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

Asmuth, E.G.J. von; Neven, B.; Albert, M.H.; Mohseny, A.B.; Schilham, M.W.; Binder, H.; ... ; Lankester, A.C.

Transplantation and Cellular Therapy

journal homepage: www.astctjournal.org

Full Length Article
Pediatric

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

Erik G.J. von Asmuth[1,*], Bénédicte Neven[2], Michael H. Albert[3], Alexander B. Mohseny[1], Marco W. Schilham[1], Harald Binder[4], Hein Putter[5], Arjan C. Lankester[1]

[1] Willem Alexander Children's Hospital, Leiden University Medical Center, Leiden, The Netherlands
[2] Pediatric Hematology and Immunology Unit, Necker Hospital for Sick Children, Assistance Publique-Hopitaux de Paris, Paris, France
[3] Department of Pediatrics, Dr. von Hauner Children's Hospital, University Hospital LMU Munich, Germany
[4] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany
[5] Department of Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands

Article history:
Received 15 August 2023
Accepted 7 September 2023

Key words:
Hematopoietic stem cell transplantation
Inborn errors
Pediatrics
Machine learning
Survival modeling

ABSTRACT

Allogeneic hematopoietic stem cell transplantation (HSCT) is a curative treatment for many inborn errors of immunity, metabolism, and hematopoiesis. No predictive models are available for these disorders. We created a machine learning model using XGBoost to predict survival after HSCT using European Society for Blood and Marrow Transplant registry data of 10,888 patients who underwent HSCT for inborn errors between 2006 and 2018, and compared it to a simple linear Cox model, an elastic net Cox model, and a random forest model. The XGBoost model had a cross-validated area under the curve value of .73 at 1 year, which was significantly superior to the other models, and it accurately predicted for countries excluded while training. It predicted close to 0% and >30% mortality more often than other models at 1 year, while maintaining good calibration. The 5-year survival was 94.7% in the 25% of patients at lowest risk and 62.3% in the 25% at highest risk. Within disease and donor subgroups, XGBoost outperformed the best univariate predictor. We visualized the effect of the main predictors—diagnosis, performance score, patient age and donor type—using the SHAP ML explainer and developed a stand-alone application, which can predict using the model and visualize predictions. The risk of mortality after HSCT for inborn errors can be accurately predicted using an explainable machine learning model. This exceeds the performance of models described in the literature. Doing so can help detect deviations from expected survival and improve risk stratification in trials.

INTRODUCTION

Allogeneic hematopoietic stem cell transplantation (HSCT) provides an often-curative treatment option for a diverse group of inborn errors, including inborn errors of immunity, metabolism, and hematopoiesis [1]. Within the European Society for Blood nd Marrow Transplant (EBMT), more than 1500 patients undergo HSCT annually for an inborn error, the most common nonmalignant indication for HSCT [2].

Outcomes of HSCT for inborn errors are highly dependent on the disease, with an 88% 2-year overall survival for thalassemia [3], an 81% 2-year survival for severe combined immunodeficiency (SCID) [4], and lower survival for familial hemophagocytic lymphohistiocytosis [5].

Models predicting survival after HSCT have various purposes, such as in trials, where multiple scores can be used to evaluate whether groups are equally distributed for mortality risk or to stratify patients according to this risk. In benchmarking, such models can adjust for differences in patient population among centers to fairly identify deviations from expected survival [6].

For malignancies, multiple algorithms exist for stratifying and predicting survival after HSCT, including algorithms that predict using underlying diagnosis and disease stage [7,8] or comorbidity [9]. Models also have been developed to adjust for multiple patient, donor, and transplantation characteristics, such as the EBMT risk score, which assigns points based on 5

Financial disclosure: See Acknowledgments on page 775.e7.

*Correspondence and reprint requests: Erik G. J. von Asmuth, Willem Alexander Children's Hospital, Leiden University Medical Center, Leiden, The Netherlands

E-mail address: e.g.j.von_asmuth@lumc.nl (E.G.J. von Asmuth).

https://doi.org/10.1016/j.jtct.2023.09.007

parameters [10], or an alternating decision tree model focused on 100-day mortality [11]. However, no such scores exist for inborn errors.

Survival prediction for patients with inborn errors is difficult, as patient groups are often small and risk factors may be disease- or disease category-specific. For example, in SCID, pretransplantation infection is a major factor influencing survival, requiring HSCT early in the first year of life [4,12], whereas in hemoglobinopathies, overall survival decreases with age throughout childhood into adulthood [3,13]. However, both disease categories follow the same donor hierarchy, with preference for a matched sibling donor.

A predictive model that integrates all inborn errors should be able to identify prognostic factors shared among all diagnoses, be able to adjust for them, and thus be able to determine remaining factors predictive of a disease category or a single disease more accurately, allowing for better predictions while working with small numbers of patients per diagnosis.

To achieve this, we evaluated 2 machine learning approaches: (1) a random survival forest, which fits many decision trees by fitting them on randomized subsets of the data and attempts to predict a patient-specific survival curve, and (2) an XGBoost model, which fits many decision trees through boosting, each one attempting to improve predictions of the previous one, and uses the Cox loss function to predict a patient-specific hazard ratio. We compared these approaches to each other and to the more generally used linear Cox model.

