RESEARCH ARTICLE

AJH WILEY

Time-dependent prediction of mortality and cytomegalovirus reactivation after allogeneic hematopoietic cell transplantation using machine learning

Lisa Eisenberg[1,2] | the XplOit consortium | Christian Brossette[3] |
Jochen Rauch[4] | Andrea Grandjean[5] | Hellmut Ottinger[6] | Jürgen Rissland[7] |
Ulf Schwarz[8] | Norbert Graf[3] | Dietrich W. Beelen[6] | Stephan Kiefer[4] |
Nico Pfeifer[1,2] | Amin T. Turki[6]

[1]Department of Computer Science, University of Tübingen, Tübingen, Germany

[2]Institute of Bioinformatics and Medical Informatics (IBMI), University of Tübingen, Tübingen, Germany

[3]Department of Pediatric Oncology and Hematology, Saarland University, Homburg, Germany

[4]Department of Biomedical Data & Bioethics, Fraunhofer Institute for Biomedical Engineering (IBMT), Sulzbach, Germany

[5]Averbis GmbH, Freiburg, Germany

[6]Department of Hematology and Stem Cell Transplantation, University Hospital Essen, Essen, Germany

[7]Institute of Virology, Saarland University Medical Center, Homburg, Germany

[8]Institute for Formal Ontology and Medical Information Science (IFOMIS), Saarland University, Saarbrücken, Germany

Correspondence
Amin T. Turki, Department of Hematology and Stem Cell Transplantation & Computational Hematology Lab, University Hospital Essen, Essen 45147, Germany. Nico Pfeifer, Department of Computer Science, University of Tübingen, Tübingen 72076, Germany.
Email: amin.turki@uk-essen.de; nico.pfeifer@uni-tuebingen.de

Abstract

Allogeneic hematopoietic cell transplantation (HCT) effectively treats high-risk hematologic diseases but can entail HCT-specific complications, which may be minimized by appropriate patient management, supported by accurate, individual risk estimation. However, almost all HCT risk scores are limited to a single risk assessment before HCT without incorporation of additional data. We developed machine learning models that integrate both baseline patient data and time-dependent laboratory measurements to individually predict mortality and cytomegalovirus (CMV) reactivation after HCT at multiple time points per patient. These gradient boosting machine models provide well-calibrated, time-dependent risk predictions and achieved areas under the receiver-operating characteristic of 0.92 and 0.83 and areas under the precision–recall curve of 0.58 and 0.62 for prediction of mortality and CMV reactivation, respectively, in a 21-day time window. Both models were successfully validated in a prospective, non-interventional study and performed on par with expert hematologists in a pilot comparison.

---

1 | INTRODUCTION

Allogeneic hematopoietic cell transplantation (HCT) is an effective and potentially curative treatment for patients suffering from high-risk hematological malignancies and other non-malignant and congenital disorders.[1] Despite its success and continuous improvement over the past decades,[2,3] the treatment-related non-relapse mortality (NRM) after HCT remains high. HCT recipients are at risk for multiple potentially life-threatening complications, such as graft-versus-host disease (GVHD) or cytomegalovirus (CMV) reactivation. Accurate risk assessment and an appropriate choice of prophylactic and pre-emptive treatments are crucial to minimize these risks.[4,5] Registries such as the databases of the European Society for Blood and Marrow Transplantation (EBMT) or of the Center for International Blood and Marrow Transplant Research (CIBMTR) collect individual patients' pre-HCT and outcome data from centers via standardized reporting forms.[6,7] Using these databases, the prevalence and risk factors of HCT complications can be analyzed on a large scale. Due to the data collection process, registry data per patient is limited to a set of categorical variables. While time-dependent endpoint data are available regarding the time of relapse or death, continuously measured laboratory values from electronic health records (EHR), or unstructured data from reports cannot yet be integrated into these registries. Since the 2000s, a number of relevant predictive risk scores have been developed utilizing static registry data to improve outcome assessment before HCT and to adjust the toxicity of the intervention by reducing the conditioning intensity. The hematopoietic cell transplantation-specific comorbidity index (HCT-CI)[8] is to date the most relevant and utilized score to predict NRM. Other Cox-regression models based on categorical, pre-HCT variables, such as the EBMT risk score[9] or the disease risk index,[10] have additionally improved pre-HCT and relapse risk assessment for different hematologic malignancies. However, the overwhelming majority of existing methods for assessing such HCT-specific risks offer only a single risk assessment before HCT.

Across medical areas, machine learning (ML) techniques have proven their value as powerful tools for diagnosis[11–14] or risk assessment.[15–17] ML models are ideally suited to discover associations in large datasets and can automatically identify important parameters and relationships between them without the need for a predefined model shape. In recent years, several ML models have been proposed for HCT-specific risk assessment at a single point in time.[18–20] For instance, an alternating decision tree model produced more accurate predictions of 100-day mortality after HCT than the EBMT score for acute leukemia patients,[18] demonstrating that ML can improve standard scores for pre-HCT risk assessment.

The endothelial activation and stress index (EASIX) measured before conditioning therapy is associated with overall survival after HCT, highlighting the potential of including laboratory parameters in pre-HCT risk assessment.[21] Additionally, EASIX measured at the onset of acute GVHD predicts overall survival after GVHD onset.[22] Despite its added value, EASIX is calculated from a limited set of three parameters (creatinine, platelets, LDH) using a predefined formula, and each study only evaluated its prognostic value at a single point in time.

Integrating time-dependent measurements into ML models can not only improve predictive performance but also allows to update risk assessments whenever new data become available. For instance, early-warning systems developed for intensive care units (ICU) continuously monitor patient data and predict critical events such as acute kidney injury[16] or circulatory failure,[17] which may help physicians to react earlier to critical events or to prevent them. Given the high variability of individual outcomes after HCT and the importance of optimal patient management, we hypothesized that ML-based models for precise, time-dependent risk prediction after HCT may provide a valuable tool to support treatment decisions. Compared with the large, annotated, public EHR datasets of ICU patients,[17,23] time-dependent HCT data are scarce. Their use in ML models is further challenged by high variability in laboratory measurement frequencies and a characteristic nonlinear development of laboratory values after HCT, which requires context-dependent evaluation of identical numerical results. In addition, longer observation times may entail missing values and censored data. Major national and international efforts are currently directed toward digitizing medicine,[24,25] developing unified standards for data management, and facilitating the increasingly widespread use of EHR systems. As a consequence, we expect the accessibility and usability of health data to improve, with impacts on different fields of medicine including HCT care.

In this article, we describe the development and prospective validation of ML models, which accurately predict death and early CMV reactivation at multiple time points after HCT. These are the first models for continuous time-dependent risk assessment of these outcomes after HCT.

2 | METHODS

2.1 | Patients

Between January 2005 and June 2020, 2191 patients with hematologic malignancies, inherited stem cell disorders, or acquired bone marrow failure underwent allogeneic HCT in the Department of Hematology and Stem Cell Transplantation of the West-German Cancer Center at University Hospital Essen (UHE). Patients with HCT before September 1, 2017 were included in the retrospective cohort. Patients with HCT between September 2017 and June 2020 were prospectively recruited into the non-interventional XplOit validation study (Figure S1). We excluded patients with multiple allogeneic HCTs, with hemoglobinopathies or without data on relevant laboratory tests, resulting in retrospective and prospective cohort sizes of

1710 and 403 patients, respectively. For models and analyses related to CMV reactivation, we additionally excluded patients without CMV data before day +30 after HCT. Donors were HLA-matched related donors (MRD, 23.0%), haploidentical related donors (haplo, 3.8%), 10/10 HLA-A-, -B, -C, -DRB1, -DQB1 matched unrelated donors (MUD, 53.6%), or mismatched unrelated donors (MMUD, 19.6%; Table S1). HLA-DPB1 was not considered for donor–recipient matching. Typically, patients were followed up for 60 months after transplantation. Long-term surviving patients were censored. Early supportive and follow-up care was identical for all patients. In the retrospective cohort, the predominant calcineurin inhibitor based GVHD prophylaxis consisted of Ciclosporin A plus Methotrexate. Patients with higher GVHD risk were assigned to additional in vivo T cell depletion using anti-T-Lymphocyte globulin (ATG) based on standardized clinical treatment protocols.

2.2 | Ethics

This study was conducted in accordance with German legislation and the revised Helsinki Declaration. Study design and data acquisition were evaluated by the institutional review board (IRB) of the University Duisburg-Essen (Protocol No. 17-7576-BO) and by the IRB of the medical association of the Saarland (Protocol No. 33/17). All patients included in the prospective, non-interventional XplOit study (registered in the German Clinical Trials Register (DRKS), registration No. DRKS00026643) have given written consent to collection, electronic storage, and scientific analysis of pseudonymized HCT-specific patient data.

