Supplemental Table 1. List of candidates included and excluded from predictor variable selection

Included variables prior to transplant

- Age (every 10-year increase)
- Sex (male, female)
- Race (white, non-white)
- Body mass index (categorical with BMI <30, 30-34.9, 35-39.9, 40+)
- Prior autologous transplant (yes, no)
- Disease type (lymphoma = NHL/HD, leukemia and others)
- Donor match/relatedness (matched related, matched unrelated, mismatched related, mismatched unrelated, umbilical cord)
- Conditioning regimen (myeloablative, non-myeloablative)
- GVHD prophylaxis (tacrolimus, cyclosporine, calcineurin inhibitor + sirolimus, other)

Included variables on or prior to day 30 (index date)

- VTE history (PE/LE-DVT, catheter associated DVT, none)
- GVHD history (grade 3-4, grade 0-2)
- Bacterial infection history (yes, no)
- Fungal infection history (yes, no)
- CMV reactivation history (yes, no)
- Admission status at day 30 (inpatient, outpatient)

Included continuous laboratory variables at day 30 +/- 7d (index date)

- All candidate laboratory variables will be tested as either continuous variables or categorical variables at 50^{th} and 90^{th} percentile cut-offs
- White blood cell count (<5, 5-10.9, 11+)
- Hemoglobin (<10, 10-11.9, 12+)
- Platelet (<100, 100-199, 200+)
- Creatinine (<1, 1-1.49, 1.5+)
- Total bilirubin (<0.7, 0.7-1.69, 1.7+)

Excluded variables (and rationale)

- HCT comorbidity index (HCI-CI): missingness >5%
- Karnofsky performance status: missingness >5%
- Timing of historical VTE: colinear with type of VTE (i.e. most of the catheter associated DVTs occurred very recently whereas most of the PE/LE-DVT occurred very remotely)
- Absolute neutrophil count at day 30: colinear with WBC
- Coagulation labs (INR, PT, PTT, or D-dimer): missingness >5%

Supplemental Table 2. Variables selected from stepwise logistic regressions after locking in history of VTE and acute GVHD based on prior knowledge

Model #1: use original variable parametrization (c-statistic = 0.72)

Proposed Risk Factor	OR (95% CI)	Standard Error
History of VTE		
None (n=1570)	Baseline	
CR-DVT (n=81)	2.10 (0.80-5.52)	1.036
PE or LE-DVT (n=52)	2.60 (0.94-7.19)	1.350
Acute GVHD before 30d		
None or mild GVHD (n=1576)	Baseline	
Grade 3-4 (n=127)	1.69 (0.75-3.82)	0.702
Inpatient admission (30d)		
No (n=1423)	Baseline	
Yes (n=280)	2.07 (1.08-3.98)	0.690
Diagnosis of lymphoma		
No (n=1494)	Baseline	
Yes (n=209)	3.46 (1.88-6.36)	1.074
Obesity		
BMI <30 (n=1268)	Baseline	
BMI 30-34.9 (n=289)	1.11 (0.54-2.31)	0.414
BMI 35-39.9 (n=85)	2.52 (1.00-6.33)	1.184
BMI 40+ (n=61)	2.63 (0.96-7.16)	1.345
WBC at day 30		
WBC <5 (n=837)	Baseline	
WBC 5-10.9 (n=664)	1.30 (0.70-2.41)	0.410
WBC 11+ (n=202)	2.21 (1.05-4.65)	0.839
Constant	0.013	0.004

Model #2: use continuous variables without categorization (c-statistic = 0.72)

Proposed Risk Factor	OR (95% CI)	Standard Error
History of VTE		
None (n=1570)	Baseline	
CR-DVT (n=81)	2.08 (0.79-5.44)	1.020
PE or LE-DVT (n=52)	2.47 (0.89-6.89)	1.294
Acute GVHD before 30d		
None or mild GVHD (n=1576)	Baseline	
Grade 3-4 (n=127)	1.70 (0.75-3.85)	0.709
Inpatient admission (30d)		
No (n=1423)	Baseline	
Yes (n=280)	2.12 (1.11-4.04)	0.698
Diagnosis of lymphoma		
No (n=1494)	Baseline	
Yes (n=209)	3.43 (1.87-6.31)	1.065
BMI at baseline		
For every 1 increase	1.05 (1.01-1.10)	0.022
WBC at day 30		
For every 1 increase	1.02 (0.99-1.07)	0.020
Constant	0.004	0.003

