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

Abstract

Achievement of complete remission signifies a crucial milestone in the therapy of acute myeloid leukemia (AML) while refractory disease is associated with dismal outcomes. Hence, accurately identifying patients at risk is essential to tailor treatment concepts individually to disease biology. We used nine machine learning (ML) models to predict complete remission and 2-year overall survival in a large multicenter cohort of 1,383 AML patients who received intensive induction therapy. Clinical, laboratory, cytogenetic and molecular genetic data were incorporated and our results were validated on an external multicenter cohort. Our ML models autonomously selected predictive features including established markers of favorable or adverse risk as well as identifying markers of so-far controversial relevance. *De novo* AML, extramedullary AML, double-mutated *CEBPA*, mutations of *CEBPA-bZIP*, *NPM1*, *FLT3*-ITD, *ASXL1*, *RUNX1*, *SF3B1*, *IKZF1*, *TP53*, and *U2AF1*, t(8;21), inv(16)/t(16;16), del(5)/del(5q), del(17)/del(17p), normal or complex karyotypes, age and hemoglobin concentration at initial diagnosis were statistically significant markers predictive of complete remission, while t(8;21), del(5)/del(5q), inv(16)/t(16;16), del(17)/del(17p), double-mutated *CEBPA*, *CEBPA-bZIP*, *NPM1*, *FLT3-ITD*, *DNMT3A*, *SF3B1*, *U2AF1*, and *TP53* mutations, age, white blood cell count, peripheral blast count, serum lactate dehydrogenase level and hemoglobin concentration at initial diagnosis as well as extramedullary manifestations were predictive for 2-year overall survival. For prediction of complete remission and 2-year overall survival areas under the receiver operating characteristic curves ranged between 0.77–0.86 and between 0.63–0.74, respectively in our test set, and between 0.71–0.80 and 0.65–0.75 in the external validation cohort. We demonstrated the feasibility of ML for risk stratification in AML as a model disease for hematologic neoplasms, using a scalable and reusable ML framework. Our study illustrates the clinical applicability of ML as a decision support system in hematology.

Introduction

Acute myeloid leukemia (AML) is the most common form of acute leukemia in adults and its incidence has been increasing in the past decades. The long-term survival rate of AML patients in the overall patient population is poor.[1] Achievement of complete remission (CR) or complete remission with incomplete hematologic recovery (CRi) signifies a crucial milestone in AML therapy as it is associated with significantly improved patient outcome.[2] For intermediate- and high-risk patients with good performance status, allogeneic hematopoietic stem cell transplantation in first CR is a curative option.[3] However, refractory disease is associated with dismal overall survival (OS) rates, and relapse and death are frequent in patients with primary refractory disease even after allogeneic hematopoietic stem cell transplantation.[4] Therefore, efforts have been made to establish predictive markers in order to identify patients at risk of primary treatment failure and predict reduced OS after intensive induction therapy. Potential predictors include patient age,[5] high-risk cytogenetics such as complex karyotypes (≥3 abnormalities),[6] and molecular genetics.[7] However, most recent studies have been based on hypothesis-driven models that require *a priori* a hypothesized connection between selected variables to be tested on the given data.[8] Machine learning (ML) is a branch of computer science that can process large data sets for a plethora of purposes.[9] The underlying mechanism does not necessarily begin with a manually drafted hypothesis model. Rather, ML can detect patterns in pre-processed data and derive abstract information, predictions and similarities.[10] Their translation to AML risk assessment has shown the potential for refined prognostic indices and unveiled novel insights into disease biology.[11]

In this study, we retrospectively analyzed a large cohort of 1,383 newly diagnosed and intensively treated AML patients according to their clinical characteristics and molecular genetics. We evaluated nine different ML models to predict achievement of CR as well as 2-year OS rate, assessed features that were automatically identified by the ML models according to their predictive value and validated our results in an external cohort of 664 AML patients.