2.3 | Data preparation, endpoint assessment, and statistical analysis

The sections on data preparation, endpoint assessment, and statistical analysis are detailed in the supplementary material.

2.4 | Preprocessing

We selected 60 features for model development, including all static features available in structured format, the prediction day, and 34 of the most frequently performed laboratory tests (Table S3). For time-dependent laboratory tests, we only used the most recent value of each parameter at the time point of prediction. Static and time-dependent features were preprocessed separately and concatenated directly before model training. Preprocessing is detailed in the supplement.

2.5 | Prediction times and classification target

We aimed for the application scenario where models are executed once per day whenever new time-dependent data become available.

Therefore, we considered all days between HCT and the event of interest (or censoring) where any laboratory measurements were reported as potential prediction days.

For each event (death or CMV reactivation), we defined binary classification targets based on two different window sizes d of 7 and 21 days, respectively. Each time point was labeled with 1 (positive) if the event occurred within the following d days and 0 (negative) otherwise. We excluded time points where patients were censored in this time window or where more than 30% of time-dependent features were missing after forward filling. For prediction of CMV reactivation, we considered only events in the first 100 days and excluded prediction days after day $100 - d$. The final number of time points is listed in Table S5 for each prediction task.

2.6 | Machine learning models and training

We trained gradient boosting machine (GBM) models using LightGBM, which provides an efficient implementation of a gradient boosted ensemble of decision trees.[26] For the comparative L2-regularized logistic regression (LR) model and baseline we used the LogisticRegressionCV class in scikit-learn.[27]

For both model types, we optimized hyperparameters with grid search and five-fold cross-validation (CV). CV folds were defined on patient level to ensure their independence and were stratified by the maximum label per patient. We selected the parameters producing the highest mean sample-AUPRC and retrained on the full training set with these parameters. For GBM models, we used early stopping during CV to determine the number of boosted trees in the ensemble. For each combination of hyperparameters, model training was stopped early when the mean logistic loss over CV folds did not improve for 50 iterations. When retraining on the full training set, we used the number of boosted trees, which produced the lowest logistic loss during CV. The exact parameter choices, grids, and optimal values are provided in Table S6.

To evaluate model performance and variability on retrospective data, we repeatedly split the patients of the retrospective cohort into two-thirds training and one-third test set (stratified by the maximum label per patient). We ran the entire training process, including imputation, normalization, and hyperparameter search, on each training set independently and evaluated model performance on the corresponding test set using AUROC, sample-AUPRC, and event-AUPRC. Here, sample-AUPRC is the standard area under the precision–recall curve, where recall is defined as the fraction of correctly predicted samples (i.e., time points) with positive label (sample recall). In contrast, event-AUPRC defines recall as the fraction of events, which were correctly predicted on at least one of the positive labeled time points (event recall) and was previously introduced for time-dependent event prediction.[17] Unless specified otherwise, model performance on retrospective data is reported as mean and SD over 10 random splits into training and test set. Using the same methodology, we additionally trained a final model on the entire retrospective cohort for prospective validation.
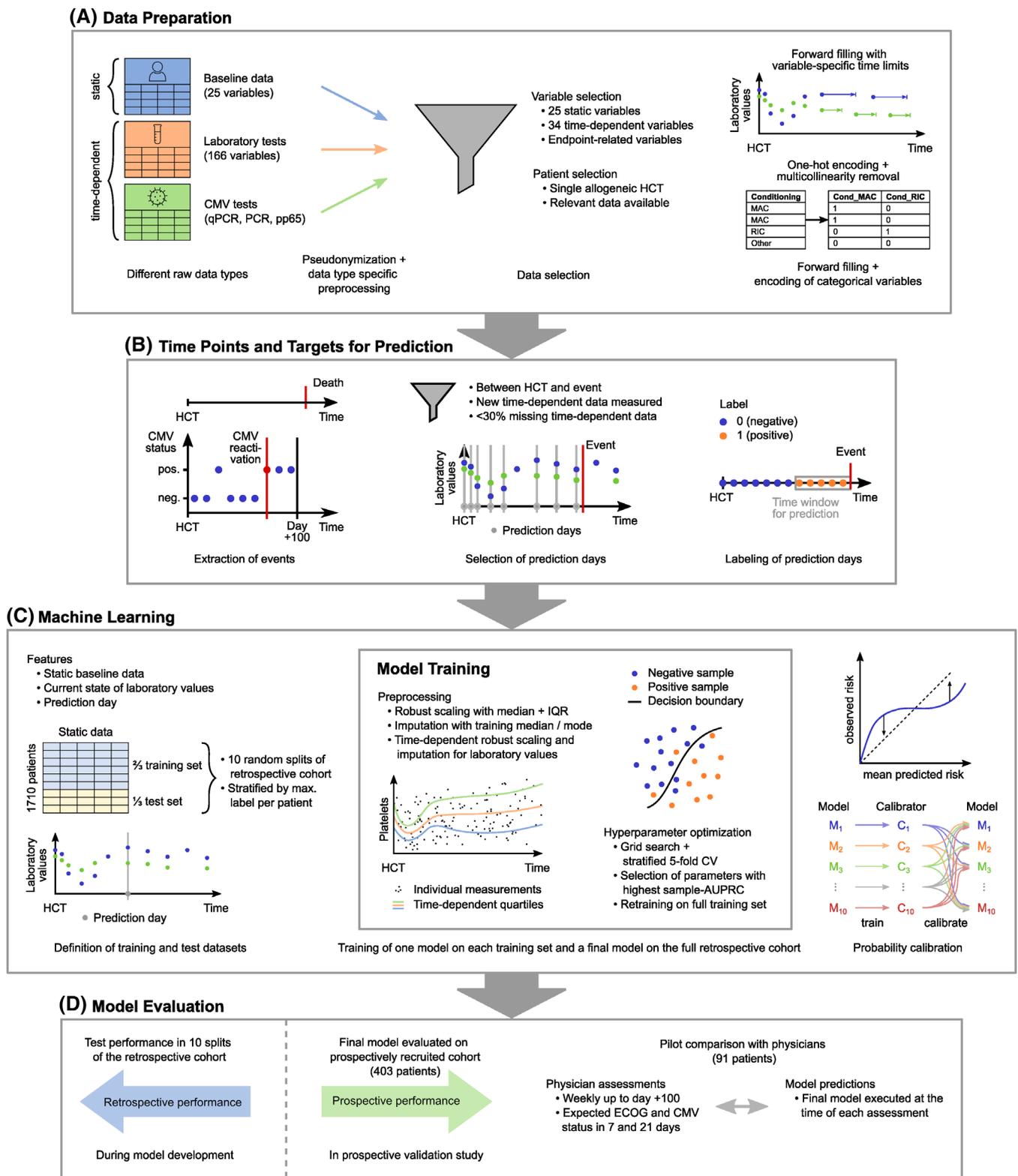
FIGURE 1 Legend on next page.

2.7 | Models with additional features

We evaluated whether additional information from unstructured medical letters or information on the history of laboratory values improved the performance of survival and CMV prediction. For this purpose, we trained two further GBM models per task, which received additional input features (Table S8). Since the added features led to little or no performance improvement on the retrospective data (Figure S15), we

selected only the simpler models with the initial feature set for prospective validation. An overview of all developed models and the included features is provided in Table S9.

2.8 | Model calibration

We calibrated all trained models as a postprocessing step using isotonic regression. For this purpose, we trained a separate calibrator for each split of the retrospective cohort into training and test set, using the raw model predictions on the test set. To apply calibration to any of the models trained for these splits, we averaged the output of the nine calibrators trained on the remaining splits (Figure 1C). The predicted probabilities of the final model trained on the entire retrospective cohort were calibrated using the average over all 10 calibrators.

2.9 | Prospective validation

In order to prospectively validate the developed models on an independent cohort, we recruited 408 patients to the prospective non-interventional XplOit study (inclusion criteria: first allogeneic HCT, ≥18 years, written informed consent) from September 2017 to June 2020. We applied the final GBM and LR models to generate predictions on the prospective cohort, selecting prediction times with the same methodology described for the retrospective cohort. Throughout the prospective study, both physicians and patients were blinded for the model predictions.

We compared model predictions to the observed outcome and measured performance with the same metrics used on retrospective data. To assess variability in performance measures, we applied bootstrapping with 10 000 bootstrap samples on the prospective dataset. During bootstrapping, we kept the total number of positive labeled samples fixed at its original value and adjusted the number of negative labeled samples to obtain the same positive fraction as observed in the retrospective dataset to enable a direct performance comparison between retrospective and prospective cohort.