METHODS

We included patients from the EBMT registry who underwent first allogeneic HSCT for an inborn error between 2006 and 2018 to investigate determinants of survival. We retrieved survival data and patient, donor, and transplantation characteristics known at the time of HSCT from the registry. Given that some of the terminology discussed below will be unfamiliar to physicians without extensive knowledge of machine learning, a glossary explaining these terms has been included in the Supplementary Data.

To predict survival, we evaluated a simple linear Cox model without regularization or term interactions, an elastic net regularized Cox model with relevant interactions [14], a random survival forest model to fit nonparametric survival curves [15], and an XGBoost model optimizing the Cox negative log-likelihood [16]. Covariate selection was based on data availability and known effects in the literature (Supplementary Table S2).

For all models except the XGBoost model, MissForest imputation was used to account for missing data [17]. The XGBoost model can natively incorporate missing values. The full model configuration, including hyperparameter search space and optimization strategy, can be found in Supplementary Tables S1 and S2. In cases of near ties, hyperparameters that led to the least complex model were preferred. Iteration counts for both the XGBoost model and random forest model were determined by inspecting learning curves.

To compare models, we determined the area under the receiver operator characteristic curve (AUC) and calculated the changes in AUC, CIs, and *P* values using the riskRegression R package [18]. We used 10-fold cross-validation for internal validation and used country-wise cross-validation among 4 countries with the most inclusions to assess the ability of the model to extrapolate geographically, by refitting the model while excluding a country and then evaluating model fit on the excluded country.

To provide a benchmark on which models should improve, we assessed predictors univariately, and determined the best predictor in each disease category using Kaplan-Meier curves. To evaluate model benefit within disease categories, we compared the model AUC for that category against the best predictor for it.

To express the effect of the final model in a more intuitive way, we categorized cross-validated risks using the interquartile range (IQR) as low risk (below the IQR), intermediate risk (within the IQR), or high risk (above the IQR); determined survival per category; and investigated associations between risk categories and patient, donor, and transplantation characteristics. We used hierarchical disease coding to allow the decision tree based models to fit effects that hold for an entire disease category (eg, all hemoglobinopathies), a subgroup of a category (eg, all non-SCID inborn errors of immunity) or a single disease (Supplementary Data).

To explain machine learning predictions, we used SHAP values [19], an explainer that determines the influence of single covariates on model predictions for each patient while accounting for interactions, and evaluated these both globally and within single patients as examples. We also calculated the mean absolute SHAP value for each covariate to determine the importance of each covariate. Model reporting was done in accordance with the TRIPOD statement (Supplementary Data) [20].

RESULTS

Patient Characteristics

A total of 10,888 patients with survival data who underwent HSCT for an inborn error were included. Diagnosis groups consisted of hemoglobinopathies and inborn errors of immunity and metabolism. Among these, 28% of the patients had thalassemia, 27% had a non-SCID inborn error of immunity, 16% had SCID, 8% had a histiocytic disorder, 10% had an inborn error of metabolism, and 1% had a congenital bone marrow failure syndrome. Donors were mostly HLA-identical siblings (43%) and unrelated donors (38%). Bone marrow was the most common stem cell source (61% of cases), with peripheral blood used in 25% of cases and cord blood used in 12% of cases (Table 1).

When splitting data in groups and predicting using survival curves, diagnosis was the best predictor of survival, with an AUC of .63 (95% CI, .62 to .65) at 1 year, followed by donor type, performance score at SCT, and stem cell source (Supplementary Table S3). Within diagnosis subgroups, the best predictor varied, with AUC between .55 and .63 and donor type, stem cell source and CMV matching the most common best predictors within subgroups (Supplementary Table S4).

Model Comparison

We fit and compared a random forest model, an XGBoost model, and 2 linear Cox models and investigated model performance. At 1 year, the XGBoost model was the best-performing model, with a cross-validated AUC of .728 (95% CI, .714 to .742), followed by the random forest model at .713 (95% CI, .698 to .727). The XGBoost model yielded significantly more accurate predictions than the Cox, elastic net Cox, and random forest models at 1 year ($P < .001$ against all models). At 5 years, the XGBoost and random forest model performed equally (ΔAUC, -.004; 95% CI, -.004 to .013: $P = .3$). Both the simple and elastic net Cox models performed markedly worse at all time points. In all models, a drop in accuracy over time can be seen (Table 2).

Cross-validated model calibration was excellent for predicted mortality between 0 and 25% for all models, above which the Cox-based models overestimated mortality, whereas the random forest model underestimated mortality

Table 1
Demographic Characteristics of the Included Patients (N = 10,888)

Characteristic	Value
Disease category, n (%)	
Inborn errors of immunity	4682 (43)
Hemoglobinopathies	4144 (38)
Inborn errors of metabolism	1055 (9.7)
Histiocytic disorders	877 (8.1)
Congenital bone marrow failure	130 (1.2)
Age at transplantation, yr, median (IQR)	4 (1-10)
Unknown, n	1
Stem cell source, n (%)	
Bone marrow	6655 (61)
PB	2731 (25
CB	1290 (12)
Combined grafts	212 (1.9)
Donor type, n (%)	
Identical sibling	4613 (43)
Unrelated	4127 (38)
Mismatched relative	1247 (12)
Matched other relative	855 (7.9)
Unknown	46
Performance status, n (%)	
100	3435 (45)
90	2292 (30)
80	1140 (15)
70-10	724 (9.5)
Unknown	3297
CMV serostatus, patient/donor, n (%)	
+/+	4458 (49)
-/-	1913 (21)
+/-	1335 (15)
-/+	1323 (15)
Unknown	1859
Patient sex, n (%)	
Male	6609 (61)
Female	4279 (39)
Donor sex, n (%)	
Male	5698 (54)
Female	4874 (46)
Unknown	316

