Prediction of complete remission and survival in acute myeloid leukemia using supervised machine learning

Jan-Niklas Eckardt,[1] Christoph Röllig,[1] Klaus Metzeler,[2] Michael Kramer,[1] Sebastian Stasik,[1] Julia-Annabell Georgi,[1] Peter Heisig,[3] Karsten Spiekermann,[4] Utz Krug,[5] Jan Braess,[6] Dennis Görlich,[7] Cristina M. Sauerland,[7] Bernhard Woermann,[8] Tobias Herold,[4] Wolfgang E. Berdel,[9] Wolfgang Hiddemann,[4] Frank Kroschinsky,[1] Johannes Schetelig,[1] Uwe Platzbecker,[2] Carsten Müller-Tidow,[10,11] Tim Sauer,[10] Hubert Serve,[12] Claudia Baldus,[13] Kerstin Schäfer-Eckart,[14] Martin Kaufmann,[15] Stefan Krause,[16] Mathias Hänel,[17] Christoph Schliemann,[9] Maher Hanoun,[18] Christian Thiede,[1,11] Martin Bornhäuser,[1,11,19] Karsten Wendt[2] and Jan Moritz Middeke[1]

[1]Department of Internal Medicine I, University Hospital Carl Gustav Carus, Dresden; [2]Medical Clinic and Policlinic I Hematology and Cell Therapy. University Hospital, Leipzig; [3]Institute of Software and Multimedia Technology, Technical University Dresden, Dresden; [4]Laboratory for Leukemia Diagnostics, Department of Medicine III, University Hospital, LMU Munich, Munich; [5]Medical Clinic III, Hospital Leverkusen, Leverkusen; [6]Hospital Barmherzige Brueder Regensburg, Regensburg; [7]Institute for Biometrics and Clinical Research, University Münster, Münster; [8]Department of Hematology, Oncology and Tumor Immunology, Charité, Berlin; [9]Department of Internal Medicine A, University Hospital Münster, Münster; [10]Department of Medicine V, University Hospital Heidelberg, Heidelberg; [11]German Consortium for Translational Cancer Research DKFZ, Heidelberg; [12]Department of Medicine 2, Hematology and Oncology, Goethe University Frankfurt, Frankfurt; [13]Department of Hematology and Oncology, University Hospital Schleswig Holstein, Kiel; [14]Department of Internal Medicine 5, Paracelsus Medical Private University Nuremberg, Nuremberg; [15]Department of Hematology, Oncology and Palliative Care, Robert-Bosch Hospital, Stuttgart; [16]Department of Internal Medicine 5, University Hospital Erlangen, Erlangen; [17]Department of Internal Medicine 3, Klinikum Chemnitz GmbH, Chemnitz; [18]Department of Hematology and Stem Cell Transplantation, University Hospital Essen, Essen and [19]National Center for Tumor Diseases (NCT), Dresden, Germany

Correspondence: J-N. Eckardt
jan-niklas.eckardt@uniklinikum-dresden.de

Received: September 15, 2021.
Accepted: March 31, 2022.
Early view: June 16. 2022.

https://doi.org/10.3324/haematol.2021.280027

Machine Learning Pipeline

To enable a customizable and reusable technological approach and to achieve optimal results, we designed a data-driven software platform. The embedded, automated ML pipeline integrates state-of-the-art software technology for data management, feature transformation, ML models and training algorithms and use-case specification (such as specific result exports), and consists of five subsequent steps, which were executed in an iterative manner to approach step-wisely the optimal configuration. 1. Data import & modeling: data from different sources were aggregated and stored in a MySQL (Oracle, Austin, Texas, USA) database, allowing efficient data access and format alignment. In that way, pooled data from the above-mentioned clinical trials and the SAL patient registry were collected and 212 multimodal variables (clinical data, laboratory parameters as well as molecular and cytogenetic genetic data) became available (see Tab. S4 for a full list of variables used in the model). 2. Model enhancement: Relevant attributes were selected by domain experts (physicians) and dimensionality was reduced by excluding sparse features (cut-off 1%). This way, redundancies were removed and the risk of collinearities and overfitting was reduced. 3. Data transformation: the resulting object graph was transformed in a uniform and robust representation for ML models, i.e. as the data included a variety of numerical values with different ranges, feature scaling was performed by standardizing numerical values to the z-score. Nominal and ordinal variables were one-hot encoded. As not all ML models can compute missing values and since we aimed to evaluate a variety of ML models for their capabilities of predicting CR and 2-year OS, an imputation of missing values was essential and thus, integrated. Missing ordinal, discrete and continuous variables were imputed with the median of the respective variable. Missing nominal values were labeled ‘unknown’. To reduce dimensionality and thereby the risk of overfitting, dynamic feature selection was used, i.e. features were selected according to their support by five feature selection metrics: linear correlation, chi-square test, recursive feature elimination, lasso regularization and random

forest ranking. To be included in an ML model, a variable had to pass a pre-determined threshold of overall predictive power determined by summing the normalized scores of these five feature selection algorithms. Precisely, each single feature selection metric evaluated every single feature for its prognostic impact resulting in a score ranging from 0 to 1, where 0 means no impact on outcome and 1 means high association with outcome. As an example: Potentially, a feature could reach a prognostic score of 2.5. That could mean that two feature selection metrics gave a score of 1, one metric gave 0.4, one metric 0.1 and finally the last metric 0. Alternatively, the feature could have been graded with 0.5 from every single feature selection metric. Essentially, this resembles a mathematical representation of a Venn diagram where the overlap for a single feature between the metrics are expressed numerically ranging from 0 – 5 (very low to very high prognostic impact). By using five rather than just one feature selection metric and summing the resulting prognostic score we aimed to reduce bias introduced by individual algorithms. Subsequently, an automated cut-off was used for including the scored features in the classification models. This cut-off was iteratively determined by maximizing the average AUROCs of all classification algorithms, i. e. the number of features included on the model was determined by cutting off less predictive features when the classification algorithms reached their peak performance in the test set. For both CR and 2-year OS, this point was achieved at a prognostic score of 0.5. Including features below 0.5 again decreased classification performance likely due to introduction of random noise. In that way, features of high redundancy or low entropy were automatically filtered out. In contrast to upfront regression analysis of all 212 parameters, the proposed ML method controls for potential type I and II errors in addition to agnostic and data-driven analysis rather than hypothesis-based parameter testing. As multiple testing greatly increases type I error rate, especially for such a multidimensional data set, conventional approaches require post-hoc correction, e. g. using Bonferroni correction. This would introduce a very conservative significance level, especially in the context of 212 variables, which in turn would