Methods

Data set

We retrospectively identified 1,383 patients who had been diagnosed and treated in previously reported multicenter trials (AML96,[12] AML2003,[13] AML60+,[14] and SORAML[15]) or were enrolled in the multicenter German Study Alliance Leukemia (SAL) AML registry (NCT03188874) encompassing 59 centers specialized in the treatment of hematologic malignancies. A short summary of individual trial durations and protocols is provided in the *Online Supplementary Material* (*Online Supplementary Table S1*). Eligibility criteria were newly diagnosed AML according to World Health Organization (WHO) criteria,[16] age ≥18 years, potentially curative treatment with intensive therapy regimens and available diagnostic biomaterial. Patients with acute promyelocytic leukemia were excluded. All mentioned studies were previously approved by the Institutional Review Board of the Technical University Dresden. All participants gave their written informed consent according to the Declaration of Helsinki. AML status was defined as *de novo* (in patients with no prior hematologic malignancy), secondary (in patients with prior myeloid entities such as myelodysplastic syndromes) and treatment-related (in patients previously exposed to radiotherapy and/or chemotherapy). CR and CRi were defined according to the European LeukemiaNet (ELN) 2017 recommendations.[17] Death was defined as death from any cause. Of the 1,383 patients studied, 91 (6.56%) died within 30 days of initial diagnosis. All patients were included in the analysis for both CR and 2-year OS. We used 2-year OS because the data set was balanced for this cut-off time with 610/1,383 (44.11%) of patients surviving 2 years or longer, which supports training of a binary classifier. Pre-treatment bone marrow or peripheral blood samples from all patients were screened using next-generation sequencing with the Illumina TruSight Myeloid Sequencing Panel covering 54 genes (*Online Supplementary Table S2*) that are associated with myeloid neoplasms, as described in detail recently.[18] A 5% variant allele frequency mutation calling cut-off was used. An external validation cohort was obtained from the AML Cooperative Group (AMLCG) encompassing 664 newly diagnosed AML patients enrolled in clinical trials (AMLCG-1999 and AMLCG-2008)[19] to validate the trained algorithms. For this validation cohort, the same eligibility and exclusion criteria were applied as described above. This study was performed in conformity with Standards for Reporting Diagnostic accuracy studies (STARD) (*Online Supplementary Table S3*).

Data curation and machine learning pipeline

For the selection of predictive features and subsequent binary decisions for CR and 2-year OS prediction, a multi-stage ML pipeline was developed for this study (Figure 1). Data from the above-mentioned clinical trials and the SAL registry were collected and 212 multimodal variables (clinical data, laboratory parameters as well as molecular and cytogenetic data) became available (see *Online Supplementary Table S4* for a full list of variables used in the model). Features were selected according to their support by five-feature selection algorithms: linear