(Supplementary Figure S1A, B). The simple Cox and elastic net Cox model had a similar prediction distribution, whereas the XGBoost model identified a very low-risk group, as well as a more high-risk group. The random forest model mainly differentiated between intermediate-risk and high-risk patients (Supplementary Figure S1C, D).

In the XGBoost model, patients are assigned a single hazard, as it uses the Cox loss function. However, for the random survival forest model, the risk for one patient compared to another can change over time, because the model predicts nonparametric survival curves. This makes the predictions of the XGBoost model easier to interpret and apply, and thus we further dissected its performance.

When performing country-wise cross-validation, the XGBoost model performed consistently with internal validation. The model performed better than internal validation in France and did not show a drop in accuracy over time for that country; however, it performed slightly worse than internal validation in the United Kingdom (Table 3).

When analyzing performance between countries, we found that performance varied across disease groups, with exceptionally high model accuracy for hemoglobinopathies in France and high accuracy for inborn errors of immunity and metabolism in Turkey. Within the UK, performance in each disease group was in line with internal validation (Supplementary Table S5); however, survival was significantly higher for the inborn errors of immunity and metabolism group ($P < .001$), whereas survival of other disease groups was in line with other countries. This could cause the model to incorrectly compare patients with an inborn error of immunity and metabolism with patients with other diagnoses, explaining why discriminative performance was reduced when globally evaluating the model but not when looking at specific disease groups (Supplementary Table S6).

Model Performance

When categorizing cross-validated risk as low (below the IQR), intermediate (within the IQR), and high (above the IQR), 1-year survival was 96.2% (95% CI, 95.4% to 97.0%) in the low-risk group, 87.7% (95% CI, 86.8% to 88.6%) in the intermediate-risk group, and 69.0% (95% CI, 67.2% to 70.9%) in the high-risk group, an 8.1-fold higher mortality compared with the low-risk group. At 5 years, survival in the 3 groups was 94.7% (95% CI, 93.7% to 95.7%), 83.1% (955 CI, 81.9% to 84.3%), and 62.3% (95% CI, 6.2% to 64.4%), respectively. The simple Cox model revealed a ratio of only 6:1 for similarly categorized predictions between the high-risk and low-risk groups at 1 year (Figure 1A, Supplementary Table S7).

The low-risk group consisted mostly of hemoglobinopathies, with 75% of sickle cell patients and 48% of thalassemia patients included in this group. Only 6% of identical sibling transplants fell in the high-risk group, whereas 63% of mismatched relative transplants were categorized as such. However, when examining risk within specific diagnosis and donor categories, considerable additional variation caused by other covariates remained (Figure 1B, C, Supplementary Table S8).

When analyzing which covariates had the most influence on the model, we used mean absolute SHAP values, which

Table 2
Discriminative Performance of Evaluated Models, Measured as AUC, at 1, 2, and 5 Years Post-HSCT and the Decrease in AUC when Choosing this Model vs the Best Performing Model at that Time Point, Calculated Using 10-Fold Cross-Validation

Model	AUC at 1 yr	Performance vs best predictor at 1 yr	AUC at 2 yr	Performance vs best predictor at 2 yr	AUC at 5 yr	Performance vs best predictor at 5 yr
Simple Cox model	.697 (.683-.712)	.031 (.022-.040); $P \leq .001$	.688 (.673-.703)	.028 (.019-.038); $P \leq .001$	.679 (.662-.696)	.019 (.010- 028); $P \leq .001$
Elastic net Cox model	.698 (.684-.713)	.030 (.021-.039); $P \leq .001$	.686 (.671-.701)	.030 (.021-.039); $P \leq .001$	.667 (.650-.684)	.031 (.022-.040); $P \leq .001$
Random forest	.713 (.698-.727)	.016 (.009-.023); $P \leq .001$	.705 (.690-.720)	.011 (.004-.018); $P = .003$	.698 (.682-.715)	Best predictor
XGBoost	.728 (.714-.742)	Best predictor	.716 (.701-.731)	Best predictor	.694 (.677-.710)	.004 (.013-.004); $P = .3$

Table 3
Geographical Validation of Discriminative Performance of the XGBoost Model for Countries with >1000 Included Subjects, Measured as AUC, at 1, 2, and 5 Years Post-HSCT

Country	AUC at 1 yr (95% CI)	AUC at 2 yr (95% CI)	AUC at 5 yr (95% CI)
France (n = 1042)	.771 (.737-.805)	.777 (.742-.811)	.768 (.724-.811)
Italy (n = 1494)	.745 (.704-.785)	.710 (.669-.751)	.674 (.630-.718)
Turkey (n = 1245)	.746 (.700-.793)	.722 (.672-.771)	.689 (.622-.756)
United Kingdom (n = 1240)	.706 (.664-.748)	.685 (.643-.728)	.653 (.608-.698)

Note that a lower AUC is not associated with better or worse survival than expected, only with less accurate survival predictions.

allowed us to integrate multiple levels of a single covariate and analyze diagnoses using hierarchical disease coding. Predictions were driven mainly by diagnosis, donor type, and patient performance status score and age. CMV serostatus matching, conditioning agents used, and stem cell source were of secondary importance (Figure 2A).