increase the risk for type II errors. By pre-selecting parameters and thereby reducing the number of univariate regression models needed for analysis, type I and II error rate are more controlled for than with upfront regression analysis for all individual parameters. 4. Machine learning classifiers: Applied ML models were Random Forest (RF), Gradient Boosting, adaptive Boosting, linear, polynomial and radial basis function kernel (RBF) support vector machines (SVM), k-nearest neighbor (KNN), logistic regression (LR) and artificial neural nets (ANN). The prepared data from step 1-3 was divided in a training and test set with a ratio of 9:1 using stratified randomization and tenfold cross-validation. That means that for each of ten iterations the data set is reshuffled and a sample is drawn completely at random where 90% of patients are allocated to the training set and 10% of patients are allocated to the test set. The test set is then strictly withheld from the training data to prevent overfitting of the classifiers. Overfitting is the notion that a classifier 'memorizes' training data rather than learning abstract feature representations derived from the data. This would result in high classification performance in the training set and poor performance (low generalizability) in the test set or with external data. To prevent this, stratified randomization ensured the 9:1 ratio for each single iteration of the tenfold cross-validation. By performing this process over ten iterations, the risk of selection bias, i. e. the notion that the patients in the training vs. the test set differ substantially e. g. with respect to risk or outcome, is greatly reduced since every patient has the chance to be allocated to either training or test set in ten different iterations. By introducing a predefined seed for the random generators before each run, reproducibility is ensured. Finally, performance for the test set was averaged. All reported performance measures are derived from averaged scores of tenfold cross-validation on test sets only. This approach enables the ML pipeline within the platform to train different ML models on the base of a stable data set, making the results comparable to search for the optimal model and configuration. 5. Visualization & analysis: Finally, the ML models' output is automatically visualized and performance can be assessed using a pre-defined cluster of performance

metrics. This way, both clinicians and ML engineers receive immediate feedback of model performance and selected features.

With the support of hyperparameter optimization, which filters parameters that do not belong to the model itself, utilizing Bayesian optimization with Gaussian processes, the entire ML pipeline was executed several times, producing an automated documentation for each model and configuration. Hyperparameter search was performed using scikit.learn version 0.23.2. including model stabilization (https://scikit-learn.org/stable/modules/model_evaluation.html#common-cases-predefined-values) and search space optimization (https://scikit-optimize.github.io/stable/modules/generated/skopt.space.space.Real.html) using default settings. Medical and ML experts collaborated to discuss the intermediate results to refine configuration, feature selection and preparation techniques, data transformation and ML technology as well as the result representation to optimize the pipeline after each run to achieve optimal results for the CR/Cri and OS use case. For external validation, pre-trained models were tested on 664 AML patients from the multi-center AML Cooperative Group bioregistry. Model building, evaluation and visualization was performed in Python 3.8 (Python Software Foundation, Fredericksburg, Virginia, USA). Python packages that were used are summarized in Tab. S6.

Code availability

Code that was generated for the purpose of this work is publicly available under https://github.com/sit-institute/sal-cr/

trial name	clinicaltrials.gov identifier	trial duration	protocol summary
AML96	NCT00180115	1996-2008	risk-adapted postremission treatment regarding allogeneic stem cell transplantation for high-risk AML and related allogeneic and autologous stem cell transplantation for standard-risk AML, and randomization between intermediate-dose and high-dose cytarabine within the first post-remission course
AML2003	NCT00180102	2003-2009	early allogeneic stem cell transplantation in post-induction aplasia for high-risk AML, factorial design with four therapy arms with two factors of two stages (intensified vs. standard therapy and cytarabine vs. cytarabine + mitoxantrone + amsacrin)
AML60+	NCT00180167	2005-2010	Patients ≥ 60 years, mitoxantron on day 1,2,3 + cytarabine on days 1,3,5,7 vs. DA 7+3
SORAML	NCT00893373	2011-2014	Standard therapy + sorafenib vs. standard therapy + placebo
SAL bioregistry	NCT03188874	2010-present	Prospective registry of AML patients
AMLCG-1999	NCT00266136	1999-2007	double induction with HAM-HAM, multiple course G-CSF or myeloablative consolidation with Bu/Cy and autologous blood stem cell transplantation instead of maintenance vs. standard therapy
AMLCG-2008	NCT01382147	2008-2012	S-HAM escalated for younger patients and

	S-HAM basis for elderly patients vs. TAD-HAM (younger) or HAM-HAM (elderly)

Table S1. Summary of trial data used for retrospective analysis

TruSight Myeloid Sequencing Panel				
ABL1	CEBPA	HRAS	MYD88	SF3B1
ASXL1	CSF3R	IDH1	NOTCH1	SMC1A
ATRX	CUX1	IDH2	NPM1	SMC3
BCOR	DNMT3A	IKZF1	NRAS	SRSF2
BCORL1	ETV6/TEL	JAK2	PDGFRA	STAG2
BRAF	EZH2	JAK3	PHF6	TET2
CALR	FBXW7	KDM6A	PTEN	TP53
CBL	FLT3	KIT	PTPN11	U2AF1
CBLB	GATA1	KRAS	RAD21	WT1
CBLC	GATA2	MLL	RUNX1	ZRSR2
CDKN2A	GNAS	MPL	SETBP1	

Table S2. Summary of the 54 genes targeted by the TruSight Myeloid Sequencing Panel (Illumina, San Diego, CA, USA).

Section & Topic	**No**	**Item**	**Reported on page #**
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	2-3
	4	Study objectives and hypotheses	2-3
METHODS			
Study design	5	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	3-4 + supplements
Participants	6	Eligibility criteria	3-4
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	3-4
	8	Where and when potentially eligible participants were identified (setting, location and dates)	3-4 + supplements

	9	Whether participants formed a consecutive, random or convenience series	3-4 + supplements
Test methods	10a	Index test, in sufficient detail to allow replication	4-5 + supplements
	10b	Reference standard, in sufficient detail to allow replication	n.a.
	11	Rationale for choosing the reference standard (if alternatives exist)	n.a.
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	5 + supplements
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	n.a.
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	n.a.
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	n.a.
Analysis	14	Methods for estimating or comparing measures of diagnostic accuracy	5
	15	How indeterminate index test or reference standard results were handled	Supplements
	16	How missing data on the index test and reference standard were handled	supplements
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	supplements
	18	Intended sample size and how it was determined	n.a.
RESULTS			
Participants	19	Flow of participants, using a diagram	Figure 1
	20	Baseline demographic and clinical characteristics of participants	Table 1 + Table S4-S5
	21a	Distribution of severity of disease in those with the target condition	Table 1 + Table S4-S5
	21b	Distribution of alternative diagnoses in those without the target condition	n.a.
	22	Time interval and any clinical interventions between index test and reference standard	n.a.
Test results	23	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	n.a.
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	6-9, Figure 2, Figure 4-6
	25	Any adverse events from performing the index test or the reference standard	n.a.
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	13-14
	27	Implications for practice, including the intended use and clinical role of the index test	12-14
OTHER INFORMATION			
	28	Registration number and name of registry	3
	29	Where the full study protocol can be accessed	n.a.
	30	Sources of funding and other support; role of funders	15