Model #3: use simplified cut-off (for final score estimation) (c-statistic = 0.71)

Proposed Risk Factor	OR (95% CI)	Standard Error
History of VTE		
None (n=1570)	Baseline	
CR-DVT (n=81)	2.10 (0.80-5.53)	1.037
PE or LE-DVT (n=52)	2.54 (0.92-7.05)	1.322
Acute GVHD before 30d		
None or mild GVHD (n=1576)	Baseline	
Grade 3-4 (n=127)	1.74 (0.77-3.91)	0 .719
Inpatient admission (30d)		
No (n=1423)	Baseline	
Yes (n=280)	2.02 (1.06-3.86)	0.666
Diagnosis of lymphoma		
No (n=1494)	Baseline	
Yes (n=209)	3.47 (1.89-6.38)	1.077
Obesity		
BMI <35 (n=1557)	Baseline	
BMI 35+ (n=146)	2.54 (1.26-5.13)	0.910
WBC at day 30		
WBC <11 (n=1501)	Baseline	
WBC 11+ (n=202)	1.95 (0.99-3.84)	0.674
Constant	0.015	0.003

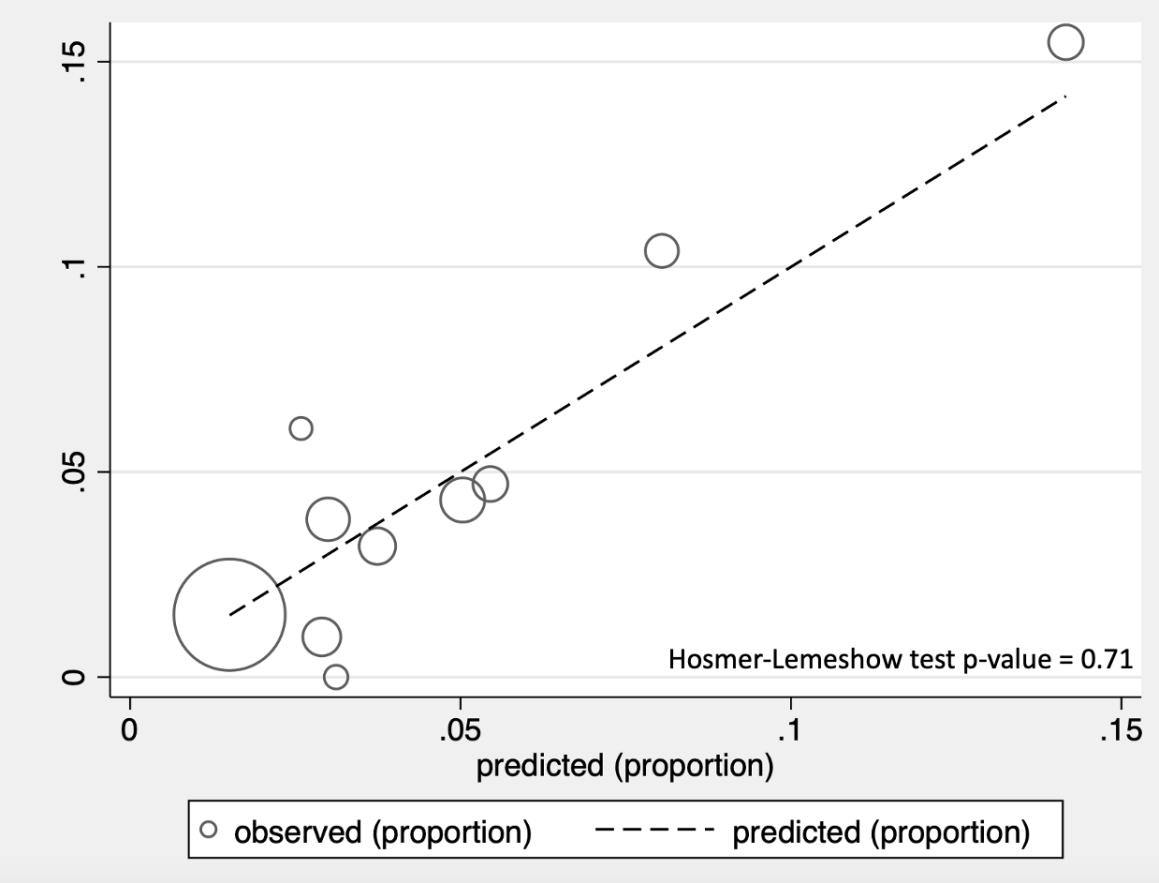
Supplemental Table 3. Variables selected from LASSO cross-validation