2.10 | Head-to-head comparison to physicians' expectations

Within the last quarter of prospectively recruited patients, we performed a pilot study to compare the performance of the developed ML models to the expectations of experienced physicians regarding early complications after HCT. For 91 patients in the prospective cohort, we prospectively assessed the expectations of the treating physicians regarding overall survival and CMV reactivation between day 0 and day +100 after HCT. Physicians were requested to estimate each patient's performance status (ECOG, 0–5) and risk to have a CMV reactivation (low, moderate, high) in 7 and 21 days after the assessment date. Assessment was performed weekly between day −7 and day +100 after HCT by physicians of the Department of Hematology and Stem Cell Transplantation at UHE. Whenever an assessment was made (starting at HCT), the GBM and LR models were executed on the most recent available data to allow for a head-to-head comparison of the predictions. Treating and risk assessing physicians were blinded for the model predictions.

To enable model predictions on the day of each assessment, we used indefinite forward filling on laboratory measurements for this analysis. Since the physicians' assessments were recorded as categories rather than probabilities, we binarized their answers and the model predictions, and compared performance measures on these binary predictions. Specifically, we compared Matthews correlation coefficients (MCC) and F1 scores, choosing the optimal binarization threshold for models and physicians, respectively. To assess variability, we repeated this evaluation on 10 000 bootstrap samples drawn from the dataset for this pilot comparison. Here, we kept the positive fraction fixed by drawing the same number of samples with positive and negative labels, respectively, as were originally in the dataset.

2.11 | Implementation

Preprocessing was in part performed within the XplOit platform (version 20201130_1700) using extract–transform–load pipelines specific to each data type. All remaining steps of preprocessing, model

FIGURE 1 Overview of model development and evaluation. (A) Data preparation. Raw data tables were pseudonymized and combined into one coherent dataset. After patient and variable selection, sparsity in laboratory values was reduced by forward filling with variable-specific time limits and categorical features were converted into a binary representation. (B) Time points and targets for prediction. Of the two considered events, death was directly documented and CMV reactivation was extracted from virological tests as the first positive CMV test, which was not an isolated positive. We selected all days between HCT and an event or censoring as prediction days where new laboratory values were measured and <30% of them were missing. Each prediction day was labeled positive if the event occurred in a fixed subsequent time window, and negative otherwise. (C) Machine learning. Models received static baseline data, current laboratory values, and the prediction day after HCT as inputs. We randomly split the retrospective cohort into training and test sets 10 times, and trained a separate model on the training set of each split and a final model on the full retrospective cohort. We defined the splits on patient level and stratified the proportion of patients with at least one positive labeled time point. Preprocessing included a time-dependent normalization and imputation of laboratory values. We trained one calibrator for each split into training and test set. To calibrate each model, we averaged over the calibrators trained on the remaining splits or over all calibrators in case of the final model. (D) Model evaluation. During model development, performance was evaluated on the test set of the 10 splits of the retrospective cohort. In a prospective validation study, we additionally evaluated the performance of the final model on 403 prospectively recruited patients and, in a subset of 91 patients, performed a pilot comparison with experienced HCT physicians.

building, and analysis were implemented in python (version 3.8.2) using scikit-learn (version 0.22.1),[27] numpy (version 1.18.1),[28] and pandas (version 1.0.3).[29] GBM models were trained with LightGBM (version 2.3.1)[26] and SHAP values for these models were computed using the TreeExplainer implemented in shap (version 0.37.0).[30]

3 | RESULTS

Using ML, we developed GBM and LR models to predict at multiple time points after HCT whether an event, that is, death or CMV reactivation, would occur in a subsequent time window of 21 or 7 days (Figure 1A–C). Each model received a combination of routinely collected static and time-dependent HCT data as input and was trained to predict a continuous risk score for one specific event. We then validated these ML models in the prospective non-interventional XplOit study, which also included a pilot comparison between ML model predictions and prospectively collected outcome expectations of experienced HCT physicians (Figure 1D).

3.1 | Assembling an extensive longitudinal HCT dataset

Utilizing the XplOit data integration platform for medical research,[31] we assembled an extensive, well-annotated retrospective dataset incorporating static and time-dependent data of 1710 HCT patients to form the basis of model development. Based on their relevance, we selected 60 parameters as input features for the ML models (Figure 1), including static pre-HCT constellations, such as diagnosis, conditioning regimen, and donor information, as well as the day of the prediction and current laboratory values (Table S3). During the non-interventional XplOit validation study, we additionally recruited 403 patients for prospective model validation.

Relevant baseline characteristics were balanced between the development and validation cohort and are detailed in Table S1. As expected, the largest fraction of patients presented with acute myeloid leukemia for HCT. Cyclosporin A (CSA) was the predominant calcineurin inhibitor for baseline immunosuppression. Following changes in HCT practices, such as the introduction of post-transplant cyclophosphamide, the prospective cohort had a higher proportion of patients with tacrolimus-based immunosuppression. Time-dependent laboratory values were available at 163 425 and 31 889 time points in the retrospective and prospective cohort, respectively, comprising more than 5.4 million individual measurements in total. In accordance with international best-practice HCT guidelines, the measurement intervals were shortest during the inpatient care of 35–40 days and were extended for outpatients (Figure S5).

The endpoints of this study were adequately covered by the analyzed data. The time of death was known for 1134 patients (53.7%), and 925 patients (43.8%) developed an early CMV reactivation (within 100 days after HCT), with the median first episode of CMV

reactivation at day +34. After 24 months, the overall survival (OS) rate was 55% in the retrospective cohort (Figure S2a), which is representative of HCT outcomes across different risk groups in real-world data. After a median follow-up of 14.4 months, the median overall survival was not reached in the prospective XplOit study (Figure S2b). While the cumulative incidence of NRM was comparable between the retrospective and prospective cohort, overall survival differed significantly consistent with reduced relapse rates in recent HCT (Figure S2c). The GBM model predicts 21-day mortality with an AUROC of 0.92 and an event-AUPRC of 0.58.

We evaluated model performance using the standard area under the receiver-operating characteristic (AUROC) and two versions of the area under the precision–recall curve (AUPRC), event-AUPRC and sample-AUPRC. While sample-AUPRC is based on the standard recall on individual samples, event-AUPRC defines recall as the fraction of correctly predicted events and specifically addresses time-dependent event prediction.[17] Following data preprocessing, as detailed in the Methods section, the retrospective dataset for the development of 21-day mortality models contained 143 669 time points of 1695 patients, 7354 of these time points (5.14%) were labeled positive (death occurred within 21 days).

The developed GBM model for 21-day mortality prediction achieved a very high AUROC of 0.918 and good event AUPRC of 0.584 (Figure 2A,B). It outperformed the LR model, which had an AUROC of 0.900 and an event-AUPRC of 0.524. To assess the value of including time-dependent data for outcome prediction, we compared these models with a baseline LR model receiving only static input data. The time-dependent GBM and LR models both vastly outperformed the static LR baseline, which achieved an AUROC of only 0.594 and event-AUPRC of 0.085. The same trend was observed in sample-AUPRC (Figure S6). After calibration, we obtained very close agreement between predicted and observed risk, with areas of 0.04 and 0.06 between the line representing ideal calibration and the calibration curve of the GBM and the LR model, respectively (Figure S7).