Table S3. STARD checklist

Variable	SAL (training and testing)		AMLCG (external validation)	
	n	%	n	%
clinical				
age				
height				
weight				
sex				
AML type, de novo	1180	85.32%	570	85.84%
AML type, secondary	146	10.56%	59	8.89%
AML type, therapy-related	40	2.89%	35	5.27%
extramedullary disease	202	14.61%	16/270	5.93%
Fever during induction phase	361	26.10%	n.a.	
laboratory values				
hemoglobin				
white blood cell count				
platelet count				
bone marrow blast count				
peripheral blood blast count				
fibrinogen level				
LDH level				
molecular genetics				
ASXL1	124	8.97%	73	10.99%
ATRX	3	0.22%	n.a.	
BCOR	61	4.41%	46	6.93%
BCORL1	49	3.54%	15	2.26%
BRAF	6	0.43%	1	0.15%
CALR	1	0.07%	n.a.	
CBL	27	1.95%	13	1.96%
CBLB	2	0.14%	n.a.	
CDKN2A	3	0.22%	2	0.30%
CEBPA, monoallelic (TAD)	41	2.96%	8	1.20%
CEBPA, monoallelic (bZIP)	30	2.17%	11	1.66%
CEBPA, double-mutated	91	6.58%	27	4.07%
CSF3R	20	1.45%	13	1.96%
CUX1	34	2.46%	2	0.30%
DNMT3A	396	28.63%	211	31.78%
ETV6	9	0.65%	15	2.26%
EZH2	53	3.83%	28	4.22%
FBXW7	3	0.22%	2	0.30%
FLT3-ITD	280	20.25%	178	26.81%
FLT3-ITD ratio				
FLT3-TKD	62	4.48%	n.a.	
GATA2	80	5.78%	27	4.01%
HRAS	2	0.14%	1	0.15%
IDH1	122	8.82%	45	6.78%

IDH2	197	14.24%	93	14.01%
IKZF1	36	2.60%	n.a.	
JAK2	18	1.30%	8	1.20%
KDM6A	9	0.65%	13	1.96%
KIT	73	5.28%	27	4.07%
KRAS	79	5.71%	41	6.17%
MPL	5	0.36%	n.a.	
MYD88	2	0.14%	n.a.	
NOTCH1	24	1.74%	8	1.20%
NPM1	466	33.69%	221	33.28%
NRAS	229	16.56%	144	21.69%
PDGFRA	1	0.07%	n.a.	
PHF6	41	2.96%	16	2.41%
PTEN	3	0.22%	1	0.15%
PTPN11	100	7.23%	68	10.24%
RAD21	50	3.62%	37	5.57%
RUNX1	134	9.69%	102	15.36%
SETBP1	7	0.51%	3	0.45%
SF3B1	41	2.96%	23	3.46%
SMC1A	22	1.59%	17	2.56%
SMC3	18	1.30%	23	3.46%
SRSF2	72	5.21%	65	9.79%
STAG2	71	5.13%	44	6.63%
TET2	247	17.86%	102	15.36%
TP53	102	7.38%	63	9.49%
U2AF1	36	2.60%	27	4.07%
WT1	102	7.38%	86	12.95%
ZRSR2	19	1.37%	5	0.75%
cytogenetics				
Karyotype, complex	152	10.99%	75	11.29%
Karyotype, neither normal nor complex	463	33.48%	259	39.02%
Karyotype, normal	709	51.27%	330	49.69%
t(6;9)	5	0.36%	5	0.75%
t(11;19)	1	0.07%	n.a.	
abn(3q)	21	1.52%	n.a.	
t(1;3)	5	0.36%	n.a.	
t(3;21)	4	0.29%	n.a.	
t(2;3)	1	0.07%	n.a.	
del(3q)	5	0.36%	n.a.	
add(3q)	2	0.14%	n.a.	
t(3;4;3)	1	0.07%	n.a.	
t(3;8)	1	0.07%	n.a.	
t(3;6)	1	0.07%	n.a.	
t(3;7)	1	0.07%	n.a.	
add(7q)	1	0.07%	n.a.	
del(7q)	18	1.30%	36	5.42%

+8	79	5.71%	n.a.	
-Y	14	1.01%	n.a.	
del(9q)	12	0.87%	n.a.	
del(20q)	5	0.36%	n.a.	
inv(3)	7	0.51%	13	1.96%
-5	7	0.51%	n.a.	
del(5q)	44	3.18%	54	8.13%
-7	33	2.39%	n.a.	
-17	2	0.14%	n.a.	
t(v;11)(v;q23)	13	0.94%	n.a.	
add(11q23)	1	0.07%	n.a.	
t(6;11)	1	0.07%	n.a.	
t(10;11)	2	0.14%	n.a.	
t(1;11)	1	0.07%	n.a.	
t(11;17)	1	0.07%	n.a.	
inv(11)	2	0.14%	n.a.	
t(5;11)	1	0.07%	n.a.	
t(9;10;11)	1	0.07%	n.a.	
t(3;11;15)	1	0.07%	n.a.	
abn(17p)	6	0.43%	n.a.	
add(17p)	1	0.07%	n.a.	
del(17p)	32	2.31%	39	5.87%
inv(16)	58	4.19%	18	2.71%
del(9p)	1	0.07%	n.a.	
del(11q)	6	0.43%	n.a.	
del(12p)	4	0.29%	n.a.	
del(16q)	4	0.29%	n.a.	
del(10p)	3	0.22%	n.a.	
del(21q)	1	0.07%	n.a.	
del(6q)	1	0.07%	n.a.	
del(17q)	2	0.14%	n.a.	
del(1p)	2	0.14%	n.a.	
del(15q)	1	0.07%	n.a.	
del(13q)	2	0.14%	n.a.	
del(1q)	1	0.07%	n.a.	
del(3p)	1	0.07%	n.a.	
del(4q)	1	0.07%	n.a.	
-22	1	0.07%	n.a.	
-13	2	0.14%	n.a.	
-18	2	0.14%	n.a.	
-X	5	0.36%	n.a.	
-15	1	0.07%	n.a.	
add(20p)	1	0.07%	n.a.	
add(18q)	1	0.07%	n.a.	
add(12p)	2	0.14%	n.a.	
add(14q)	2	0.14%	n.a.	
add(9p)	1	0.07%	n.a.	

add(15q)	1	0.07%	n.a.	
add(19p)	1	0.07%	n.a.	
add(21q)	2	0.14%	n.a.	
add(8q)	1	0.07%	n.a.	
add(22q)	1	0.07%	n.a.	
add(17q)	1	0.07%	n.a.	
+6	3	0.22%	n.a.	
+11	9	0.65%	n.a.	
+9	3	0.22%	n.a.	
+14	3	0.22%	n.a.	
+4	10	0.72%	n.a.	
+19	6	0.43%	n.a.	
+13	10	0.72%	n.a.	
+22	20	1.45%	n.a.	
+21	14	1.01%	n.a.	
+1	1	0.07%	n.a.	
+5	2	0.14%	n.a.	
+12	1	0.07%	n.a.	
+7	1	0.07%	n.a.	
+10	2	0.14%	n.a.	
+X	1	0.07%	n.a.	
+Y	3	0.22%	n.a.	
+15	1	0.07%	n.a.	
+20	2	0.14%	n.a.	
+23	1	0.07%	n.a.	
+3	1	0.07%	n.a.	
+r	2	0.14%	n.a.	
mar	13	0.94%	n.a.	
XXYY	1	0.07%	n.a.	
dup(21)(q22q22)	1	0.07%	n.a.	
dup(17)(q21q25)	1	0.07%	n.a.	
dup(8)	1	0.07%	n.a.	
t(9;11)	20	1.45%	17	2.56%
t(4;14)(q11;q32)	1	0.07%	n.a.	
t(8;9)	2	0.14%	n.a.	
t(5;18)(q35;q21)	1	0.07%	n.a.	
t(16;16)	18	1.30%	2	0,3%
t(9;21)	2	0.14%	n.a.	
t(4;21)(q11;q11)	1	0.07%	n.a.	
t(1;4)(q25;q12)	1	0.07%	n.a.	
t(3;5)	4	0.29%	n.a.	
t(8;11)	1	0.07%	n.a.	
t(2;15)	1	0.07%	n.a.	
t(7;14)	1	0.07%	n.a.	
t(7;9)	1	0.07%	n.a.	
t(6;12)	1	0.07%	n.a.	
t(2;14)	2	0.14%	n.a.	