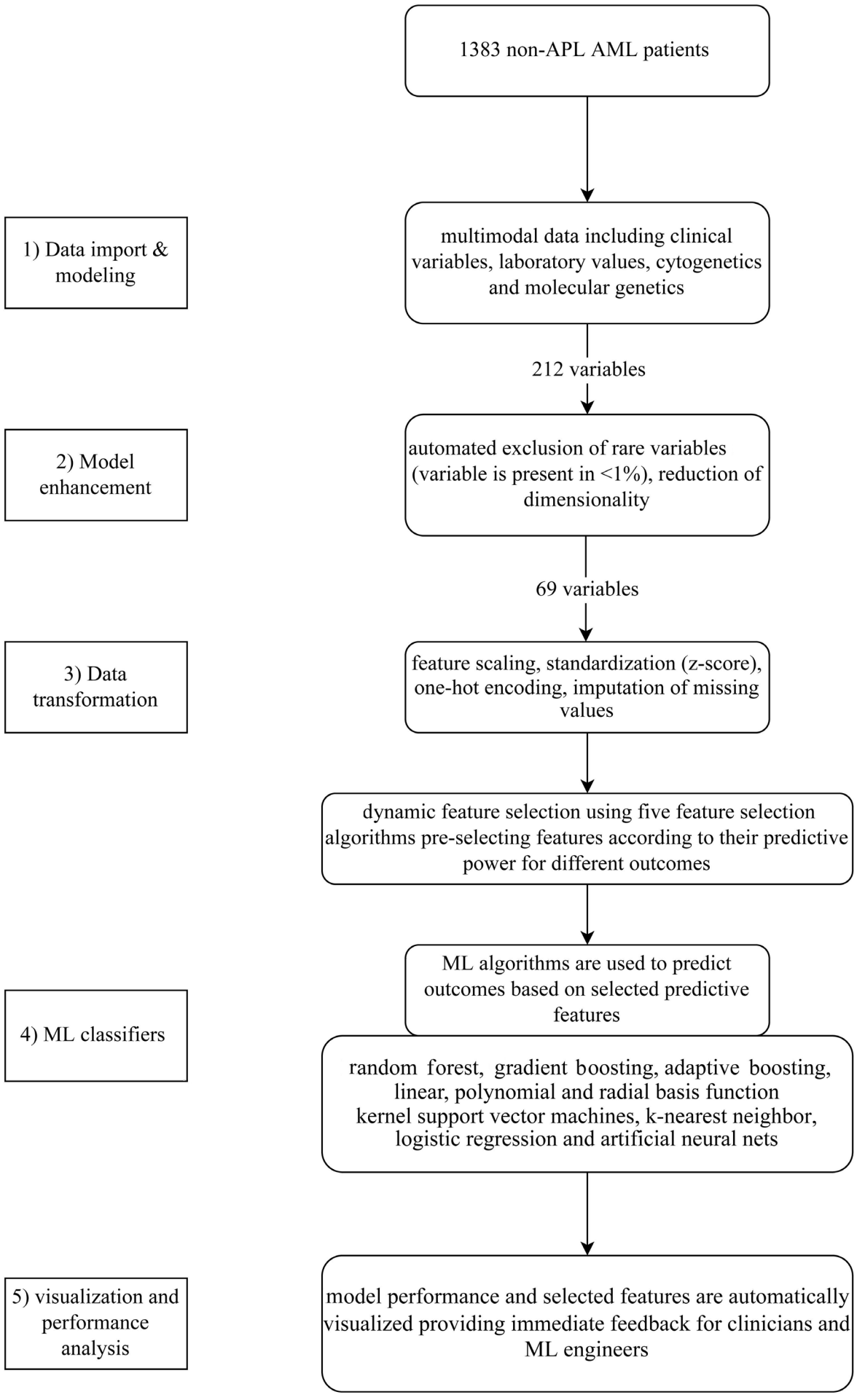


Figure 1. Iterative workflow of the machine learning pipeline. For the purpose of this study, 1,383 patients with acute myeloid leukemia from previous multicenter clinical trials and the German Study Alliance Leukemia bioregistry were analyzed. Multimodal clinical, laboratory, cytogenetic and molecular genetic data (1) were available. To remove redundancies and reduce dimensionality, rare features were excluded (2). Data were transformed, scaled and standardized and missing values were imputed (3). Dynamic feature selection was used to identify predictive parameters which were then included for analysis by nine supervised machine learning classifiers (4). Individual model performance and selected features were subsequently put out by the pipeline for interpretation (5). APL: acute promyelocytic leukemia; AML: acute myeloid leukemia.

correlation, chi-square test, recursive feature elimination, lasso regularization and random forest ranking. To be included in a ML model, a variable had to pass a predetermined threshold of overall predictive power determined by summing the normalized scores of these five-feature selection algorithms. Features below the threshold were automatically excluded from the ML models for the respective iteration. In that way, relevant attributes were selected and dimensionality was reduced by excluding sparse features (cut-off 1%). After automated feature selection, binary decision models of the following types were trained: random forest, gradient boosting, adaptive boosting, linear, polynomial and radial basis function kernel (RBF), support vector machines (SVM), k-nearest neighbor, logistic regression, and artificial neural nets using a 9:1 training-to-test split. All test data were strictly withheld from the training stage in order to avoid information leakage and overfitting. The

best performing models were optimized in a subsequent hyperparameter-optimization step. A more detailed explanation of the ML pipeline is given in the *Online Supplementary Material*.

Performance evaluation and statistical analysis

To analyze the performance of the ML models we used F1-score, precision and recall as well as precision-recall-curves as standard ML performance metrics, as well as receiver operating characteristics (ROC) with the area under the curve (AUC). Precision (positive predictive value) is the fraction of true positives among all positive predictions while recall (sensitivity) is the fraction of all positive predictions among all true positives and F1-score is the harmonized mean of precision and recall. To account for the imbalance of the data set, micro-averaging AUROC was calculated as it computes the total number of cumulative true positives, true negatives, false positives and false negatives globally instead of calculating metrics for each class independently and then averaging them (macro-averaging) which may lead to inaccurate metrics for imbalanced data sets. Additional statistical analysis and visualizations were performed using STATA BE 16.0 and R 3.6.3. Odds ratios and 95% confidence intervals for the binary decision of achieving or failing to achieve CR as well as surviving 2 years or longer were obtained using logistic regression. Statistical significance was determined using a significance level α of 0.05.

Results

We utilized nine ML models to predict CR and 2-year OS in a large data set of 1,383 newly diagnosed and intensively treated AML patients with a median age of 54 years (interquartile range, 43–64). A total of 1,008 patients (72.9%) achieved CR/CRi with induction therapy, while 375 (27.1%) failed to achieve CR/CRi. Of the 1,008 patients who achieved CR/Cri, 755 (74.9 %) did so after two courses of induction therapy, while 253 (25.1 %) received only one course of induction therapy. The median OS was 17.1 months and 44,1% of patients survived 2 years or longer after initial diagnosis. The patients' baseline characteristics are summarized in Table 1. Detailed information on the characteristics of patients from the different trials of both the internal training and testing cohort as well as the external validation cohort are summarized in *Online Supplementary Table S5*.