We then analyzed the performance of the GBM model for 21-day mortality prediction over time in more detail. As expected, the fraction of correctly predicted events increased with shorter time to the event (Figure S3a). This finding was independent of the exact threshold chosen to convert continuous risks into binary event predictions. With a threshold chosen to obtain an overall event recall of 0.8, the majority of events was predicted at least 2 weeks in advance. The predicted continuous risks evolved similarly with a steady increase as patients approached an event (Figure 2C), which supports the plausibility of the model. Compared with the average risk predicted for negatively labeled time points, that is, without any event in the subsequent 21 days, this increase was detectable as early as 85 days before the event. Although the GBM model recognized initial signs of an impending event much earlier than 21 days before, these were not yet sufficient for a confident event prediction. Analyzing GBM model performance as a function of the prediction day after HCT, we found that AUROC increased slightly over time (Figure 2D). Sample-AUPRC varied more noticeably; it was lower early after HCT and highest
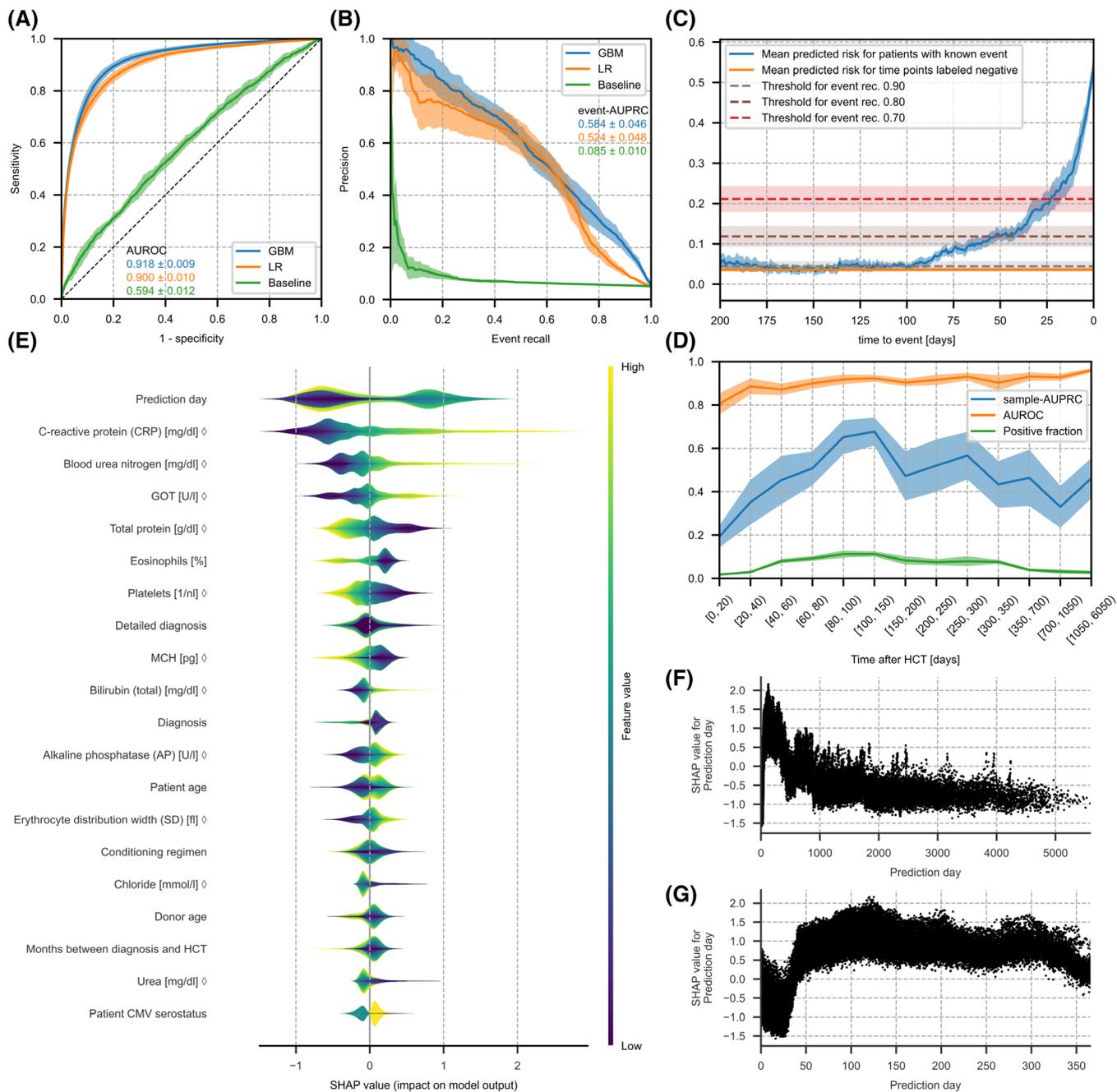
FIGURE 2 Performance and feature importance of the GBM model for 21-day mortality prediction. (A) Receiver-operating characteristic of GBM and LR model, which received a combination of static and time-dependent input features, and a baseline model which received only static features. (B) Precision–recall curve for the same models as shown in (A), based on event recall, that is, the fraction of events that were correctly predicted on any of the previous 21 days. (C) Mean predicted risk of the GBM model as a function of time to event. For reference, the orange horizontal line indicates the mean predicted risk over all time points labeled negative. Dashed horizontal lines indicate the thresholds to reach the target event recall stated in the figure legend. (D) AUROC and sample-AUPRC of the GBM model and fraction of samples with positive label as functions of time after HCT. Bin size increases because fewer samples were available late after HCT. (A–D) Lines and shaded areas show the mean ± SD on the test set over 10 random splits into training and test data. (E) Layered violin plot of SHAP values of the GBM model for the 20 features with highest mean absolute SHAP value. The thickness of the violins corresponds to the estimated density of each feature's SHAP values, colors show the magnitude of feature values (percentiles). For features marked with ◇, the feature value is the time-normalized score that the model received as input, not the raw value in its original unit. For categorical features, the colors are based on an integer representation and should not be interpreted as ordered. All SHAP values were computed based on raw model output in log-odds space. (F–G) Scatter plots of individual SHAP values over feature values. Shown are plots for the feature prediction day after HCT on the entire range of feature values (F) and zoomed in on the first year after HCT (G).

between days 80 and 150. This correlated with the fraction of positive labeled samples at different times after HCT since a small positive fraction makes it difficult to achieve a high precision score.

3.2 | Prediction day, CRP, and urea nitrogen had the highest impact on mortality predictions

Using SHapley Additive exPlanations (SHAP values),[30] we analyzed the impact of individual features on GBM model predictions (Figure 2E). SHAP values indicate how much the value of a feature has contributed to the prediction generated for a specific sample. High values (>0) indicate that the feature value increased the predicted risk, while low values (<0) indicate that it reduced the predicted risk. For the GBM model predicting 21-day mortality, the most important features were the day of the prediction (in days after HCT), C-reactive protein (CRP), blood urea nitrogen, glutamate oxaloacetate transaminase (GOT), and protein levels (Figure 2E). Especially high blood levels of CRP, urea nitrogen, and GOT as compared with other patients at the same time after HCT led the model to predict an increased mortality risk. In contrast, high values of total protein led to a lower predicted risk. These features are markers of inflammation or infection, or reflect liver or kidney function. For the prediction day after HCT, the relationship between feature value and SHAP value was more complex. Within the first year after HCT, the prediction day appeared to increase the predicted risk, while after 1 year the SHAP values continuously decreased, falling below zero about 3 years after HCT (Figure 2F). A closer inspection of the first year after HCT revealed that prediction days up to day +40 decreased the predicted risk, while all later prediction days of the first year had constantly high SHAP values (Figure 2G).

For 7-day mortality prediction, the GBM and LR models both had a higher AUROC and lower event- and sample AUPRC than the corresponding 21-day models (Figure S8). As a consequence of the narrower time window, fewer samples were labeled positive (1.88% for 7-day prediction), which can partially explain the lower event and sample-AUPRC. Detailed results for the 7-day prediction models are provided in the supplementary material (Figures S8 and S9). While these models focused on all-cause mortality to enable prediction for all HCT patients, independently of their relapse status, we also tested if our modeling approach would result in comparable prediction performance for NRM, which was indeed confirmed (Table S10b).

3.3 | The performance of 21-day mortality prediction models remained high on prospective data

In the second step, we validated the developed ML models on an independent prospectively recruited cohort ($n = 403$) from the same HCT center (Table 1A). Depending on the time window for prediction, we observed specific differences in the performance of mortality prediction on prospective data. The models for 21-day mortality

prediction remained relatively stable; AUROC and event-AUPRC of the GBM model faded only slightly from 0.918 to 0.895 and from 0.584 to 0.522, respectively. Responding to changes in HCT practices, we additionally compared subgroups of the two main distinct immunosuppressive regimens (CSA and TAC) within the prospective cohort (Table S2), and found no major differences between these subgroups. However, for 7-day prediction, we observed a quite pronounced decrease in model performance on prospective data, with AUROC and event-AUPRC of the GBM model dropping from 0.951 to 0.931 and from 0.525 to 0.372, respectively. Here, model performance was noticeably higher for patients with CSA instead of TAC immunosuppression, which were better represented in the retrospective cohort. Model calibration remained appropriate on prospective data (Figure S10).

Despite some differences between retrospective and prospective patient outcomes and model performance, the AUROC of both GBM and LR models remained high on prospective data. Event- and sample-AUPRC were also acceptable given the low fraction of positive labeled time points. Next, we tested if the models trained to predict all-cause mortality could also be leveraged to predict NRM. The validation of the GBM model for 21-day mortality on the subgroup of 361 prospectively recruited patients without relapse resulted in a comparably high AUROC of 0.900, an event-AUPRC of 0.536, and a sample-AUPRC of 0.428 (Table S10a). Thus, the developed ML models were successfully validated on the prospective dataset for both all-cause mortality and NRM.