Examining SHAPs for the most important predictors univariately revealed that the model properly captured well-known facts: hemoglobinopathies are lower risk than other inborn errors, a higher performance score at time of transplantation is favorable, and having an identical sibling donor improves survival. We found that younger age appeared favorable, but that additional heterogeneity was caused by other covariates (Figure 2B, C).

Within the univariate SHAP plots, chronic granulomatous disease (CGD) and Wiskott- Aldrich syndrome (WAS) appeared to result in favorable predictions. When analyzing risk groups, 348 out of 2924 patients with a non-SCID inborn error of immunity were in the low-risk group, with 242 of these patients having either CGD or WAS, indicating that these diagnoses have more favorable outcomes after HSCT compared to other non-SCID inborn errors of immunity.

When inspecting interactions, the effect of age varied with the graft source, with the increased risk at a older age reduced for patients receiving a peripheral blood stem cell graft, whereas patients receiving such a graft at a younger age had a reduced benefit of their younger age (Figure 2C). We also saw that inborn errors of immunity or metabolism were associated

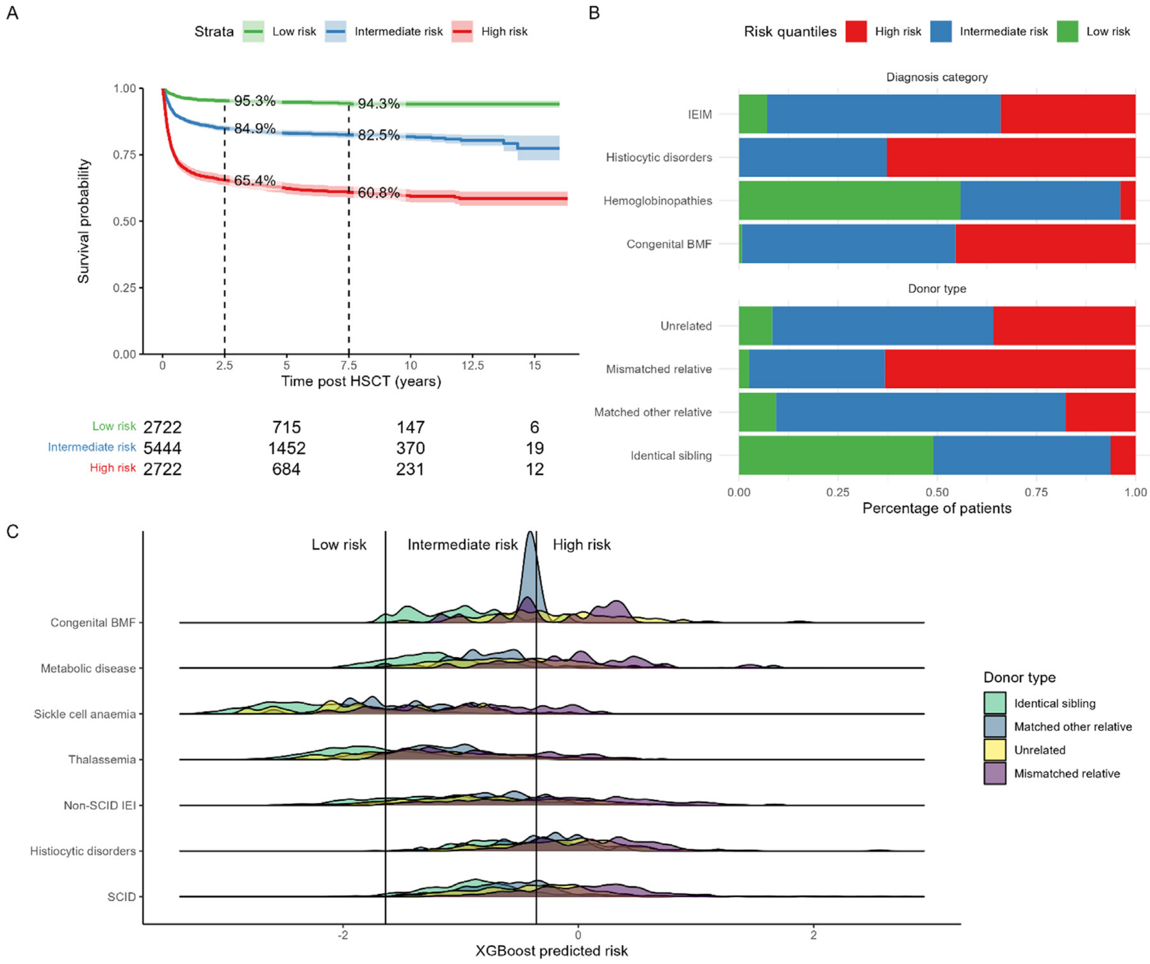


Figure 1. (A) Survival according to cross-validated predicted risk quantiles using the XGBoost model, including number at risk and survival probabilities at 2.5 and 5 years. (B) Predicted risk quantiles by diagnosis and donor group. (C) Risk density plot for the linear predictor by disease subcategory and donor relation, with lines representing the cutoffs for risk quantiles.