ID	lambda	No. of nonzero coef.	CV mean deviance		
2	.0165138	1	1.368851	A 0.lymphoma	
5	.0124921	2	1.31162	A 0.inpatient_day30	
6	.0113823	3	1.286448	A 0.vte_hx2	
8	.0094498	4	1.239995	A 0.agvh_severe_day30	
9	.0086103	6	1.218326	A 2.vte_hx2	4.bmi_cat
10	.0078454	7	1.196694	A 3.bmi_cat	
12	.0065134	8	1.166528	A cr_30d	
14	.0054075	9	1.149977	A wbc_30d	
* 15	.0049271	9	1.149564	U	
16	.0044894	13	1.151643	A 1.race2 plt_30d	2.race2
17	.0040906	15	1.155249	A age_cat	2.regimen
18	.0037272	18	1.161854	A 2.ppx2	3.ppx2
19	.0033961	19	1.168556	A 2.don_match	
21	.0028195	20	1.180657	A 0.aspergillus_day30	
23	.0023408	21	1.192089	A 0.regimen	
26	.0017707	23	1.207116	A 3.race2	3.don_match
32	.0010133	24	1.229605	A 0.bacteremia_day30	
35	.0007665	25	1.237122	A 0.prior_auto	
37	.0006364	26	1.241255	A 0.cmv_day30	

* lambda selected by cross-validation.

Lasso penalized regression was used as an alternative to stepwise regression for variable selection to avoid overfitting. The lasso penalty parameter lambda was selected through 10-fold cross-validation (CV) to minimize the CV mean deviance. The covariate with non-zero coefficients at optimal lambda included all the variables selected in the stepwise regression in addition to creatinine values at 30 days. This prompted us to investigate creatine further using various continuous and categorical transformation. However, we could not detect either a statistical signal or to improve fit the final multivariable model. As creatinine remained a weak covariate chosen by lasso, we decided against fitting in into the final model.

Supplemental Figure 1. Calibration plot and Hosmer-Lemeshow goodness of fit test for the final model



TRIPOD Checklist: Prediction Model Development

Section/Topic	Item	Checklist Item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	3
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	4
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	5
Methods			
Source of data	4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	5
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	5, 6
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5
	5b	Describe eligibility criteria for participants.	5, 6
	5c	Give details of treatments received, if relevant.	--
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	7
	6b	Report any actions to blind assessment of the outcome to be predicted.	--
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	Supp Table
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	--
Sample size	8	Explain how the study size was arrived at.	Fig. 1
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	8
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.	8
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	8, 9
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	8, 9
Risk groups	11	Provide details on how risk groups were created, if done.	8, 9
Results			
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	Fig 1
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	9, 10
Model development	14a	Specify the number of participants and outcome events in each analysis.	10, 11
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	--
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	Table 2, Supplemental Table 2
	15b	Explain how to the use the prediction model.	11, 12
Model performance	16	Report performance measures (with CIs) for the prediction model.	12
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	17
Interpretation	19b	Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence.	13, 14, 15
Implications	20	Discuss the potential clinical use of the model and implications for future research.	13, 16, 17
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	--
Funding	22	Give the source of funding and the role of the funders for the present study.	18

We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.