3.4 | For 21-day mortality prediction the GBM models performed similar to HCT physicians

In a pilot study, which was part of the prospective validation, we additionally compared the predictive performance of the final GBM and LR models during the first 100 days after HCT to the outcome expectations of experienced HCT physicians. Within the last year of the prospective XplOit study, each treating physician was requested once per week to estimate their patients' expected Eastern Cooperative Oncology Group (ECOG) performance score and risk of CMV reactivation (low, medium, high) in 7 and 21 days. In total, we collected 649 forms containing post-HCT assessments for 91 patients. In parallel, we executed GBM and LR models at the time of each assessment with the latest available time-dependent data. All physicians were blinded to the model predictions.

The results of this comparison are displayed in Table 1B. For 21-day mortality prediction, GBM model and physicians showed a similar performance, as measured by MCC values of 0.461 ± 0.086 and 0.488 ± 0.089, respectively. Although the differences were small compared with the SD derived from bootstrapping, trends showed a slight advantage of the physicians' expectations over the GBM model predictions and of the GBM model over the LR model. For 7-day prediction, the physicians achieved a very high MCC and F1 score of 0.796 ± 0.180 and 0.767 ± 0.214, respectively,

TABLE 1 Model performance on prospective data and comparison of the prediction performance of ML models and treating physicians

A. Comparison of model performance on retrospective and prospective cohort[a]

Prediction task	Model	Performance metric	Retrospective cohort	Prospective cohort
Mortality 21 days	GBM	AUROC	0.918 ± 0.009	0.895 ± 0.005
		Event-AUPRC	0.584 ± 0.046	0.522 ± 0.023
		Sample-AUPRC	0.488 ± 0.042	0.414 ± 0.015
	LR	AUROC	0.900 ± 0.010	0.866 ± 0.006
		Event-AUPRC	0.524 ± 0.048	0.549 ± 0.021
		Sample-AUPRC	0.445 ± 0.043	0.413 ± 0.015
Mortality 7 days	GBM	AUROC	0.951 ± 0.006	0.931 ± 0.006
		Event-AUPRC	0.525 ± 0.038	0.372 ± 0.029
		Sample-AUPRC	0.410 ± 0.034	0.303 ± 0.021
	LR	AUROC	0.940 ± 0.008	0.894 ± 0.009
		Event-AUPRC	0.464 ± 0.038	0.348 ± 0.026
		Sample-AUPRC	0.375 ± 0.023	0.269 ± 0.020
CMV 21 days	GBM	AUROC	0.825 ± 0.006	0.846 ± 0.004
		Event-AUPRC	0.620 ± 0.040	0.574 ± 0.011
		Sample-AUPRC	0.565 ± 0.025	0.549 ± 0.009
	LR	AUROC	0.793 ± 0.013	0.818 ± 0.004
		Event-AUPRC	0.532 ± 0.050	0.515 ± 0.012
		Sample-AUPRC	0.502 ± 0.033	0.496 ± 0.009
CMV 7 days	GBM	AUROC	0.846 ± 0.010	0.875 ± 0.005
		Event-AUPRC	0.335 ± 0.023	0.323 ± 0.015
		Sample-AUPRC	0.295 ± 0.017	0.302 ± 0.012
	LR	AUROC	0.777 ± 0.014	0.802 ± 0.006
		Event-AUPRC	0.192 ± 0.017	0.176 ± 0.007
		Sample-AUPRC	0.188 ± 0.014	0.181 ± 0.006

B. Comparison of the prediction performance of ML models and treating physicians[b]

Prediction task	Performance metric	Physicians	GBM	LR
Mortality 21 days	MCC	0.488 ± 0.089	0.461 ± 0.086	0.417 ± 0.087
	F1 score	0.453 ± 0.086	0.427 ± 0.085	0.360 ± 0.084
Mortality 7 days	MCC	0.796 ± 0.180	0.377 ± 0.064	0.304 ± 0.069
	F1 score	0.767 ± 0.214	0.272 ± 0.077	0.204 ± 0.069
CMV 21 days	MCC	0.234 ± 0.051	0.329 ± 0.062	0.266 ± 0.023
	F1 score	0.289 ± 0.055	0.322 ± 0.049	0.281 ± 0.026
CMV 7 days	MCC	0.170 ± 0.067	0.147 ± 0.033	0.143 ± 0.042
	F1 score	0.168 ± 0.063	0.110 ± 0.025	0.117 ± 0.030

[a]For the retrospective cohort, the table displays mean ± SD on the test set over 10 random splits into training and test data. For the prospective cohort, it shows the performance of the final models, trained on the entire retrospective cohort, as mean ± SD over 10 000 bootstrap samples.
[b]Performance of models and physicians was measured using Matthews correlation coefficient (MCC) and F1 score after binarization with the respective optimal threshold. Displayed is the mean ± SD over 10 000 bootstrap samples.

outperforming both ML models. However, the dataset for comparing predictive performance over a 7-day window in this pilot sub-study was limited due to a low number of fatalities preceded by prospective assessments. In addition, these deceased patients were less representative of the training cohort since they received TAC immunosuppression.

3.5 | The GBM models for 21-day CMV prediction had AUROC 0.83 and event-AUPRC 0.62

For two reasons, the dataset for the development of models predicting early CMV reactivation was smaller than for mortality prediction: first, we focused on the first 100 days after HCT, where the

earliest episode of CMV reactivation almost exclusively occurs in the absence of prophylaxis. Second, we excluded patients without CMV testing during the first 30 days after HCT since the earliest CMV episode could have been missed without regular tests. For CMV prediction over 21 days, the dataset contained 52 008 time points from 1561 patients, of which 12 413 (23.87%) were labeled positive.

Here, the GBM model also had the best performance with an AUROC of 0.825 compared with 0.793 and 0.779 for LR and baseline, respectively (Figure 3A). The same trend was observed in event-AUPRC (Figure 3B), which was 0.620, 0.532, and 0.473 for GBM, LR, and baseline models, respectively, and in sample-AUPRC (Figure S6). For CMV prediction, the gap between models using time-dependent data (GBM and LR) and the static baseline was much smaller than for mortality prediction. The primary reason is that even the CMV models with access to time-dependent data relied on static features for their predictions, while time-dependent laboratory values had only a minor impact (Figure 3E). Calibrated predictions agreed closely with the observed risk; GBM and LR models both had an area of 0.05 between the calibration curve and the line representing perfect calibration (Figure S11).

We performed the same analysis of GBM model performance over time for 21-day CMV prediction as described for 21-day mortality prediction. Again, the fraction of correctly predicted events increased while approaching the event, and this trend was independent of the exact decision threshold chosen (Figure S3b). With a threshold offering an event recall of 0.8, the GBM model predicted 60% of events at least 2 weeks before they occurred. For patients approaching a CMV event, the mean predicted risk rose almost linearly, starting about 40 days beforehand (Figure 3C). While AUROC remained nearly constant over time after HCT, sample-AUPRC dropped after day +40 post-HCT as fewer events occurred (Figure 3D).

3.6 | The CMV predictions were mainly based on prediction day and static features

SHAP value analysis of the GBM model for 21-day CMV prediction revealed that patient CMV serostatus had the highest impact on model predictions, followed by prediction day after HCT and underlying hematologic disorder (Figure 3E). Conditioning regimen, anti-thymocyte globulin as GVHD prophylaxis, donor CMV serostatus, and patient age were also relevant. Interestingly, the time-dependent laboratory values had only a minor role in the predictions of this CMV model, with the exception of the percentage of lymphocytes, which ranked among the top 10 features. Consequently, the CMV model relied predominantly on static data. The joint analysis of feature values and SHAP values confirmed that a positive patient CMV serostatus led to a strongly increased risk prediction, while a negative serostatus reduced the predicted risk (Figure 3F). This dichotomy was even more pronounced among patients who received additional T cell depletion with anti-thymocyte globulin as GVHD prophylaxis. The

SHAP values for the prediction day after HCT peaked between days +20 and +50, indicating a typical timing for early CMV reactivation events (Figure 3G). This peak was most pronounced for patients with recipient-positive CMV serostatus. Interestingly, donor age did not have a differential impact on the risk of CMV reactivation predicted by the GBM model, except for very young donors (<17 years) (Figure S4b). However, these samples were limited in our dataset and were also associated with young patient age.

For prediction of CMV reactivation over 7 days, the GBM and LR models both had a similar AUROC but considerably lower event- and sample-AUPRC than the corresponding models for prediction over 21 days. Again, this may be influenced by the lower positive fraction of 7.50% with the narrower 7-day time window. An analysis of model performance over time and of the impact of individual features on predictions of the 7-day GBM are included in the supplementary material (Figures S12 and S13).