t(5;21)	1	0.07%	n.a.	
t(7;11)	2	0.14%	n.a.	
t(7;21)	1	0.07%	n.a.	
t(3;11)	1	0.07%	n.a.	
t(13;21)	1	0.07%	n.a.	
t(1;17)	1	0.07%	n.a.	
t(5;9)	1	0.07%	n.a.	
t(10;11)	1	0.07%	n.a.	
t(8;21)	52	3.76%	26	3.92%
t(12;22)	1	0.07%	n.a.	
t(4;22)	1	0.07%	n.a.	
t(1;8;16)	1	0.07%	n.a.	
t(2;5;10)	1	0.07%	n.a.	
t(7;12;12)	1	0.07%	n.a.	
ins(21)	1	0.07%	n.a.	
i(17)(q10)	6	0.43%	n.a.	
i(22)(q10)	1	0.07%	n.a.	
idic(X)	1	0.07%	n.a.	
inv(8)	1	0.07%	n.a.	
inv(9)	3	0.22%	n.a.	
inv(17)	1	0.07%	n.a.	
inv(10)	1	0.07%	n.a.	
inv(11)(1)	1	0.07%	n.a.	
der(16)t(1;16)	2	0.14%	n.a.	
der(1;7)	1	0.07%	n.a.	
der(2)(p23)	1	0.07%	n.a.	
der(10)	1	0.07%	n.a.	
der(9)	3	0.22%	n.a.	
der(19)	1	0.07%	n.a.	
der(18)	2	0.14%	n.a.	
der(1;14)	1	0.07%	n.a.	
der(12)	1	0.07%	n.a.	

Table S4. Multimodal data including clinical data, laboratory values, molecular genetics and cytogenetics were available for dynamic feature selection and subsequent model building.

Variables	AML96	AML2003	AML60+	SORAML	Validation
N of patients	943	191	53	196	664
age, median (IQR)	60 (47 – 67)	48 (39 – 55)	69 (66 – 73)	50 (44 – 55)	57 (44 – 66)
sex, n (%)					
Female	439 (46.6)	90 (47.1)	33 (62.3)	99 (50.5)	328 (49.4)
Male	504 (53.4)	101 (52.9)	20 (37.7)	97 (49.5)	336 (50.6)
AML status, n (%)					
de novo	773 (82.0)	181 (94.8)	49 (92.5)	177 (90.3)	570 (85.8)
Secondary	123 (13.0)	1 (0.5)	3 (5.7)	14 (7.1)	59 (8.9)
therapy-associated	31 (3.3)	2 (1.0)	0 (3.0)	5 (2.6)	35 (5.3)
missing, n (%)	16 (1.7)	7 (3.7)	1 (1.9)	0	
FAB classification, n (%)					
M0	39 (4.1)	2 (1.0)	0	8 (4.1)	35 (5.4)
M1	201 (21.3)	59 (30.9)	22 (41.5)	44 (22.4)	157 (23.6)
M2	323 (34.3)	57 (29.8)	20 (37.7)	58 (29.6)	178 (26.8)
M3	0	0	0	0	0
M4	169 (17.9)	44 (23.0)	3 (5.7)	32 (16.3)	163 (24.5)
M5	141 (15.0)	15 (7.9)	2 (3.8)	31 (15.8)	83 (12.5)
M6	33 (3.5)	5 (2.6)	0	8 (4.1)	19 (2.9)
M7	6 (0.6)	0	0	0	3 (0.5)
missing, n (%)	31 (3.3)	9 (4.7)	6 (11.3)	15 (7.7)	26 (3.9)
ELN2017 category, n (%)					
Favorable	307 (32.6)	120 (62.8)	17 (32.1)	74 (37.8)	231 (34.8)
Intermediate	378 (40.1)	41 (21.5)	14 (26.4)	77 (39.3)	166 (25.0)
Adverse	205 (21.7)	2 (1.0)	1 (1.9)	33 (16.8)	250 (37.7)
missing, n (%)	53 (5.6)	28 (14.7)	21 (39.6)	12 (6.1)	17 (2.6)
Complex karyotype (≥ 3 abnormalities), n (%)	122 (12.9)	8 (4.2)	0	22 (11.2)	75 (11.3%)
missing, n (%)	0	0	0	114 (58.2)	0
Extramedullary disease, n (%)	181 (19.2)	4 (2.1)	4 (7.5)	12 (6.1)	16 (5.9)
missing, n (%)	132 (14.0)	3 (1.6)	2 (3.8)	0	379 (57.1)
WBC, median (IQR) in GPt/l	12.0 (2.9 – 49.2)	13.7 (3.3 – 48.1)	8.1 (2.2 – 32.1)	8.8 (2.4 – 28.3)	23.8 (6.4 – 60.3)
Hb, median (IQR) in mmol/l	5.7 (4.9 – 6.5)	5.7 (4.9 – 6.6)	5.7 (5.1 – 6.3)	9.0 (8.0 – 10.3)	5.6 (5.0 – 6.3)
Plt, median (IQR) in GPt/l	51 (28 – 98)	53 (29 – 96.5)	49 (30 – 97)	58 (30 – 110)	53 (30 – 102)
LDH, median (IQR) in U/l	408 (254 – 745)	463 (283 – 827)	409 (251 – 698)	354 (221 – 527)	466 (291 – 787)
BM blasts, median (IQR) in %	61 (42 – 78)	60.5 (37 – 78.5)	56 (35 – 76.5)	63 (42 – 80)	80 (58 – 90)

PB blasts, median (IQR) in %	30 (7 – 67)	27 (6 – 64)	20 (2 – 57.5)	20 (4 – 57)	23 (4.5 – 67)
Achieved CR after induction therapy, n (%)	610 (64.7)	182 (95.3)	46 (86.8)	170 (86.7)	445 (67.0)
Median OS (months)	12.2	41.4	9.4	17.1	17.3
OS ≥ 2 years, n (%)	335 (35.5)	134 (70.2)	17 (32.1)	124 (63.3)	290 (43.7)

Table S5. Baseline patient characteristics according to individual trials used in training and testing as well as external validation. FAB: French-American-British Classification; ELN2017: European Leukemia Net 2017; WBC: white blood cell count; Hb: hemoglobin; Plt: platelet count; BM: bone marrow; OS: overall survival; PB: peripheral blood; CR: complete remission; n/N: number; IQR: interquartile range; n.a. – not available

package	version
click	7.1.2
coverage	5.3
flake8	3.8.4
matplotlib	3.3.2
missingno	0.4.2
numpy	1.18.5
numpydoc	1.1.0
pandas	1.1.4
pytablewriter	0.58.0
python-dotenv	0.15.0
scikit-learn	0.23.2
seaborn	0.11.0
sklearn-pandas	2.0.2
Sphinx	3.3.0
sphinx-rtd-theme	0.5.0
PyYAML	5.3.1
xgboost	1.2.1
yellowbrick	1.2
imbalanced-learn	0.7.0
sphinxcontrib-images	0.9.2
scikit-optimize	0.8.1
tune-sklearn	0.1.0
ray[tune]	1.0.1.