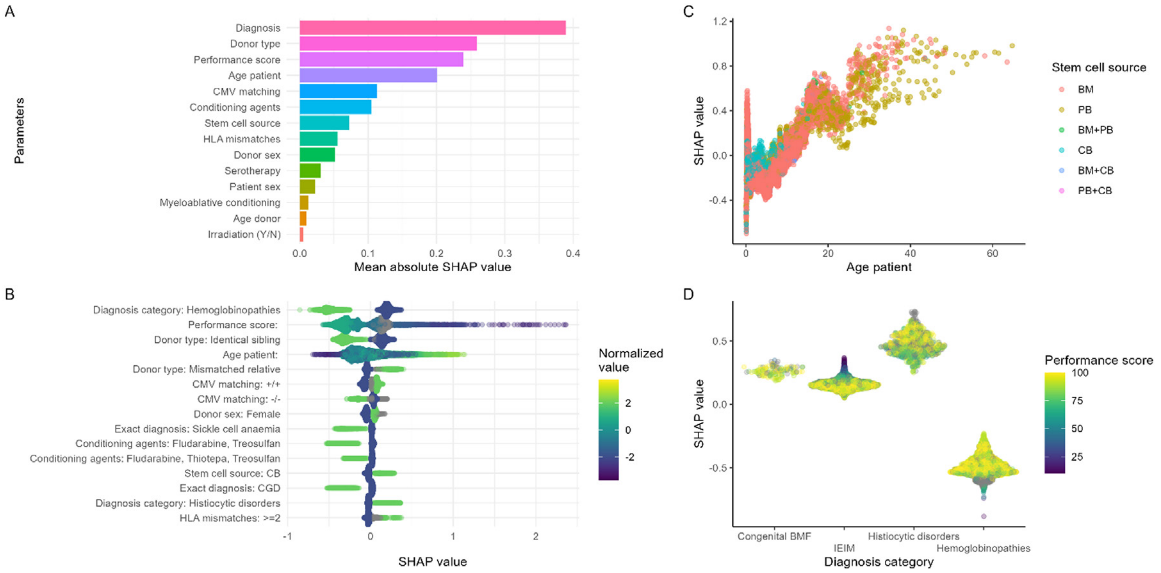


Figure 2. (A) Mean absolute SHAP value for the top 15 parameters with the most influence in the model. Note that the model can consider HLA matching only for patients with an unrelated or matched other related donor, which decreases its effect when considering the model as a whole. Parameters with a high missing variable count (eg, donor age) also will be less informative and thus have less influence on the model. (B) SHAP value for each patient for each one-hot encoded parameter for the top 15 parameters. SHAP values are equivalent to a change in the XGBoost linear predictor; negative values mean a lower risk of death and thus better survival. (C) SHAP plot for the influence of age on predicted survival, color-coded by graft source. Of interest is that when a peripheral blood stem cell graft is used, the SHAP value increases at age <20 years but decreases at age >20 years. (D) SHAP plot for the influence of diagnosis category, color-coded by performance score. A lower performance score appears to have a more negative effect in the inborn error of immunity and metabolism group than in other diagnosis categories.

with especially poor outcome in patients with a low performance status score at time of transplantation (Figure 2D).

Application Development

To make the model usable for future research, we have made the complete model object available, including code on how to use the model. To allow researchers inexperienced with machine learning to use the model, we include a standalone offline application for the model that can predict HSCT mortality risk, visualize these predictions, and provide both global and disease category-specific references on where this prediction falls relative to other predictions included in the research cohort (Supplementary Data).

The application is structured as follows. On the left is a data entry panel, which allows users to enter data as entered in the EBMT registry (Figure 3A). To allow the user to view the contribution of parameters to the current prediction, the tool offers a SHAP waterfall plot of the top 6 contributing parameters. The parameters shown can change, depending on the data, as interactions can change the role of specific parameter values (Figure 3B).

The application also draws an overall survival curve, which plots the estimated survival given the entered data and curves at point predictions of the median risk and the 10% and 90% risk quantiles. It also reports the risk level of the patient in relation to both the overall cohort and patients within the same disease category (Figure 3C, D).