3.7 | CMV models were successfully validated and performed similar to HCT physicians

In the prospective validation cohort ($n = 398$), the performance of all CMV models remained very close to their performance on retrospective data (Table 1A). Compared with the retrospective cohort, the AUROC of the GBM model for 21-day CMV prediction increased slightly from 0.825 to 0.846, while its event-AUPRC decreased slightly from 0.620 to 0.574. This performance remained stable across patient subgroups with distinct immunosuppressive regimens (Table S2). In contrast, the 21-day LR model had a higher performance for patients who received CSA instead of TAC immunosuppression. For prediction over 7 days, both models demonstrated very similar performance on retrospective and prospective data, and a trend toward higher performance for patients with CSA immunosuppression. All CMV models remained well calibrated on prospective data, concluding the successful prospective validation (Figure S10).

In a pilot study, we compared the predictive performance of the ML models to the risk of CMV reactivation estimated by experienced HCT physicians. The results are shown in Table 3B. For 21-day prediction, the GBM model had the best performance, with an MCC of 0.329 ± 0.062 compared with 0.266 ± 0.023 and 0.234 ± 0.051 for LR model and physicians, respectively. On the other hand, the physicians had a small lead over both ML models for prediction over 7 days. In both cases, these differences in average performance were not decisive, given the limited dataset for this comparison.

4 | DISCUSSION

In response to persisting difficulties to predict relevant complications in HCT patients and to support clinical assessment, we developed and validated the first ML models for time-dependent prediction of mortality and CMV reactivation after HCT. These ML models accurately predict patient-specific event risks within a specified time window
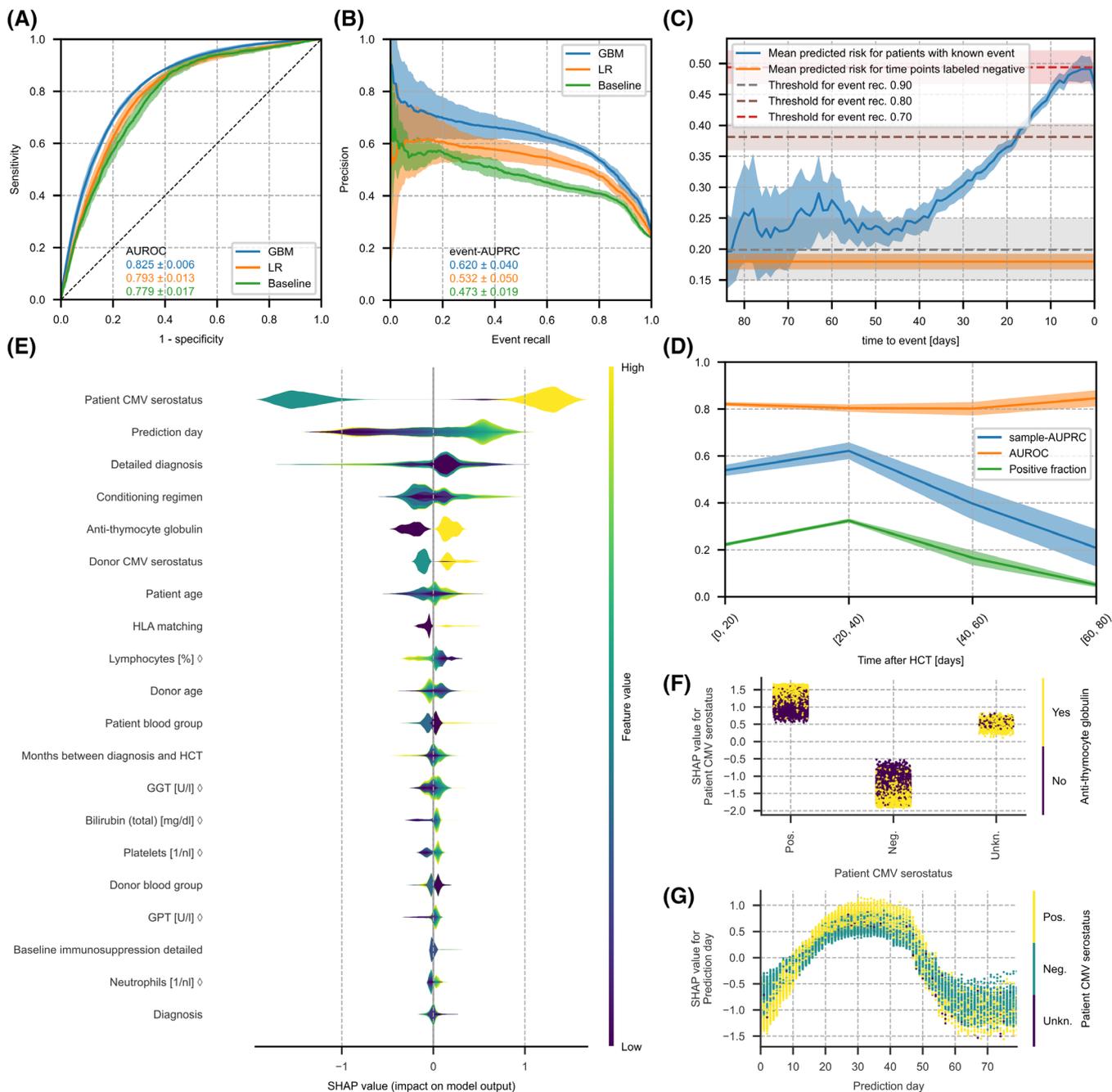
FIGURE 3 Performance and feature importance of the GBM model for 21-day prediction of CMV reactivation. (A) Receiver-operating characteristic of GBM and LR model, which received a combination of static and time-dependent input features, and a baseline model which received only static features. (B) Precision–recall curve for the same models shown in (A) based on event recall, i.e. the fraction of events that were correctly predicted on any of the previous 21 days. (C) Mean predicted risk of the GBM model as a function of time to event. For reference, the orange horizontal line indicates the mean predicted risk over all time points labeled negative. Dashed horizontal lines as in Figure 2. (D) AUROC and sample-AUPRC of the GBM model and fraction of samples with positive label as functions of time after HCT. (A–D) Lines and shaded areas show the mean ± SD on the test set over 10 random splits into training and test data. (E) Layered violin plot of SHAP values of the GBM model for the 20 features with highest mean absolute SHAP value. The thickness of the violins corresponds to the estimated density of each feature's SHAP values, colors show the magnitude of feature values (percentiles). For features marked with ◇, the feature value is the time-normalized score that the model received as input, not the raw value in its original unit. For categorical features, the colors are based on an integer representation and should not be interpreted as ordered. All SHAP values were computed based on raw model output in log-odds space. (F–G) Scatter plots of SHAP values over feature values. Samples are colored by the value of a second feature to reveal interactions, which show as vertical color patterns. Displayed are plots for the feature patient CMV serostatus colored by anti-thymocyte globulin (F) and prediction day after HCT colored by patient CMV serostatus (G).

and at multiple time points after HCT and pave the way toward clinical decision support systems for transplantation medicine. While existing predictive models[18–20] and scores[8,9,32] for HCT-specific risk assessment predominantly focus on pre-HCT assessment to support treatment and donor selection, time-dependent risk assessment may enable physicians to refine and individually adjust treatments and preventive measures after HCT to obtain the best possible outcome for each patient.

Our ML models combine static patient information as used in previous HCT ML models[18] with longitudinal laboratory data and update their predictions whenever new time-dependent data become available. Although this study builds on previous research on ICU data,[17] our ML models prove the applicability of this new approach in the field of HCT and on a much larger time scale with varying data granularity, which underlines the relevance of this study beyond the field of transplantation.

Recent ML models in patients with leukemia combined static patient data at diagnosis with time series of laboratory measurements to predict patient outcome at a single point in time.[33] While these models included HCT as an input parameter, they neither predicted the outcome after HCT nor at multiple time points. Another ML study using longitudinal HCT data integrated patients' vital signs and predicted graft-versus-host disease by day +100 with a modest AUROC of 0.66,[34] allowing for a single prediction on day +10 after HCT. Personalized ML survival models for HCT patients refined prognosis at the time of HCT but exclusively relied on static pre-HCT data as input parameters without adapting to complications occurring after HCT.[35] Most recently, the integration of multiple time-dependent variables into an ML model improved the prediction of acute GVHD (AUROC 0.78) in HCT recipients.[36]

Although our final models update their predictions whenever new data become available, they use only the most recent laboratory result for each prediction. On large EHR databases, recurrent deep neural networks, for example, using long short-term memory (LSTM) units, have demonstrated high prediction performances utilizing entire time series as model input.[16,37,38] A limitation of LSTMs is, however, their dependence on very large training data, which are not available in all medical domains. For instance, LSTMs did not outperform GBM models for the time-dependent prediction of circulatory failure based on a large single-center ICU dataset.[17] Since additional features describing the history of laboratory values did not improve the performance of our GBM models (Figure S15), we did not pursue more complex approaches for time series data.