Table S6. Python packages used for model building

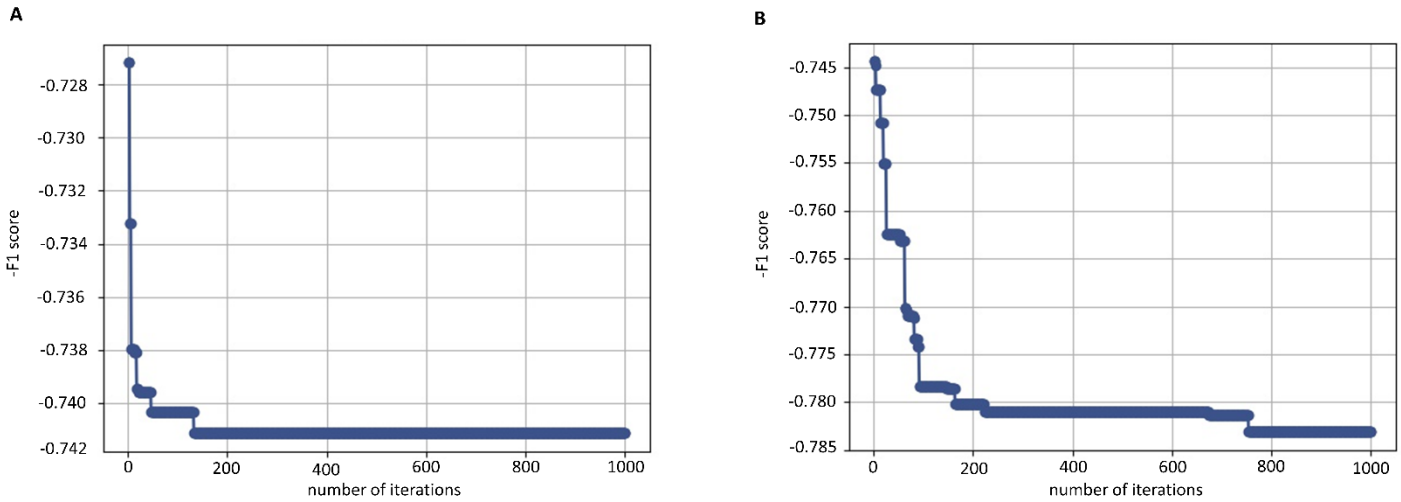

Figure S1 Convergence plot for prediction of complete remission (CR). Logistic regression (A) and Random Forest (B) were selected for hyperparameter tuning for CR classification. Both converged over 1000 iterations achieving a final F1-score of 0.7411 (A) and 0.7831 (B), respectively.

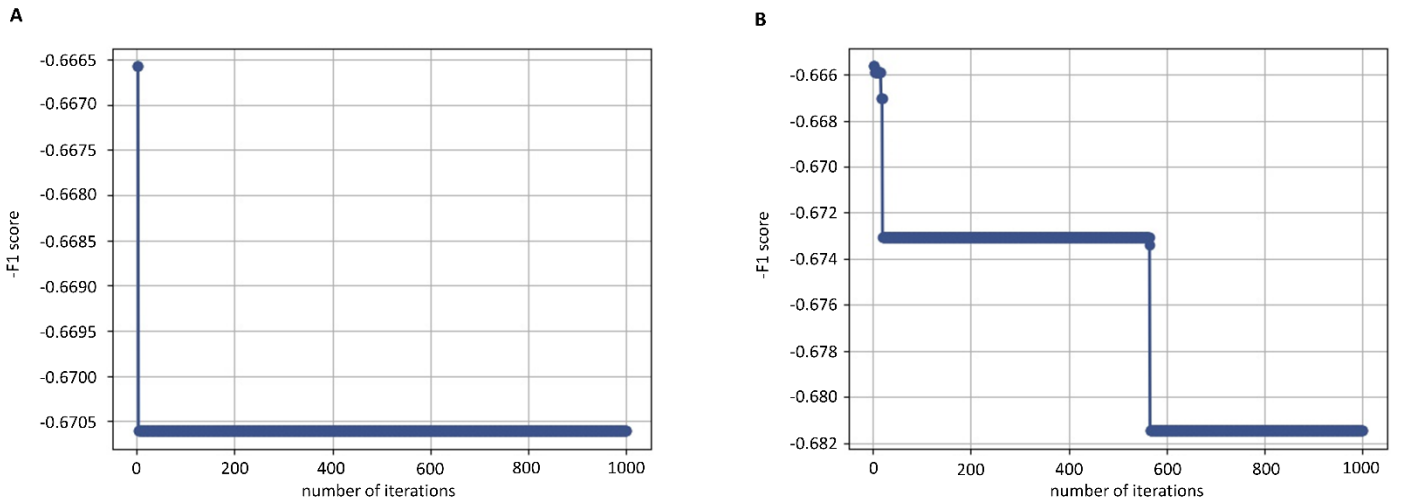

Figure S2 Convergence plot for prediction of overall survival (OS). Logistic regression (A) and Random Forest (B) were selected for hyperparameter tuning for classification of OS above 24 months. Both converged over 1000 iterations achieving a final F1-score of 0.6706 (A) and 0.6815 (B), respectively.

