Risk group, CD34 dose, ANC recovery, Recipient age, Recipient CMV, Extensive chronic GVHD, Recipient ABO, Donor age, Antigen, Recipient Gender, Donor CMV, HLA Matching, Recipient BMI
Rifat et al., 2023 [50]	Overall survival	187	Children	extcGVHD, aGVHDIIIIV, Txpostrelapse, RecipientRH, Stemcellsource, Riskgroup, Relapse, Recipientgender, Recipientageit, IIIV, Recipientage10, CD34kgx10d6, CD3dCD34, CD3dkgx10d8, Rbodymass, ANCrecovery, PLTrecovery, Recipientage, survival tie
Hossain et al., 2022 [51]	Relapse, GVHD	187	Children	CD34 dose, CD3 dose, HLA match, HLA group, Allel, Stem cell source, Disease, ABO match, Risk group, Recipient CMV, Donor CMV, txpost relapse, Donor age below 35, Recipient gender, Gender match, Recipient rh, Plt recovery, Recipient age below 10, Recipient BMI
Gourisaria et al., 2023 [52]	Overall survival	187	Children	Only known they selected 11 features out of the following variables (Donor age below 35, Donor ABO, Donor CMV, Recipient age, Recipient gender, Recipient ABO, Recipient rh, Recipient CMB, Disease, HLA match, Risk group, Stem cell source, txpost relapse, CD3 dose, CD34 dose, CD3/CD34, ANC recovery, PLT recovery, extensive chronic GVHD, Relapse, Survival time, Survival status)
Afanaseva et al., 2023 [53]	Overall survival	4652	(16-70)	Patient age, Patient sex, Performance status, CMV antibody, Active bacterial or fungal infection at HSCT, Phenotype of disease (B-cell, T-cell), Philadelphia chromosome, Refined Disease Risk Index, Time between diagnosis to HSCT, Donor sex, Sex mismatch between donor and patient, ABO blood type mismatch, HLA mismatch, Graft source, Prophylaxis, Transplantation year
Spellman et al., 2024 [54]	3-years overall survival and Event (cGvHD, Relapse and Rejection) -free survival	11818	(01-82)	Recipient Age, Recipient Sex, Recipient Race, Recipient Ethnicity, Donor Age, Donor Parity, Sex Matching, CMV Serostatus, DQB1 Match, DPB1 Match or Permissive Mismatch, Graft Type, Karnofsky

				Score, Disease, Disease Stage, Conditioning Regimen, GVHD Prophylaxis, HCT-Comorbidity Index, Time from Diagnosis to HCT, Transplant Year
Choi et al., 2022 [55]	Overall survival, Relapse, and Relapse-free survival	1470	(15-75)	Diagnosis and disease, Disease risk, WBC count at diagnosis, Extramedullary disease at diagnosis, Extramedullary disease at HCT, Karyotype at diagnosis, Karyotype at HCT, CMV serostatus of recipient, CMV serostatus of donor, Hepatic score of HCT-CI, Total score of HCT-CI, Conditioning regimen, Donor type, Recipient HLA type, Donor HLA type, RBC transfusion before HCT, Platelet transfusion before HCT
Okamura et al., 2021 [56]	1-year overall survival, progression-free survival, relapse/progression, relapse-free survival	363	(17-69)	Recipients' age, Refined Disease Risk Index, Hematopoietic Cell Transplantation Comorbidity Index, Performance status, Donor source, HLA compatibility, Conditioning intensity, Number of allogeneic hematopoietic cell transplantations
Iwasaki et al., 2021 [57]	GVHD and Relapse-free survival	2207	All age groups	Donor Relation, Donor source, Donor type, Sex, Age, Sex mismatch, Diagnosis, Disease stage, DRI, HCT-CI, Performance Status
Echecopar et al., 2024 [58]	1-year overall survival,	201	Children	Disease Status, HLA Compatibility, Year Of HSCT, Age At HSCT, Conditioning Regimen, Time between diagnosis and HSCT, Diagnosis
McCurd et al., 2022 [59]	Overall survival	145	(22-64)	NK cell recovery, Pre-transplant status, ST2, Naive CD4, TNFR1, Graft CD3, CXCL9, Reg3a, BMT year, Plasmablast
Mussetti et al., 2023 [60]	Overall survival and Non-relapse survival	33927	(18-80)	Patient's age, Patient's sex, Karnofsky score, presence of significant comorbidities, Total number of comorbidities, Patient's CMV serostatus, Donor's sex, Donor's CMV serostatus, Donor type, Graft type, Conditioning intensity, GVHD prophylaxis, Disease risk index, Type of disease, Disease status at transplant

Appendix 3: Summary of Missing Data

Features	Count	Percentage
Recipient Gender	1	0.01%
Donor Gender	115	2.53%
CMV Status Donor	88	0.52%
CMV Status Patient	96	0.54%
Donor Type	8	0.05%
Source Stem Cells	9	0.05%
Prophylaxis	12	0.07%
TBI	33	0.20%
IMMISTepisode	14	0.09%
HCICI	16	0.01%
DRI	680	4.14%
Performance Status	1263	7.69%
HLA Matching	59	0.36%