DISCUSSION

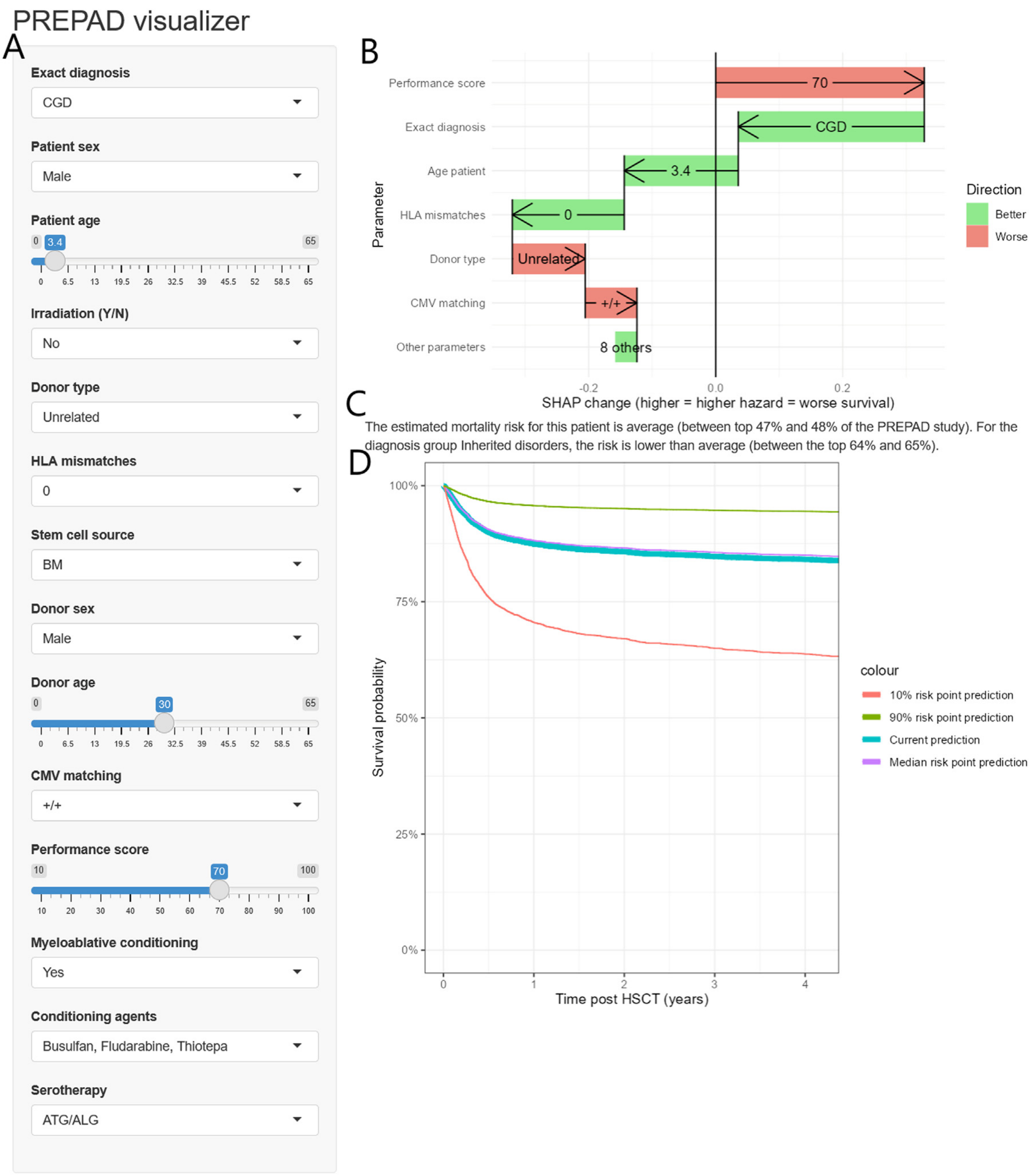
For patients with inborn errors, survival following HSCT is dependent on both general and disease-specific factors, and some factors, such as age, are not linearly associated with survival. Here we have presented a machine learning model using XGBoost, which can accurately predict survival for these patients with good calibration and geographical extrapolation, by modeling the effects of patient, donor, and transplant characteristics and their interactions using boosted decision trees.

This model is the first published prognostic model that predicts survival after HSCT for inborn errors, whereas such models for malignancies are common. This population proved especially challenging because it cannot be assumed that prognostic indicators are shared among all diagnoses, and because some diagnoses are extremely rare. Thus, we embedded knowledge on diagnosis categorization in model parameters, allowing the model to fit general effects, effects that hold for a disease category, and effects that hold only for a specific disease.

We envision several applications for this model, such as evaluating equal distribution of risk when performing retrospective trials, stratifying patients based on risk when performing interventional trials, or in quality management to detect deviations from expected survival. However, it is not a causal model and so cannot be used to infer causal relations regarding determinants of survival and their interaction. It also is not meant to be applied in patient care, which would require a costly certification process and constant evaluation of model accuracy in patients in whom it is used, which is not feasible in the context of retrospective registry studies.

In gene therapy for inborn errors as an alternative to HSCT, which is starting to become available for specific forms of SCID [21,22] and is being studied for many different inborn errors [23], model predictions could be used to compare the results of gene therapy with expected results if HSCT were performed instead, adjusting for a patient's specific risk factors, and results could be tested against multiple hypothetical donors to identify in which cases gene therapy would be most beneficial.

Regarding discriminative performance, a previous study comparing multiple prognostic scores for hematologic malignancies identified the revised pretransplantation assessment of mortality as the best performing score, reaching an AUC of .64 at 2 years, with other scores reaching AUCs between .63



PREPAD visualizer

Exact diagnosis
CGD

Patient sex
Male

Patient age
3.4

Irradiation (Y/N)
No

Donor type
Unrelated

HLA mismatches
0

Stem cell source
BM

Donor sex
Male

Donor age
30

CMV matching
+/+

Performance score
70

Myeloablative conditioning
Yes

Conditioning agents
Busulfan, Fludarabine, Thiotepa

Serotherapy
ATG/ALG

The estimated mortality risk for this patient is average (between top 47% and 48% of the PREPAD study). For the diagnosis group Inherited disorders, the risk is lower than average (between the top 64% and 65%).