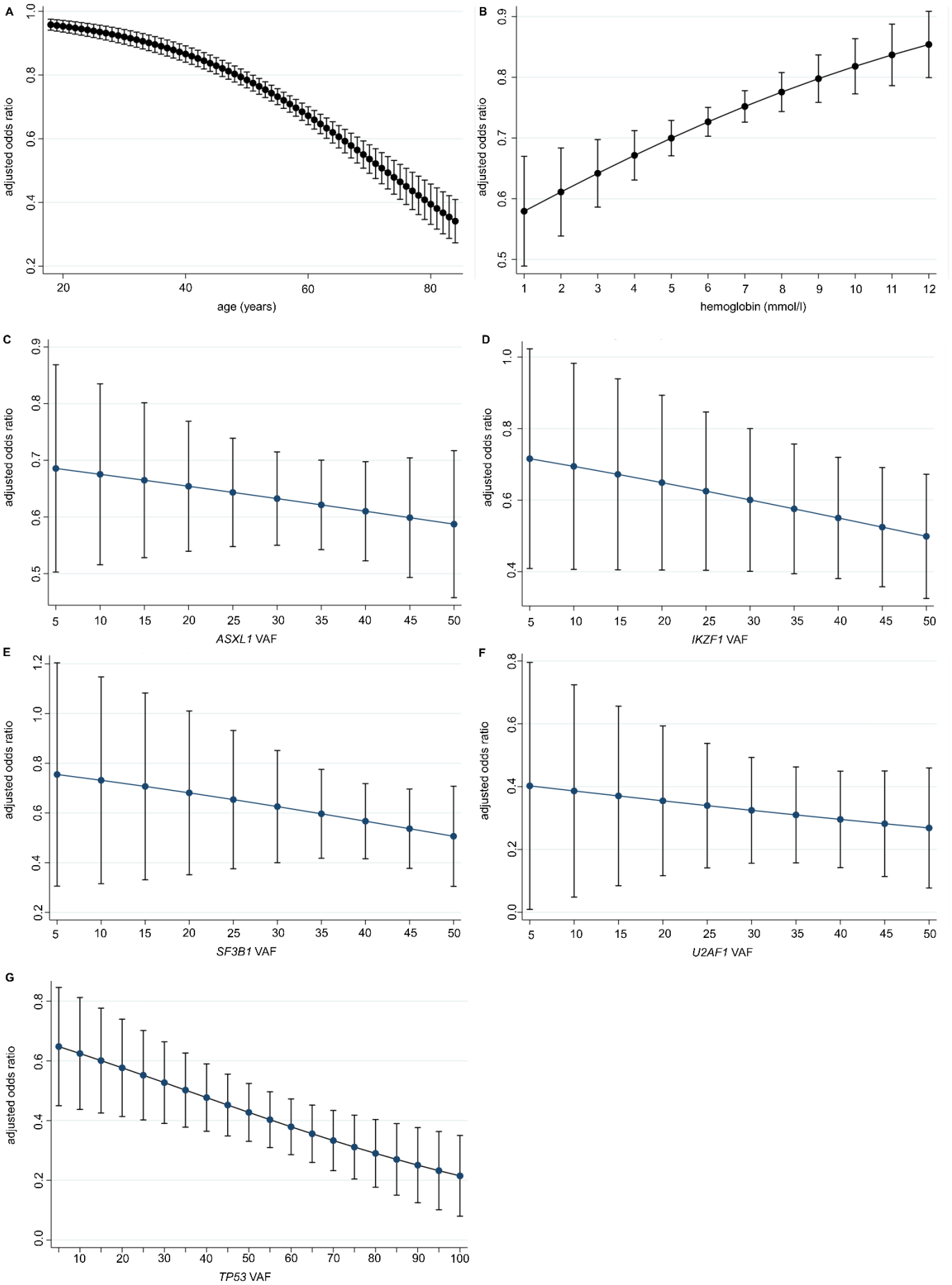
A
adjusted odds ratio
1.0
0.8
0.6
0.4
0.2
20
40
60
80
age (years)
B
adjusted odds ratio
0.9
0.8
0.7
0.6
0.5
1
2
3
4
5
6
7
8
9
10
11
12
hemoglobin (mmol/l)
C
adjusted odds ratio
0.9
0.8
0.7
0.6
0.5
5
10
15
20
25
30
35
40
45
50
ASXL1 VAF
D
adjusted odds ratio
1.0
0.8
0.6
0.4
5
10
15
20
25
30
35
40
45
50
IKZF1 VAF
E
adjusted odds ratio
1.2
1.0
0.8
0.6
0.4
0.2
5
10
15
20
25
30
35
40
45
50
SF3B1 VAF
F
adjusted odds ratio
0.8
0.6
0.4
0.2
0.0
5
10
15
20
25
30
35
40
45
50
U2AF1 VAF
G
adjusted odds ratio
0.8
0.6
0.4
0.2
0.0
10
20
30
40
50
60
70
80
90
100
TP53 VAF

Figure S3 Adjusted odds ratios for continuous variables regarding prediction of complete remission (CR). (A) Age ranged between 18 and 84 years. Increasing age was significantly associated with decreased odds for achieving CR with intensive induction therapy. **(B)** Increased hemoglobin (until normal values) was associated with increased odds of achieving CR. For molecular genetics associated with CR such as *ASXL1* (**C**), *IKZF1* (**D**), *SF3B1* (**E**), *U2AF1* (**F**), *TP53* (**G**), higher variant allele fraction (VAF) was associated with decreased CR rates. For biallelic *CEBPA* mutations and *CEBPA*-bZIP, VAF was not available for analysis.