In this article, we considered multiple endpoints and time windows for prediction. Across these tasks, GBM models consistently outperformed LR and provided well-calibrated time-dependent risk predictions. Prediction performance was best for prediction of 21-day mortality, where we obtained very high AUROC and high event-AUPRC. High predictive performance, in addition to validity and independent replication, is a core requirement for the clinical use of predictive models in decision support systems[39] since it is the first indicator of health impact and effectiveness. Yet, identifying the optimal performance threshold for effectiveness and impact is also subject

to medical,[40] technical, and ethical[41] considerations relating to the predicted outcome, potential consequences of false predictions, and implementation issues. Our pilot comparison to physicians' expectations indicates that the developed models will likely provide relevant practical use, for example, as a risk screening tool for post-HCT outpatients. Given the possibilities of intervening via anti-infective or immunosuppressive drugs and hospitalization, such warning systems might prevent fatal outcomes. The immediate availability of the features used by our models in most HCT centers, including both the static HCT parameters and the continuously measured standardized laboratory variables, is a major advantage for its clinical application for decision support. Finally, successful implementation in clinical practice can also be influenced by physicians' trust in ML models, which may be increased by providing understandable explanations for individual predictions,[40] for example, via SHAP values.

Since a direct comparison to existing scores designed for pre-HCT risk assessment is not possible, we compared our models to a baseline model, which was trained for the time-dependent prediction task but used only static input features. Interestingly, time-dependent input features proved highly valuable for mortality prediction but only offered modest improvement for CMV prediction, indicating that time-dependent outcome prediction may improve HCT-specific risk assessment beyond current standards, but possibly not for all endpoints in equal measure.

The final ML models were successfully validated on an independent, prospectively recruited cohort, as shown by the overall high predictive performance of the developed models on prospective data. For mortality prediction, model performance decreased slightly compared with the retrospective cohort, which was in part explained by changes in immunosuppression strategies. However, the slight performance drop also in patients with identical baseline immunosuppression indicates a dataset shift over time. This is well in line with a recent EBMT analysis of HCT data up to the year 2016, showing decreasing NRM over time.[3] Given the small differences in prediction performance between the retrospective and prospective cohort, the applicability of the mortality prediction models remains unaffected. The importance of prospective validation has been previously shown[42] and is also reflected in our study design. Indeed, predictive models developed for use in clinical practice require continuous monitoring and, if necessary, refinement. Possibly due to the large impact of static features, the performance of models predicting CMV reactivation was not affected by this dataset shift and remained stable.

Our exploratory head-to-head comparison with experienced HCT physicians revealed that GBM models performed approximately on par for 21-day prediction of mortality and CMV reactivation. Despite the limitations of this pilot comparison, trends showed that the physicians performed slightly better in mortality prediction while the GBM model was better in predicting CMV reactivation. Since the physicians had direct contact with their patients, and therefore access to more information than the 60 input features of the ML models, these results underline the promising potential for future use of such GBM models in clinical practice. Integrating additional features, such as vital signs or current medication, could potentially increase model

performance further. However, the current feature set used by our final models is readily available in most HCT centers, which is a prerequisite for the implementation as a clinical decision support system.

Although this is a topic of active discussion in the scientific community,[43] better interpretability or explainability of ML models in healthcare may improve trust into model predictions[44] and even the quality of decision support systems.[45] Here, SHAP values provide insight into the impact of specific features on model predictions and offer a comprehensive approach to explore underlying biological mechanisms. In the GBM models for mortality prediction, mainly features related to organ function and inflammation (CRP, urea nitrogen, GOT, protein) affected the predicted risk. In contrast, the GBM models predicting CMV reactivation strongly relied on static patient data (CMV serostatus, diagnosis, conditioning regimen). For both endpoints, the prediction day after HCT had a large impact on the predicted risks indicating a typical time period for potential complications after HCT, which is in line with previous reports.[1] While SHAP values can provide valuable insight into the features contributing to individual model predictions, it is important to note that they do not represent causal relationships.

The time-dependent prediction problems we considered were imbalanced, meaning that our data contained few samples with a positive label. In this situation, AUPRC is a more informative performance measure than AUROC.[17] However, the exact positive fraction in our data varied across prediction tasks, and we observed that event- and sample-AUPRC were strongly correlated with it. This made it difficult to compare models for different endpoints and time windows directly. Sampling methods could be used to adjust the positive fraction for such comparisons, but then performances would no longer be measured on the data distribution of a realistic application scenario, where the positive fraction is determined by the prevalence of events. By design, the positive fraction for 21-day prediction tasks was higher than for 7-day prediction. Quite unexpectedly, this made 21-day prediction the easier task for ML methods, leading to more robust results even though the distance from positive labeled prediction days to the event was longer. In addition, the 21-day prediction models have a greater potential clinical applicability because they may enable an earlier intervention to prevent or treat complications.

This study has limitations and strengths. It included only data from a single center, which may limit the general applicability of the developed models. However, the models were built on a homogeneous and large dataset of several million data points, and the patient characteristics and HCT practice standards reflected those of major international centers. The precise predictions of our models using standard laboratory features available in all HCT centers pave the way toward the implementation of decision support systems in HCT. Ultimately, its routine use as a medical device requires a prospective clinical trial for safety and efficacy, according to, for example, the EU medical device regulation (EU 2017/745). As in many previous studies,[46,47] we defined CMV reactivation events only based on detectability, combining data of different quantitative and qualitative CMV tests. However, more recent studies have demonstrated that the severity of CMV disease may be revealed by viral load

kinetics.[48,49] It would be interesting for future work to attempt time-dependent prediction of CMV reactivation with a narrower event definition based on a threshold for the viral load.

The developed ML models predict mortality and CMV reactivation for HCT patients reliably and in a time-dependent manner, and therefore may potentially improve patient outcomes once implemented as decision support systems in post-HCT care.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

CONFLICT OF INTEREST

The authors declare the following competing interests: Amin T. Turki received consulting fees from CSL Behring, MSD, JAZZ Pharmaceuticals, and MaaT Pharma; travel subsidies from Neovii Biotech. Dietrich W. Beelen received travel subsidies from Medac. The other authors declare no competing financial interests.

DATA AVAILABILITY STATEMENT

The data used in this article contains sensitive personal health information. Due to the high dimensionality and the inclusion of longitudinal data, it cannot be fully anonymized and published without the risk of re-identification. Requests for access to the data may be submitted to the University Hospital Essen and are subject to approval by data protection officer and ethics committee. Source code may be obtained from the corresponding authors upon request. To enable independent replication of our methods, we included detailed

descriptions of preprocessing and model development in the Methods section and in the supplementary material.

ORCID

Lisa Eisenberg https://orcid.org/0000-0002-1041-7948
Norbert Graf https://orcid.org/0000-0002-2248-323X
Dietrich W. Beelen https://orcid.org/0000-0001-5050-220X
Nico Pfeifer https://orcid.org/0000-0002-4647-8566
Amin T. Turki https://orcid.org/0000-0003-1347-3360

REFERENCES

1. Copelan EA. Hematopoietic stem-cell transplantation. *N Engl J Med*. 2006;354:1813-1826.
2. Gooley TA, Chien JW, Pergam SA, et al. Reduced mortality after allogeneic hematopoietic-cell transplantation. *N Engl J Med*. 2010;363: 2091-2101.
3. Penack O, Peczynski C, Mohty M, et al. How much has allogeneic stem cell transplant-related mortality improved since the 1980s? A retrospective analysis from the EBMT. *Blood Adv*. 2020;4:6283-6290.
4. Kröger N, Solano C, Wolschke C, et al. Antilymphocyte globulin for prevention of chronic graft-versus-host disease. *N Engl J Med*. 2016; 374:43-53.
5. Marty FM, Ljungman P, Chemaly RF, et al. Letermovir prophylaxis for cytomegalovirus in hematopoietic-cell transplantation. *N Engl J Med*. 2017;377:2433-2444.
6. Gratwohl A, Mohty M, Apperley J. The EBMT: history, present, and future. In: Carreras E, Dufour C, Mohty M, Kröger N, eds. *The EBMT Handbook*. Springer; 2019:11-17.
7. Phelan R, Arora M, Chen, M. *Current Use and Outcome of Hematopoietic Stem Cell Transplantation: CIBMTR US Summary Slides*. Center for International Blood and Marrow Transplant Research, Milwaukee, WI, USA; (2020).
8. Sorror ML, Sandmaier BM, Storer BE, 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;25:4246-4254.
9. 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.
10. 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.
11. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542: 115-118.
12. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, et al. Advanced machine learning in action: Identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *npj Digit Med*. 2018;1:9.
13. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. 2019;25:65-69.
14. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577:89-94.
15. Gao Y, Cai GY, Fang W, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat Commun*. 2020;11:5033.
16. Rank N, Pfahringer B, Kempfert J, et al. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. *npj Digit Med*. 2020;3:139.
17. Hyland SL, Faltys M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med*. 2020;26:364-373.