Prediction of complete remission

For CR/CRi, F1-scores ranged between 0.72 and 0.75 while AUROC ranged between 0.77 and 0.86 (Figure 2). Random forest (F1: 0.75; AUROC: 0.86), logistic regression (F1: 0.75; AUROC: 0.84) and artificial neural nets (F1: 0.73; AUROC: 0.77) were selected for hyperparameter tuning. Random forest and logistic regression converged over 1,000 iterations (*Online Supplementary Figure S1*). Hyperparameter tuning did not improve the F1 of logistic regression, but random forest achieved an improved final F1 of 0.78. Artificial neural nets did not converge over 1,000 iterations and the F1 of artificial neural nets did not improve, likely due to the requirement of a much larger sample size for deep learning in general. Features for CR/CRi prediction were selected automatically using five-feature selection algorithms that included or rejected features based on an importance score with a predefined threshold. We found the optimum performance was achieved when a summed support threshold of 0.5 was used as a cut-off for inclusion or exclusion of features. Features that were present in less than 1% of patients in the cohort were automatically excluded. Using this method, our algorithms selected 27 features for CR/CRi prediction that were uniformly used in all nine classification models. Patient age at first diagnosis was the most important feature according to our feature selection algorithm. Genetic aberrations included in our model were found in *TP53* (n=102, 7.38%), *U2AF1* (n=36, 2.60%), *NPM1* (n=466, 33.69%), *FLT3-ITD* (n=280, 20.25%), *IKZF1* (n=36, 2.6%), *CEBPA* (double-mutated n=91, 6.58% and bZIP n=30, 2.17%), *ASXL1* (n=124, 8.97%), *RUNX1* (n=134, 9.69%), *IDH1* (n=122, 8.82%), *PTPN11* (n=100, 7.23%), *SF3B1* (n=41, 2.96%), as well as t(8;21) (n=52, 3.76%), inv(16) or t(16;16) (n=76, 5.50%), del(5) or del(5q) (n=85, 6.15%), del(17) or del(17p) (n=34, 2.50%), complex karyotype (≥3 aberrations, n=152, 10.99%) or normal karyotype (no aberrations, n=707, 51.12%). These genetic features differed substantially between patients achieving CR/CRi (Figure 3A) or failing to achieve CR/CRi (Figure 3B). Clinical and laboratory parameters that were selected by our algorithm were lactate dehydrogenase concentration, white blood cell count, bone marrow blast count, peripheral blood blast count, platelet count and hemoglobin concentration at first diagnosis as well as *de novo* manifestation of AML and presence or absence of extramedullary disease. Individual feature support calculated by the five-feature selection algorithms is shown in Figure 4A. For these features we subsequently calculated univariate odds ratios to further quantify their predictive capacity for CR. At a significance level of 0.05, we found *de novo* status of AML, higher hemoglobin concentration at initial diagnosis, normal karyotype, t(8;21), inv(16) or t(16;16), double-mutated *CEBPA* or mutations in the bZIP domain of *CEBPA*, and mutations in *NPM1* and *FLT3*-ITD to be associated with significantly higher odds of achieving CR (Figure 4B). Notably, the effect of mutations in *FLT3*-ITD was confined to patients with an *FLT3*-ITD ratio <0.5 and concurrent *NPM1* mutations (odds ratio [OR]=2.01, 95% confidence interval [95% CI]: 1.09-3.71, *P*=0.024) while

Table 1. Patients' baseline characteristics.