Appendix 4: The Hyperparameters Summary for XGB, Random Forest, Elastic Net, Decision Tree, and BART

	NRM	Relapse	Rejection	aGvHD	cGvHD
XGB	{‘eval_metric’: AUC, ‘max_depth’: 6, ‘eta’: 0. 3, ‘nfold’: 10, ‘best_nrounds’: 13 }	{‘eval_metric’: AUC, ‘max_depth’: 6, ‘eta’: 0. 3, ‘nfold’: 10, ‘best_nrounds’: 14 }	{‘eval_metric’: AUC, ‘max_depth’: 6, ‘eta’: 0. 3, ‘nfold’: 10, ‘best_nrounds’: 16}	{‘eval_metric’: AUC, ‘max_depth’: 6, ‘eta’: 0. 3, ‘nfold’: 10, ‘best_nrounds’: 15 }	{‘eval_metric’: AUC, ‘max_depth’: 6, ‘eta’: 0. 3, ‘nfold’: 10, ‘best_nrounds’: 13 }
Random Forest	{‘eval_metric’: AUC, ‘mtry’: 5, ‘min.node.size’: 38, ‘sample.fraction’: 0.24 , ‘Iters’:30 ‘iters.warmup’:70 ‘num.trees’: 1000}	{‘eval_metric’: AUC, ‘mtry’: 10, ‘min.node.size’: 338, ‘sample.fraction’: 0.51, ‘Iters’:30 ‘iters.warmup’:70 ‘num.trees’: 1000}	{‘eval_metric’: AUC, ‘mtry’: 6, ‘min.node.size’: 24 ‘sample.fraction’: 0.24, ‘Iters’:30 ‘iters.warmup’:70 ‘num.trees’: 1000}	{‘eval_metric’: AUC, ‘mtry’: 9, ‘min.node.size’: 274 ‘sample.fraction’: 0.43, ‘Iters’:30 ‘iters.warmup’:70 ‘num.trees’: 1000}	{‘eval_metric’: AUC, ‘mtry’: 7, ‘min.node.size’: 173 ‘sample.fraction’: 0.56, ‘Iters’:30 ‘iters.warmup’:70 ‘num.trees’: 1000}
Elastic net	{‘lambda.min’: 0.001, ‘nfold’: 10 }	{‘lambda.min’: 0.0008, ‘nfold’: 10 }	{‘lambda.min’: 0.002, ‘nfold’: 10 }	{‘lambda.min’: 0.002, ‘nfold’: 10 }	{‘lambda.min’: 0.005, ‘nfold’: 10 }
Decision tree	{‘minsplit’: 120, ‘cp’: 0.001, ‘nfold’: 10 }	{‘minsplit’: 10, ‘cp’: 0.001, ‘nfold’: 10 }	{‘minsplit’: 30, ‘cp’: 0.001, ‘nfold’: 10 }	{‘minsplit’: 210, ‘cp’: 0.001, ‘nfold’: 10 }	{‘minsplit’: 100, ‘cp’: 0.001, ‘nfold’: 10 }

BART	{‘ntree’: 300, ‘k’: 2, ‘nskip’: 500, ‘ndpost’: 2, ‘nfold’: 10 }	{‘ntree’: 300, ‘k’: 3, ‘nskip’: 500, ‘ndpost’: 2, ‘nfold’: 10 }	{‘ntree’: 300, ‘k’: 3, ‘nskip’: 500, ‘ndpost’: 2, ‘nfold’: 10 }	{‘ntree’: 200, ‘k’: 2, ‘nskip’: 500, ‘ndpost’: 2, ‘nfold’: 10 }	{‘ntree’: 100, ‘k’: 3, ‘nskip’: 500, ‘ndpost’: 2, ‘nfold’: 10 }

Appendix 5: The Performance of Model Prediction for NRM

	AUC	Accuracy	Sensitivity	Specificity
XGB	0.607	0.595	0.561	0.601
Random forest	0.608	0.593	0.580	0.596
Decision tree	0.610	0.440	0.375	0.808
Elastic net	0.608	0.662	0.698	0.458
Logistic regression	0.607	0.746	0.339	0.818
Bayesian classifier	0.593	0.434	0.376	0.767
BART	0.618	0.541	0.517	0.675
Stacking	0.732	0.633	0.742	0.614

Appendix 6: The Performance of Model Prediction for Rejection

	AUC	Accuracy	Sensitivity	Specificity
XGB	0.660	0.573	0.730	0.559
RF	0.680	0.51	0.80	0.49
Decision tree	0.662	0.538	0.520	0.746
Elastic net	0.659	0.620	0.620	0.617
Logistic regression	0.658	0.585	0.663	0.578
Bayesian classifier	0.671	0.599	0.593	0.663
BART	0.681	0.503	0.478	0.798
Stacking	0.745	0.845	0.494	0.875