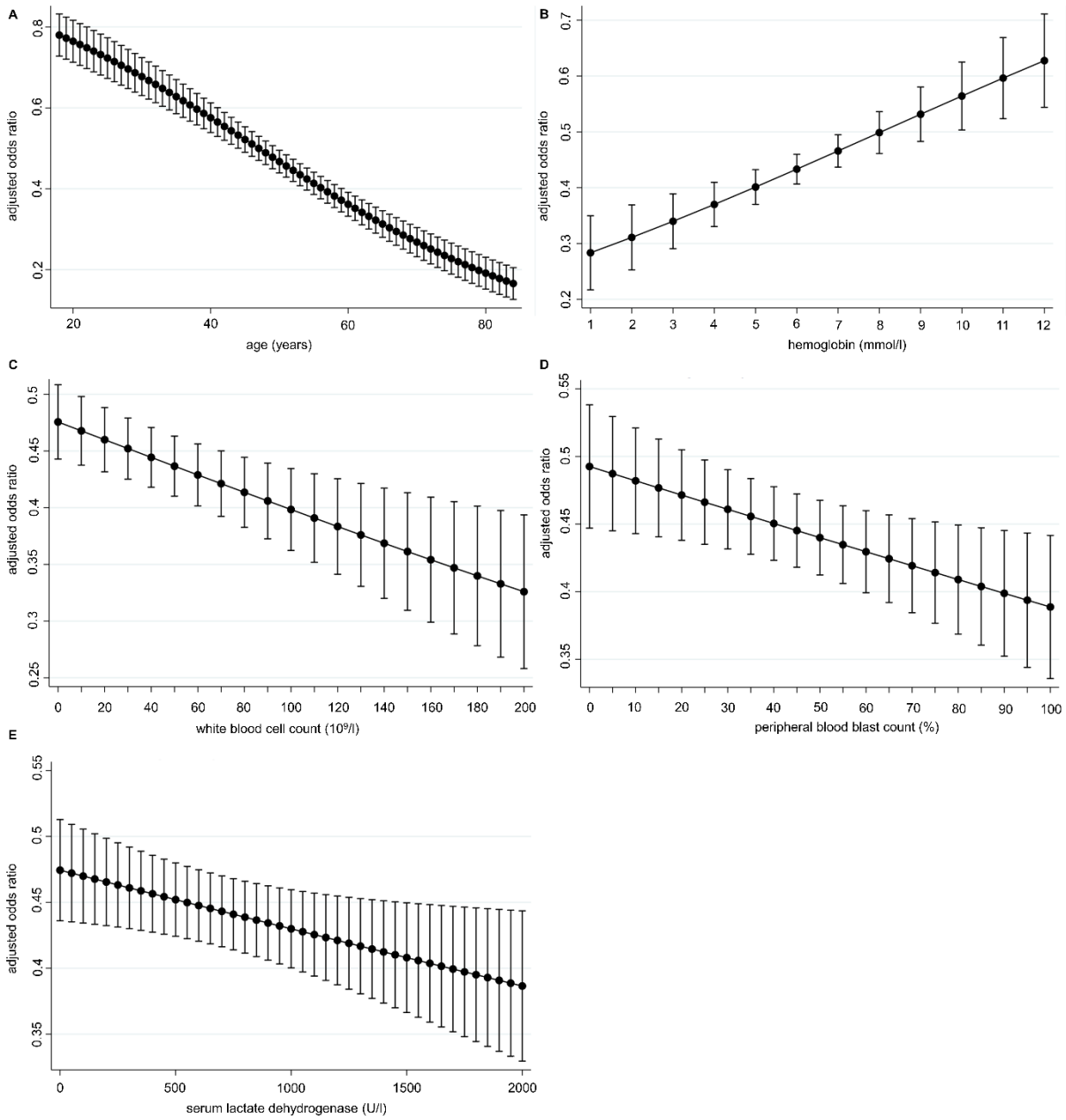

Figure S4 Adjusted odds ratios for continuous variables regarding prediction of overall survival ≥ 2 years. (A) Age ranged between 18 and 84 years. Increasing age was significantly associated with decreased odds for survival for 2 years or longer. **(B)** Increased hemoglobin (until normal values) was associated with increased odds of surviving 2 years or longer. An increase in white blood cell count (**C**), peripheral blood blast count (**D**) and serum lactate dehydrogenase (**E**) was associated with decreased odds of survival.

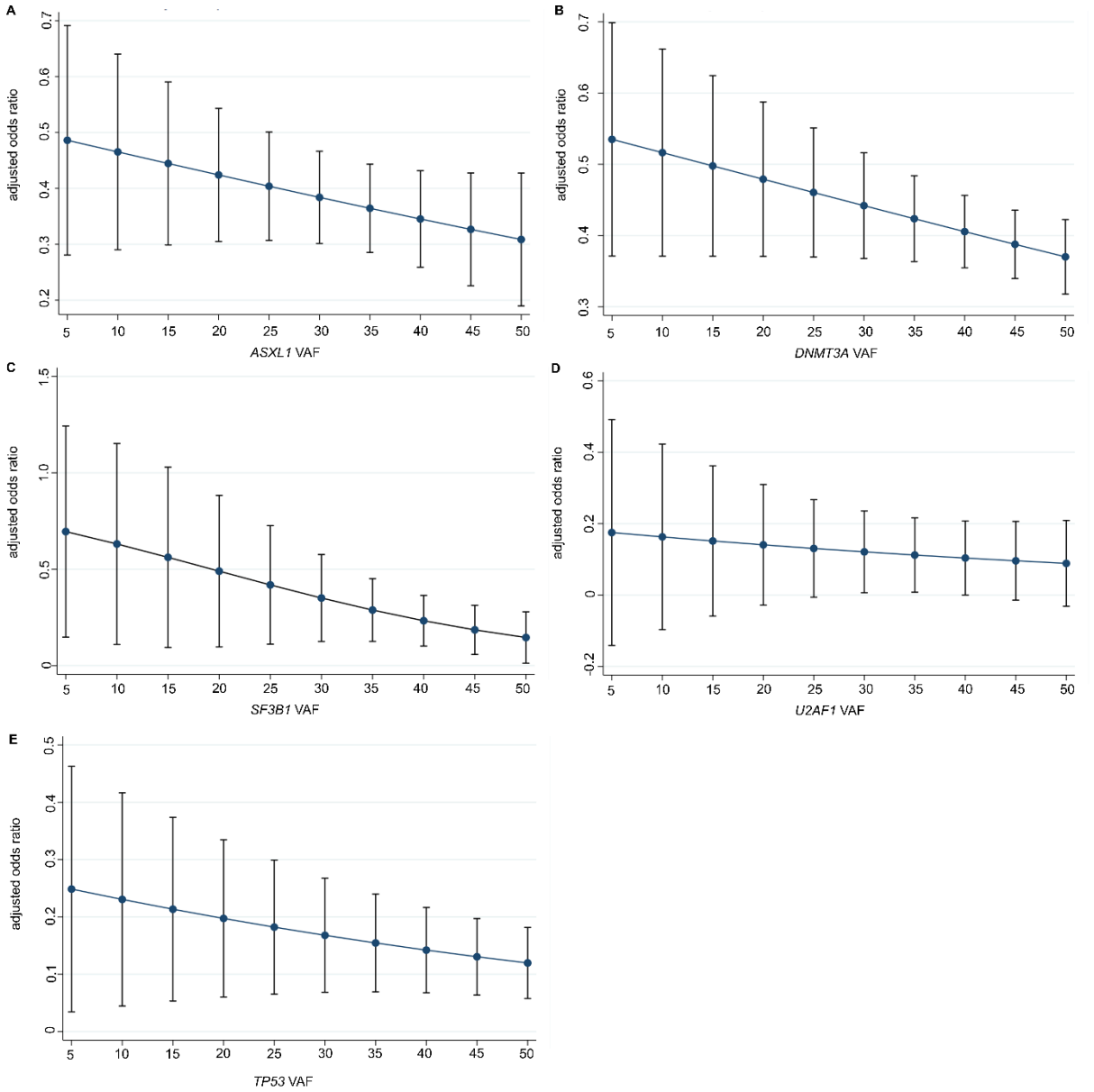

Figure S5 Adjusted odds ratios for continuous variables regarding prediction of overall survival ≥ 2 years. For molecular genetics associated with overall survival (OS) ≥ 2 years such as *ASXL1* (**A**), *DNMT3A* (**B**), *SF3B1* (**C**), *U2AF1* (**D**), *TP53* (**E**), higher variant allele fraction (VAF) was associated with decreased rates of 2-year OS. For biallelic *CEBPA* mutations and *CEBPA*-bZIP, VAF was not available for analysis.