Variables	Training/testing (SAL)	External validation (AMLCG)
N. of patients	1383	664
Age, median (IQR), in years	54 (43-64)	57 (44-66)
Sex, N (%)		
Female	661 (48)	328 (49)
Male	722 (52)	336 (51)
AML status, N (%)		
De novo	1180 (86.4)	570 (85.8)
Secondary	146 (10.7)	59 (8.9)
Therapy-associated	40 (3.0)	35 (5.3)
French-American-British classification, N (%)		
M0	49 (3.7)	35 (5.4)
M1	326 (24.6)	157 (23.6)
M2	458 (34.6)	178 (26.8)
M3	0	0
M4	248 (18.7)	163 (24.5)
M5	191 (14.4)	83 (12.5)
M6	46 (3.5)	19 (2.9)
M7	6 (0.5)	3 (0.5)
European LeukemiaNet 2017 category, N (%)		
Favorable	518 (37.8)	231 (34.8)
Intermediate	510 (37.2)	166 (25.0)
Adverse	247 (13.0)	250 (37.7)
Complex karyotype (≥3 abnormalities)	154 (11.9)	75 (11.3%)
Extramedullary disease, N (%)	201 (14.5)	16 (5.9)
White blood cell count, median (IQR), x10^9/L	20.4 (4.8-56.4)	23.8 (6.4-60.3)
Hemoglobin, median (IQR) in mmol/L	5.9 (5.0-7.0)	5.6 (5.0-6.3)
Platelet count, median (IQR) x10^9/L	52 (27-95)	53 (30-102)
Lactate dehydrogenase, median (IQR) in U/L	453 (288-821)	466 (291-787)
Bone marrow blasts, median (IQR) in %	63 (45-79)	80 (58-90)
Peripheral blood blasts, median (IQR) in %	41 (12-74)	23 (4.5-67)
Achieved CR after induction therapy, N (%)	1008 (72.9)	445 (67.0)
Median OS, in months	17.1	17.3
Overall survival ≥ 2 years, N (%)	610 (44.1)	290 (43.7)

For a division of the internal cohort by clinical trials, see *Online Supplementary Table S5.* SAL: German Study Alliance Leukemia registry; AMLCG: AML Cooperative Group; n/N: number; IQR: interquartile range; AML: acute myeloid leukemia; CR: complete remission.

patients who harbored mutated *FLT3*-ITD with a ratio ≥0.5 and concurrent *NPM1* mutations showed less favorable CR rates (OR=0.51, 95% CI: 0.28-0.94; *P*=0.03). Higher age at initial diagnosis, extramedullary manifestations, complex karyotype, del(5) or del(5q), del(17) or del(17p) as well as mutations in *ASXL1*, *SF3B1*, *RUNX1*, *IKZF1*, *TP53* and *U2AF1* were associated with significantly lower odds of achieving CR with intensive induction therapy (Figure 4B). *IKZF1*, *SF3B1*, and *U2AF1* mutations have been reported to be associated with secondary AML.[20,21] In a multivariable model adjusted for *de novo* and secondary AML, we found *IKZF1* (OR=0.39, 95% CI: 0.20-0.76; *P*=0.006), *SF3B1* (OR=0.49, 95% CI: 0.26-0.94; *P*=0.031) and *U2AF1* (OR=0.17, 95% CI: 0.08-0.35; *P*<0.001) to be independently associated with lower odds of achieving CR. In a multivariable model adjusting for double-mutated *CEBPA*, mutations of the bZIP domain of *CEBPA* were still significantly associated with increased odds of achieving CR (OR=5.95, 95% CI: 1.90-18.66; *P*=0.002). Every 1-year increase in age was associated with a 5.73% decrease in the odds of achieving CR (*Online Supplementary Figure S3A*) and every one mmol/L increase in hemoglobin at initial diagnosis (until normal values were reached) was associated with a 13.15% increase in the odds of achieving CR (*Online Supplementary Figure S3B*). For molecular genetics associated with CR such as *ASXL1*, *IKZF1*, *SF3B1*, *U2AF1* and *TP53* (*Online Supplementary Figure Table S3C-G*), higher variant allele frequency was associated with decreased odds for CR. For biallelic *CEBPA* mutations and *CEBPA*-bZIP, variant allele frequency was not available for analysis. For the remaining selected features – peripheral blood blast count, bone marrow blast count, lactate dehydrogenase level, platelet count and white blood cell count at initial diagnosis as well as mutations in *PTPN11*, and *IDH1* – no statistically significant associations with achievement of CR were found (Figure 4B).