Appendix 7: The Performance of Model Prediction for Relapse

	AUC	Accuracy	Sensitivity	Specificity
XGB	0.660	0.573	0.730	0.559
RF	0.627	0.554	0.659	0.529
Decision tree	0.605	0.606	0.623	0.531
Elastic net	0.620	0.704	0.774	0.405
Logistic regression	0.623	0.725	0.371	0.808
Bayesian classifier	0.622	0.526	0.490	0.682
BART	0.626	0.653	0.687	0.508
Stacking	0.735	0.638	0.734	0.615

Appendix 8: The Performance of Model Prediction for aGvHD

	AUC	Accuracy	Sensitivity	Specificity
XGB	0.616	0.558	0.451	0.717
RF	0.628	0.592	0.592	0.594
Decision tree	0.603	0.593	0.566	0.611
Elastic net	0.625	0.594	0.615	0.580
Logistic regression	0.625	0.596	0.583	0.617
Bayesian classifier	0.612	0.573	0.628	0.537
BART	0.634	0.585	0.700	0.508
Stacking	0.700	0.613	0.504	0.774

Appendix 9: The Performance of Model Prediction for cGvHD

	AUC	Accuracy	Sensitivity	Specificity
XGB	0.561	0.555	0.567	0.549
RF	0.562	0.565	0.478	0.620
Decision tree	0.556	0.568	0.636	0.460
Elastic net	0.566	0.562	0.599	0.506
Logistic regression	0.567	0.558	0.554	0.560
Bayesian classifier	0.567	0.579	0.660	0.451
BART	0.563	0.549	0.520	0.594
Stacking	0.630	0.615	0.513	0.680

Abstract in French

Contexte : La greffe de cellules souches hématopoïétiques allogéniques (Allo-HSCT) est un traitement curatif des hémopathies malignes. Cependant, les patients restent exposés à des complications graves, notamment la mortalité non liée à une rechute (NRM), la rechute, le rejet, la maladie aiguë du greffon contre l'hôte (aGvHD) et la maladie chronique du greffon contre l'hôte (cGvHD). Une prédiction précoce et précise de ces issues peut soutenir la prise de décision clinique et améliorer le pronostic à long terme. Cette étude utilise les données pré-greffe de 16 427 patients atteints de maladies malignes, enregistrées dans le registre de la Société européenne de transplantation de moelle et de sang (EBMT) entre 2013 et 2023.

Objectifs : Cette étude vise à utiliser des algorithmes d'apprentissage automatique pour prédire la probabilité de NRM, rejet, rechute, aGvHD et cGvHD chez les patients dans l'année suivant une Allo-HSCT.

Méthodes : Nous avons développé et évalué huit algorithmes d'apprentissage automatique, dont la régression logistique, la forêt aléatoire, XGBoost (XGB), l'arbre de décision, l'elastic net, le classificateur bayésien, les arbres de régression additifs bayésiens (BART) et un modèle d'ensemble (stacking). Les données cliniques de patients ayant reçu une Allo-HSCT pour des maladies malignes ont été utilisées. Les performances des modèles ont été évaluées à l'aide de l'Aire sous la courbe (AUC) et de la précision. L'explication additive de Shapley (SHAP) a été utilisée pour interpréter l'impact des variables individuelles.

Résultats : Le modèle d'ensemble (stacking) a obtenu la meilleure AUC pour tous les résultats, avec les meilleures performances pour le rejet (0,745), suivi de la rechute (0,735), de la NRM (0,732), de l'aGvHD (0,700) et de la cGvHD (0,630). En termes de précision, le stacking s'est également classé premier pour la cGvHD (0,615), l'aGvHD (0,613) et le rejet (0,845), tandis que les modèles les plus performants pour la rechute et la NRM étaient respectivement l'elastic net (0,704) et la régression logistique (0,746). Pour le rejet, les variables les plus influentes étaient la date de traitement, la compatibilité HLA, l'état fonctionnel et le conditionnement. Pour la rechute, la date de traitement, l'indice de risque des maladies (DRI), le conditionnement et l'utilisation de la thiotépa étaient les plus importants. Pour l'aGvHD, la date de traitement, le type de donneur et la compatibilité HLA étaient les facteurs les plus déterminants. Pour la cGvHD, les variables clés comprenaient la date de traitement, l'état fonctionnel, l'utilisation de la thiotépa, la correspondance CMV et la compatibilité HLA. Pour la NRM, l'âge au moment du traitement, l'utilisation de la thiotépa et la date de traitement figuraient systématiquement parmi les variables les plus influentes.

Conclusion : Le modèle stacking a démontré les meilleures performances globales. Parmi les cinq issues étudiées, la date de traitement s'est révélée être la variable la plus importante de manière constante. L'apprentissage automatique montre un fort potentiel en tant qu'outil d'aide à la décision clinique.

MOTS-CLÉS : Allo-HSCT, Adulte, Maladies malignes, Apprentissage automatique, Prédiction des événements indésirables