Figure 3. PREPAD visualizer application. (A) Parameter entry pane containing patient, donor, and transplantation characteristics. (B) SHAP waterfall plot of the top 6 parameters with the most influence on the current prediction. (C) Current estimated risk percentage for the patient, related to the entire cohort and the risk subgroup. (D) Estimated survival in years according to the XGBoost model, in relation to the median and 10% and 90% risk quantile point predictions.

and .58 at 2 years, and more recent studies show similar performance [8,24,25]. Considering this, an AUC of .716 at 2 years is a big step forward.

Compared to the simple Cox model, the XGBoost model increased the AUC by .028 at 2 years, with an even larger difference at 1 year. For reference, adjusting the HSCT Comorbidity Index for use in nonmalignant patients improved its predictive performance from .643 to .649 at 2 years, and the revised version of the Disease Risk Index showed a C-index of .643, compared to .637 for the original, with both indices increasing the C-index by .006 [26]. This indicates that while numerically small, the relative improvement in accuracy from using XGBoost is substantial.

A limitation of the AUC is that it is strongly dependent on the cohort, which is also demonstrated by the variability between countries in external validation. No existing scores apply to our cohort, leaving the AUC difficult to interpret, given that the only reference points are the other approaches that we tried. Model categorization is easier to interpret, with 3.8% of patients dying in the low-risk group at 1 year versus 31% in the high-risk group (an 8.1-fold difference), whereas the simple Cox model showed only a 6.1-fold difference in survival between similarly created groups.

These risk groups were created arbitrarily, because the predicted risk was distributed smoothly. Dividing the predicted risk into more groups would be equally valid and could allow for more fine-grained risk categorization. However, when applying the model, exact predicted survival should be used instead of relying on categorization to achieve the model's full discriminative potential. We ensured that this was feasible by freely providing the model object as well as an example application that can be used to enter patient characteristics and form predictions.

The added model complexity, with thousands of estimates instead of tens at most in previous models, allowed us to achieve this performance but comes at the cost of interpretability. We used SHAP values to investigate the modeled effects and their interactions. The example application allows users to enter specific patients, investigate the main determinants, see the prediction and the associated SHAP explanation, and change parameters on the fly to see the change in prediction and the explanation. However, these explainers should not be misinterpreted as causal effects. For example, a relatively modest increase in mortality at a higher patient age might be modeled owing to a less aggressive course of disease for patients of older age not needing transplantation earlier in life, and thus should not be used as an argument to delay transplantation. Similarly, the excellent results using an HLA-identical sibling donor may be augmented by the benefit of HSCT more often outweighing the risks if an identical sibling donor is available, causing low-risk patients to be treated with an identical sibling donor more often.

Although simple scores can be calculated using only weights that can be published in a table, our scores cannot be calculated in this manner, and so an application to calculate risk is needed. When designing this application, future availability was a key concern. Sustainably maintaining a web application that processes patient data available is a challenge in academia, as demonstrated by the fact that 2 previously available applications are no longer available at the time of this report [11,27]. To avoid this, we took an approach in which the application can be run both online and offline using Shiny [28], which by being open source serves as a reference implementation to easily allow future use in large datasets.

A limitation of the model lies in data collection, which ended in 2018 to ensure a decent duration of follow-up. Innovations and changes in HSCT practice over time may change the accuracy of the model, necessitating refitting. In an analysis that included year of transplantation as a covariate, prediction accuracy did not improve, but we cannot anticipate if this may change in a more recent or future dataset. For example, the increasing use of post-transplantation cyclophosphamide for HLA-mismatched transplantation may change the influence of donor type on survival probability. Coronavirus disease 2019 (COVID-19) also might have an impact, both because patients may be at risk for COVID-19 itself, and because of the increased use of cryopreserved grafts during the COVID-19 pandemic [29].

Missing data remained a challenge during model development, and the ability of XGBoost to incorporate missing data into the model and make predictions based on incomplete data may be one reason for its superior performance over the other modeling approaches, which rely on imputation of missing data both for fitting and for prediction. Ideally, we could evaluate the model against a complete and correct dataset to see whether performance would change; however, obtaining such a dataset while remaining representative of clinical practice, and thus not excluding patients, is currently unfeasible. We hope that with current efforts to increase the reliability of registry data, future models can be developed without the need to rely on handling of missing data and the inherent uncertainty associated with it, and thus become even more accurate.

ACKNOWLEDGMENTS

Financial disclosure: This study was not externally funded. The authors declare no relevant financial interests.

Conflict of interest statement: There are no conflicts of interest to report.

Authorship statement: H.P. and A.C.L. contributed equally to this work.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.jtct.2023.09.007.