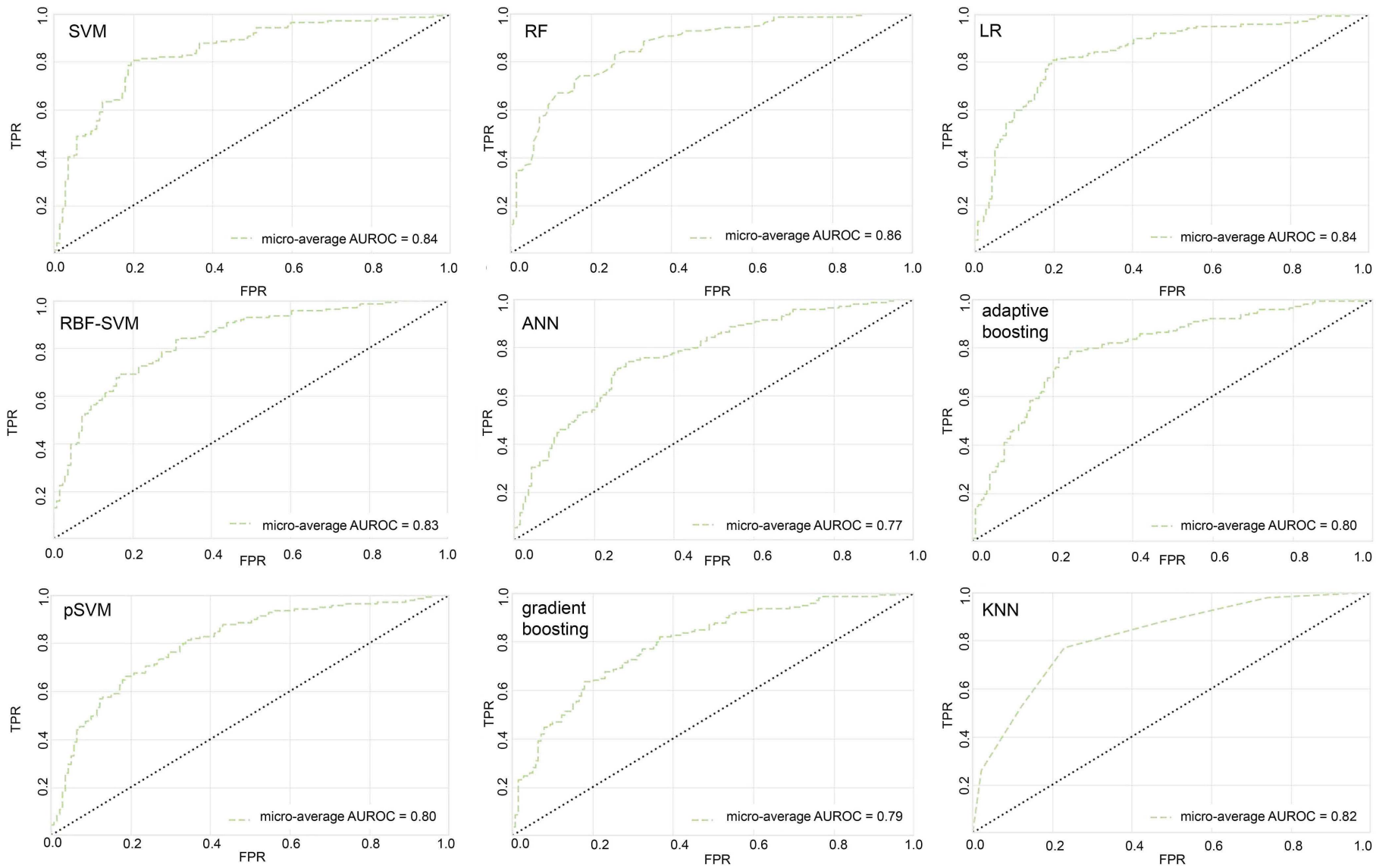


Figure 2. Performance of the machine learning algorithms for prediction of complete remission or complete remission with incomplete hematologic recovery. Nine machine learning algorithms were trained and tested on 1,383 patients for whom multimodal clinical, laboratory and cytogenetic as well as molecular genetic data were available (train-test split 9:1, 10-fold cross-validation). Micro-average area under the receiver operating characteristic curve (AUROC) was used to evaluate performance of the imbalanced data set regarding achievement or failure of complete remission after intensive induction therapy. 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.