18. 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;33:3144-3152.
19. Arai Y, Kondo T, Fuse K, et al. Using a machine learning algorithm to predict acute graft-versus-host disease following allogeneic transplantation. *Blood Adv*. 2019;3:3626-3634.
20. Logan BR, Maiers MJ, Sparapani RA, et al. Optimal donor selection for hematopoietic cell transplantation using Bayesian machine learning. *JCO Clin Cancer Inform*. 2021;5:494-507.
21. Luft T, Benner A, Terzer T, et al. EASIX and mortality after allogeneic stem cell transplantation. *Bone Marrow Transplant*. 2020;55:553-561.
22. Luft T, Benner A, Jodele S, et al. EASIX in patients with acute graft-versus-host disease: a retrospective cohort analysis. *Lancet Haematol*. 2017;4:e414-e423.
23. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035.
24. Semler SC, Wissing F, Heyder R. German medical informatics initiative. *Methods Inf Med*. 2018;57:e50-e56.
25. Kolitsi Z, Dipak K, Petra W, et al. *DigitalHealthEurope Recommendations on the European Health Data Space: Supporting Responsible Health Data Sharing and Use through Governance, Policy and Practice*. DigitalHealthEurope, 2021.
26. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30:3146-3154.
27. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
28. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with numpy. *Nature*. 2020;585:357-362.
29. McKinney, W. Data structures for statistical computing in python. In Proc. 9th Python Sci. Conf., 56–61 (2010).
30. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2:56-67.
31. Weiler G, Schwarz U, Rauch J, et al. XplOit: an ontology-based data integration platform supporting the development of predictive models for personalized medicine. *Stud Health Technol Inform*. 2018; 247:21-25.
32. Beauvais D, Drumez E, Blaise D, et al. Scoring system for clinically significant CMV infection in seropositive recipients following allogenic hematopoietic cell transplant: an SFGM-TC study. *Bone Marrow Transplant*. 2021;56:1305-1315.
33. Lu C-C, Li J-L, Wang Y-F, Ko B-S, Tang J-L, Lee C-C. A BLSTM with attention network for predicting acute myeloid leukemia patient's prognosis using comprehensive clinical parameters. In Proc. 2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., 2455–2458 (2019).
34. Tang S, Chappell GT, Mazzoli A, Tewari M, Choi SW, Wiens J. Predicting acute graft-versus-host disease using machine learning and longitudinal vital sign data from electronic health records. *JCO Clin Cancer Inform*. 2020;4:128-135.
35. Okamura H, Nakamae M, Koh S, et al. Interactive web application for plotting personalized prognosis prediction curves in allogeneic hematopoietic cell transplantation using machine learning. *Transplantation*. 2021;105:1090-1096.
36. Liu X, Cao Y, Guo Y, et al. Dynamic forecasting of severe acute graft-versus-host disease after transplantation. *Nat Comput Sci*. 2022;2: 153-159.
37. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *npj Digit Med*. 2018;1:18.
38. Ayala Solares JR, Diletta Raimondi FE, Zhu Y, et al. Deep learning for electronic health records: a comparative review of multiple deep neural architectures. *J Biomed Inform*. 2020;101:103337.

39. Tcheng JE, ed. *Optimizing Strategies for Clinical Decision Support: Summary of a Meeting Series*. The learning health system series. National Academy of Medicine; 2017.

40. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digit Med*. 2020;3:17.

41. McLennan S, Fiske A, Celi LA, et al. An embedded ethics approach for AI development. *Nat Mach Intell*. 2020;2:488-490.

42. Buturovic L, Shelton J, Spellman SR, et al. Evaluation of a machine learning-based prognostic model for unrelated hematopoietic cell transplantation donor selection. *Biol Blood Marrow Transplant*. 2018; 24:1299-1306.

43. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. *Science*. 2021;373:284-286.

44. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113:103655.

45. Bruckert S, Finzel B, Schmid U. The next generation of medical decision support: a roadmap toward transparent expert companions. *Front Artif Intell*. 2020;3:507973.

46. Teira P, Battiwalla M, Ramanathan M, et al. Early cytomegalovirus reactivation remains associated with increased transplant-related mortality in the current era: a CIBMTR analysis. *Blood*. 2016;127: 2427-2438.

47. Elmaagacli AH, Steckel NK, Koldehoff M, et al. Early human cytomegalovirus replication after transplantation is associated with a decreased relapse risk: evidence for a putative virus-versus-leukemia effect in acute myeloid leukemia patients. *Blood*. 2011;118: 1402-1412.

48. Leserer S, Bayraktar E, Trilling M, et al. Cytomegalovirus kinetics after hematopoietic cell transplantation reveal peak titers with differential impact on mortality, relapse and immune reconstitution. *Am J Hematol*. 2021;96:436-445.

49. Duke ER, Williamson BD, Borate B, et al. CMV viral load kinetics as surrogate endpoints after allogeneic transplantation. *J Clin Investig*. 2021;131:e133960.

50. Lablans M, Borg A, Ückert F. A RESTful interface to pseudonymization services in modern web applications. *BMC Med Inform Decis Mak*. 2015;15:2.

51. Seuss H, Dankerl P, Ihle M, et al. Semi-automated de-identification of German content sensitive reports for big data analytics. *Rofo*. 2017; 189:661-671.

52. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53:457-481.

53. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94:496-509.

54. Therneau, T. survival: A Package for Survival Analysis in R (2021). R package version 3.2-11.

55. Kassambara, A., Kosinski, M. & Biecek, P. survminer: Drawing Survival Curves using 'ggplot2' (2021). R package version 0.4.9.

56. Gray, B. cmprsk: Subdistribution Analysis of Competing Risks (2020). R package version 2.2-10.

57. Malone B, Garcia-Duran A, Niepert M. Learning representations of missing data for predicting patient outcomes (2018). Preprint at http://arxiv.org/pdf/1811.04752v1.

58. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks (2015). Preprint at http://arxiv.org/pdf/1511.03677v7.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

> **How to cite this article:** Eisenberg L, the XplOit consortium, Brossette C, et al. Time-dependent prediction of mortality and cytomegalovirus reactivation after allogeneic hematopoietic cell transplantation using machine learning. *Am J Hematol*. 2022;97(10):1309-1323. doi:10.1002/ajh.26671

APPENDIX A: XplOit Consortium

The members of the XplOit consortium were Lisa Eisenberg, Prof. Nico Pfeifer, Jochen Rauch, Kerstin Rohm, Dr Gabriele Weiler, Dr Stephan Kiefer, Dr Jürgen Rissland, Prof. Sigrun Smola, Dr Thorsten Pfuhl, Dr Pascal Feld, Dr Lise Lauterbach-Rivière, Dr Anna Marthaler, Dr Jörg Bittenbring, Dr Dominic Kaddu-Mulindwa, Katharina Götz, Katharina Och, Prof. Thorsten Lehr, Christian Brossette, Stefan Theobald, Yvonne Braun, Prof. Norbert Graf, Abdul Kadir, Dr Ulf Schwarz, Andrea Grandjean, Dr Matthias Ihle, Claudia Riede, Sonja Fix, Dr Amin T. Turki, MD PhD, Prof. Dietrich W. Beelen, MD, PD Dr Hellmut Ottinger, MD and The XplOit Study Team: Dr Nikolaos Tsachakis-Mück, MD, Dr Rashit Bogdanov, Prof. Michael Koldehoff, MD, Dr Nina Steckel, MD, Dr Ji-He Yi, MD, Aiste Fokaite, MD, Dr Vesna Klisanin, MD, PD Dr Lambros Kordelas, MD, Dr Diana Garay, Ximena Gavilanes, MD, Robert F. Lams, MD, Aleksandra Pillibeit, Saskia Leserer and Theresa Graf. Stefan Hilbig and Joachim Weiß kindly provided EHR-related IT support.