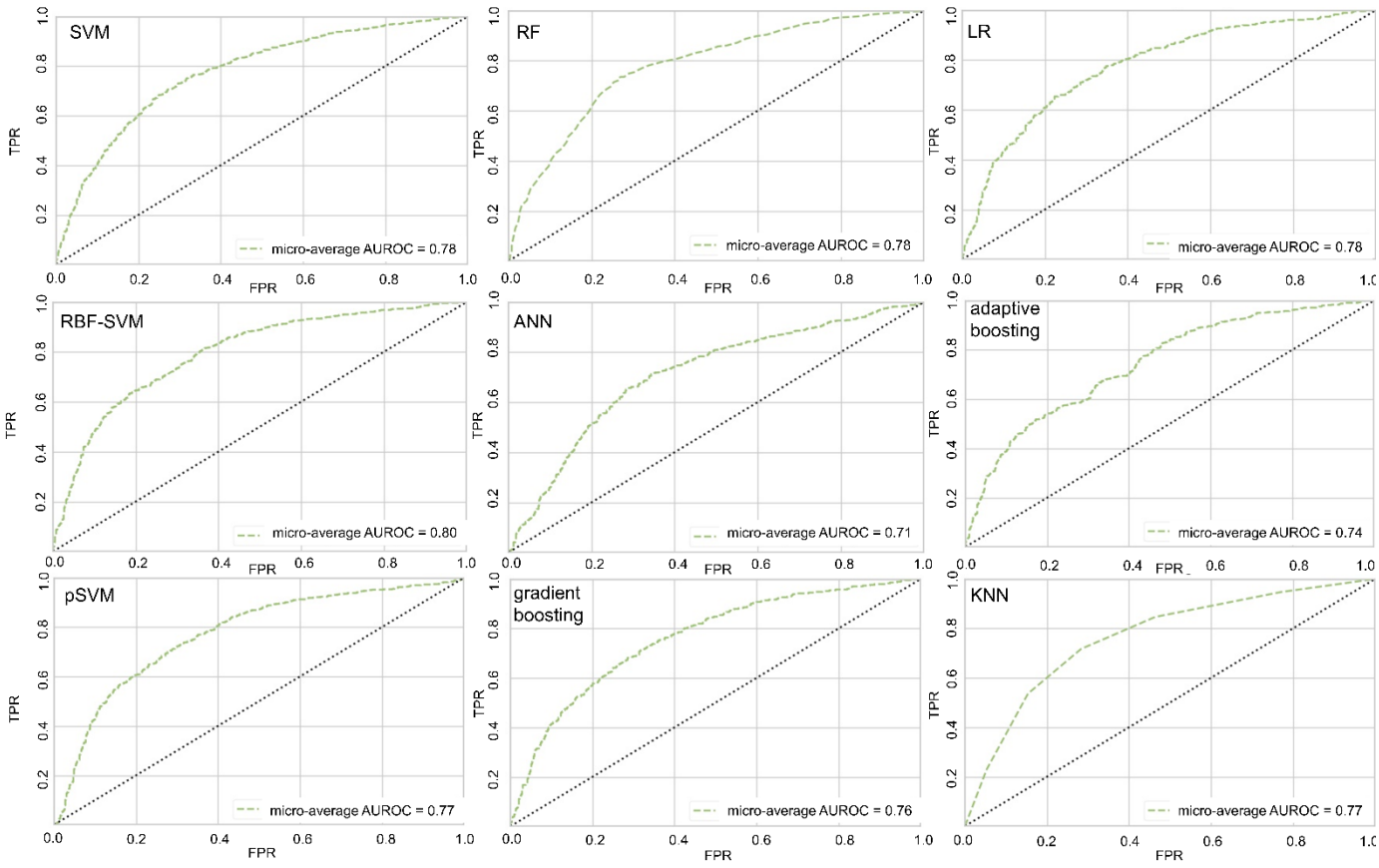

Figure S6 Performance of pre-trained machine learning models for prediction of CR/CRi on external data. The previously trained machine learning models were tested on external multi-center data encompassing 664 AML patients from the bioregistry of the AML Cooperative Group. ANN – artificial neural net; CR: complete remission; CRi: complete remission with incomplete hematologic recovery; FPR – false positive rate; KNN – k nearest neighbor; LR – logistic regression; pSVM – polynomial support vector machine; RBF-SVM – radial basis kernel function support vector machine; RF – random forest; SVM – (linear) support vector machine; TPR – true positive rate.

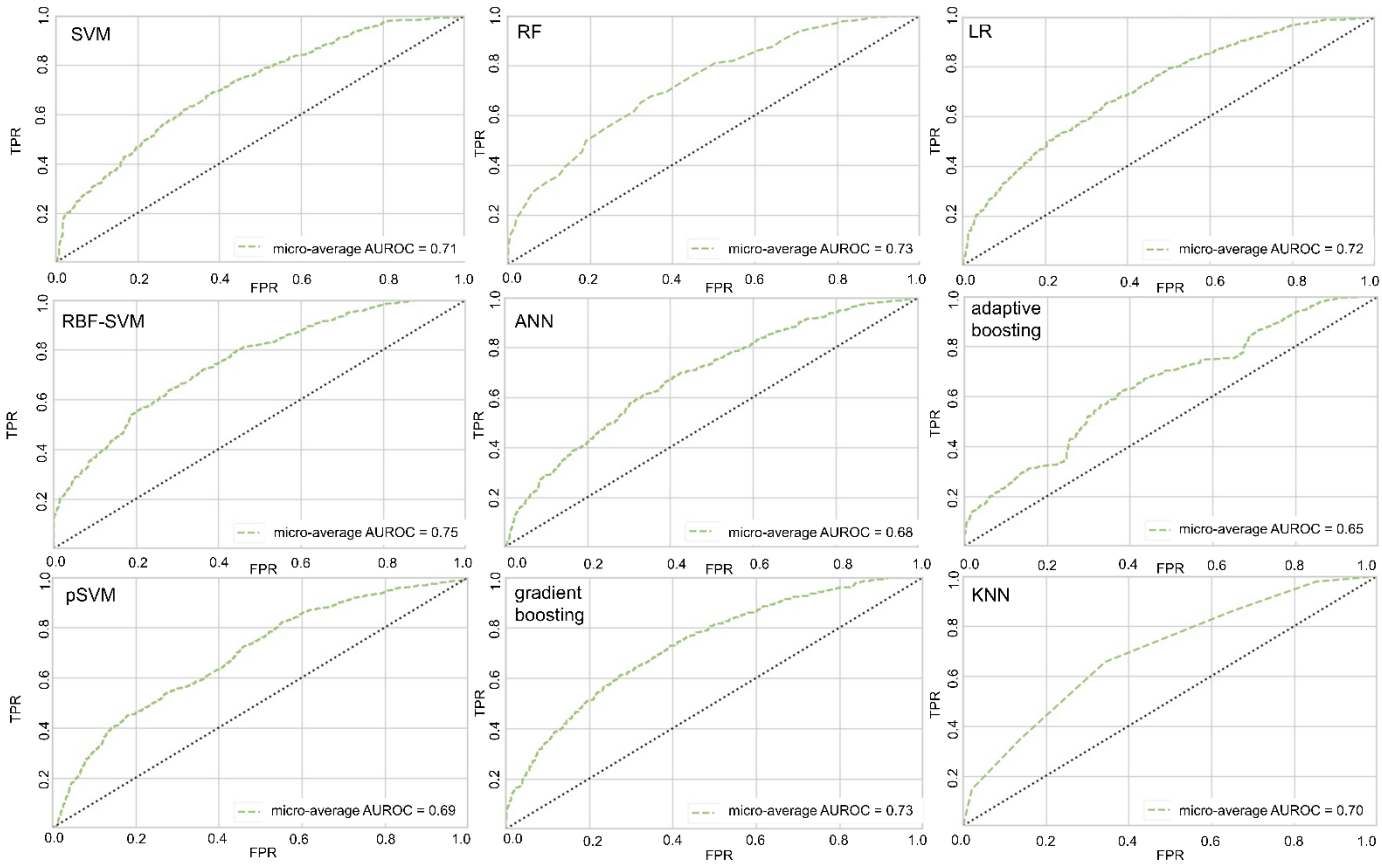

Figure S7 Performance of pre-trained machine learning models for prediction of 2-year overall survival on external data. The previously trained machine learning models were tested on external multi-center data encompassing 664 AML patients from the bioregistry of the AML Cooperative Group. ANN – artificial neural net; CR: complete remission; CRi: complete remission with incomplete hematologic recovery; FPR – false positive rate; KNN – k nearest neighbor; LR – logistic regression; pSVM – polynomial support vector machine; RBF-SVM – radial basis kernel function support vector machine; RF – random forest; SVM – (linear) support vector machine; TPR – true positive rate.