Prediction of 2-year overall survival

Analogous to CR/CRi prediction, the ML pipeline was used to predict 2-year overall survival. For OS, F1-scores ranged between 0.60 and 0.70 (Table 2) while AUROC ranged between 0.63 and 0.74 (Figure 5). Again, random forest (F1: 0.67; AUROC: 0.73), logistic regression (F1: 0.70; AUROC: 0.74) and artificial neural nets (F1: 0.63; AUROC: 0.70) were selected for hyperparameter tuning. Artificial neural nets again did not converge and F1 did not improve over 1,000 iterations. Random forest and logistic regression both converged over 1,000 iterations (*Online Supplementary Figure S2*). While F1 did not improve for logistic regression, random forest showed an increased F1 of 0.68 after hyperparameter tuning. The feature selection algorithm chose the 25 most important features based on the same threshold that was previously used for CR prediction (Figure 6A). Again, the most important feature selected by the algorithms was patient age at initial diagnosis. Selected genetic features encompassed mutations in *TP53*, *NPM1*, double-mutated *CEBPA*, mutations in the bZIP domain of *CEBPA*, *U2AF1*, *SF3B1*, *ASXL1*, *FLT3*-ITD and -TKD (n=62, 4.48%), *WT1* (n=102, 7.38%), *PTPN11*, *KRAS* (n=79, 5.71%), and *DNMT3A* (n=396, 28.63%), t(8;21), del(5) or del(5q), inv(16) or t(16;16), del(17) or del(17p), which again differed between patients who survived 2 years or longer (Figure 3C) or died within 2 years after initial diagnosis (Figure 3D). Selected clinical and laboratory features were hemoglobin concentration at initial diagnosis, white blood cell count, peripheral blood blast count, bone marrow blast count, platelet count and lactate dehydrogenase level at initial diagnosis, as well as the presence of extramedullary manifestations. Univariate logistic regression showed significantly increased odds of surviving 2 years or longer for t(8;21), inv(16) or t(16;16), double-mutated *CEBPA*, mutations in the bZIP domain of *CEBPA*, *FLT3*-ITD with low (<0.5) variant allele ratio (irrespective of *NPM1* status), mutations of *NPM1* as well as higher hemoglobin at initial diagnosis (Figure 6B).