REFERENCES

1. Snowden JA, Sánchez-Ortega I, Corbacioglu S, et al. Indications for haematopoietic cell transplantation for haematological diseases, solid tumours and immune disorders: current practice in Europe, 2022. *Bone Marrow Transplant*. 2022;57:1217-1239.
2. Passweg JR, Baldomero H, Chabannon C, et al. Hematopoietic cell transplantation and cellular therapy survey of the EBMT: monitoring of activities and trends over 30 years. *Bone Marrow Transplant*. 2021;56:1651–1664.
3. Baronciani D, Angelucci E, Potschger U, et al. Hemopoietic stem cell transplantation in thalassemia: a report from the European Society for Blood and Bone Marrow Transplantation Hemoglobinopathy Registry, 2000–2010. *Bone Marrow Transplant*. 2016;51:536–541.
4. Lankester AC, Neven B, Mahlaoui N, et al. Hematopoietic cell transplantation in severe combined immunodeficiency: the SCETIDE 2006-2014 European cohort. *J Allergy Clin Immunol*. 2022;149:1744–1754. e8.
5. Bergsten E, Horne A, Hed Myrberg I, et al. Stem cell transplantation for children with hemophagocytic lymphohistiocytosis: results from the HLH-2004 study. *Blood Adv*. 2020;4:3754–3766.
6. Snowden JA, Saccardi R, Orchard K, et al. Benchmarking of survival outcomes following haematopoietic stem cell transplantation: a review of existing processes and the introduction of an international system from the European Society for Blood and Marrow Transplantation (EBMT) and the Joint Accreditation Committee of ISCT and EBMT (JACIE). *Bone Marrow Transplant*. 2020;55:681–694.
7. Armand P, Gibson CJ, Cutler C, et al. A disease risk index for patients undergoing allogeneic stem cell transplantation. *Blood*. 2012;120:905–913.
8. Shouval R, Fein JA, Labopin M, et al. Development and validation of a disease risk stratification system for patients with haematological malignancies: a retrospective cohort study of the European Society for Blood and Marrow Transplantation registry. *Lancet Haematol*. 2021;8:e205–e215.
9. Sorror ML, Maris MB, Storb R, et al. Hematopoietic cell transplantation (HCT)-specific comorbidity index: a new tool for risk assessment before allogeneic HCT. *Blood*. 2005;106:2912–2919.
10. Gratwohl A, Stern M, Brand R, et al. Risk score for outcome after allogeneic hematopoietic stem cell transplantation: a retrospective analysis. *Cancer*. 2009;115:4715–4726.
11. Shouval R, Labopin M, Bondi O, et al. Prediction of allogeneic hematopoietic stem-cell transplantation mortality 100 days after transplantation using a machine learning algorithm: a European Group for Blood and Marrow Transplantation Acute Leukemia Working Party retrospective data mining study. *J Clin Oncol*. 2015;33:3144–3151.
12. Haddad E, Logan BR, Griffith LM, et al. SCID genotype and 6-month posttransplant CD4 count predict survival and immune recovery. *Blood*. 2018;132:1737–1749.
13. Hsieh MM, Fitzhugh CD, Tisdale JF. Allogeneic hematopoietic stem cell transplantation for sickle cell disease: the time is now. *Blood*. 2011; 118:1197–1207.
14. Simon N, Friedman JH, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw*. 2011;39:1–13.
15. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2:841–860.
16. Chen TQ, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 13-17, 2016; San Francisco, CA.
17. Stekhoven DJ, Bühlmann P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28:112–118.
18. Gerds TA, Kattan MW. *Medical Risk Prediction Models: With Ties to Machine Learning*. New York, NY: Chapman and Hall/CRC; 2021.
19. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA; 2017.

20. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162:55–63.
21. Aiuti A, Roncarolo MG, Naldini L. Gene therapy for ADA-SCID, the first marketing approval of an ex vivo gene therapy in Europe: paving the road for the next generation of advanced therapy medicinal products. *EMBO Mol Med*. 2017;9:737–740.
22. Cicalese MP, Ferrua F, Castagnaro L, et al. Update on the safety and efficacy of retroviral gene therapy for immunodeficiency due to adenosine deaminase deficiency. *Blood*. 2016;128:45–54.
23. Naldini L, Cicalese MP, Bernardo ME, et al. The EHA research roadmap: hematopoietic stem cell gene therapy. *Hemasphere*. 2022;6:e671.
24. Shouval R, Fein JA, Shouval A, et al. External validation and comparison of multiple prognostic scores in allogeneic hematopoietic stem cell transplantation. *Blood Adv*. 2019;3:1881–1890.
25. Fattinger N, Roth JA, Baldomero H, et al. External validation of the revised Pretransplant Assessment of Mortality score in allogeneic hematopoietic cell transplantation: a cohort study. *Hemasphere*. 2022;6:e704.
26. Armand P, Kim HT, Logan BR, et al. Validation and refinement of the Disease Risk Index for allogeneic stem cell transplantation. *Blood*. 2014; 123:3664–3671.
27. Au BKC, Gooley TA, Armand P, et al. Reevaluation of the Pretransplant Assessment of Mortality score after allogeneic hematopoietic transplantation. *Biol Blood Marrow Transplant*. 2015;21:848–854.
28. Chang W, Cheng J, Allaire JJ, et al. shiny: Web Application Framework for R. 2021.
29. Passweg JR, Baldomero H, Chabannon C, et al. Impact of the SARS-CoV-2 pandemic on hematopoietic cell transplantation and cellular therapies in Europe 2020: a report from the EBMT activity survey. *Bone Marrow Transplant*. 2022;57:742–752.