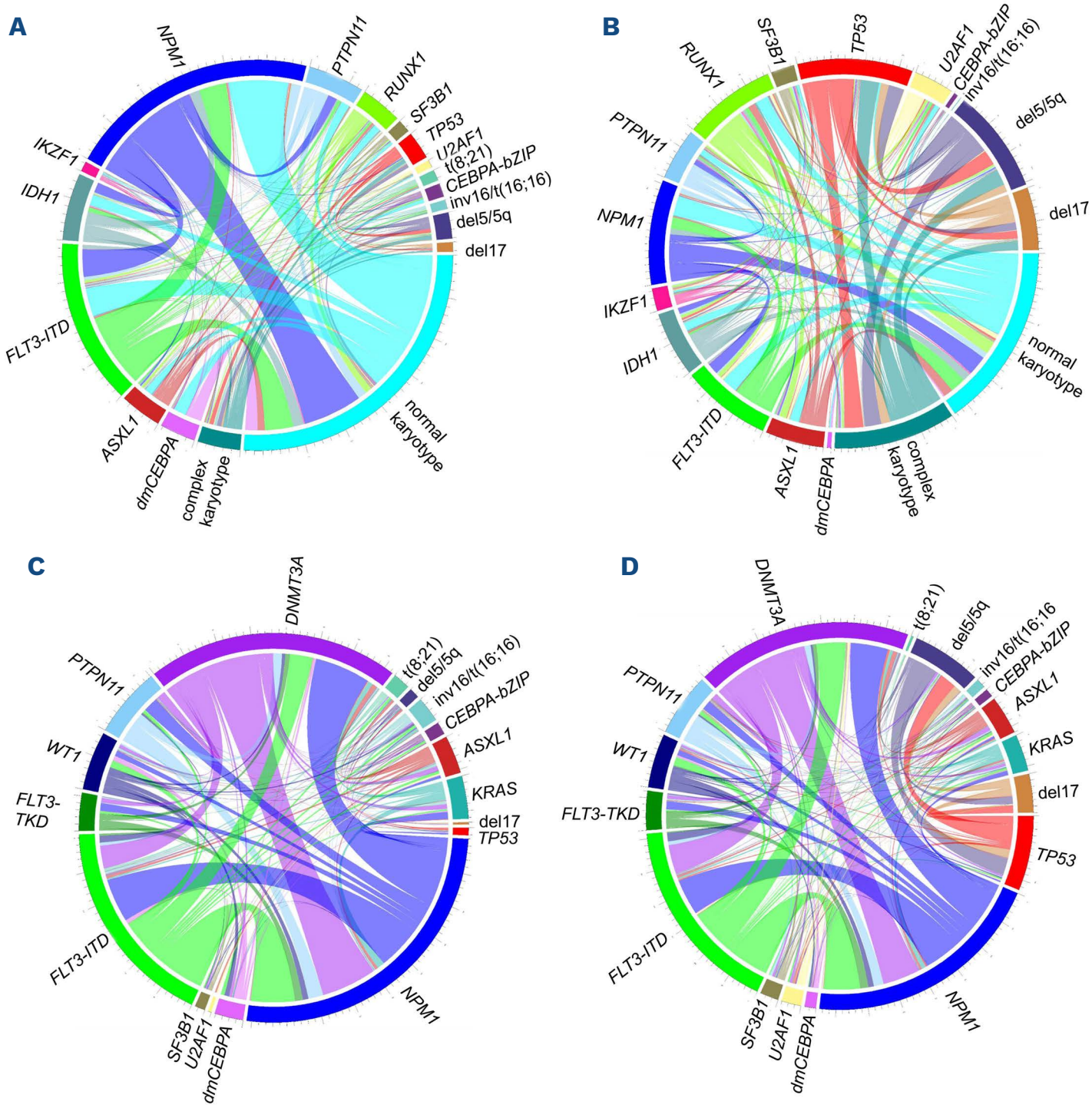


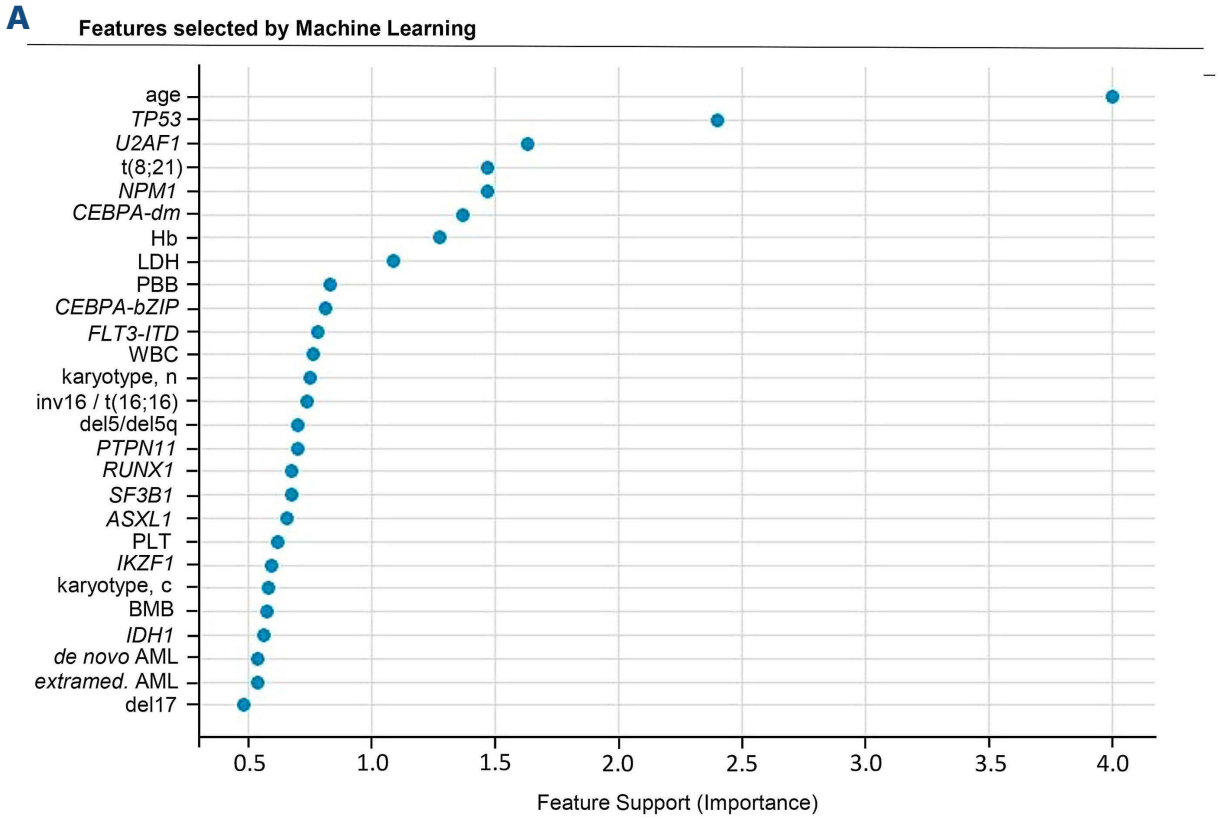
Figure 3. Mutational spectrum of aberrations selected by machine learning for prediction of complete remission and overall survival. Patients who achieved complete remission (CR)/complete recovery with incomplete hematologic recovery (CRi) after intensive induction therapy (A) showed different molecular patterns regarding molecular features selected by machine learning than patients who failed to achieve CR (B). The mutational spectrum of the cohort of patients who achieved CR largely comprised normal karyotypes (no aberrations) as well as mutations of *NPM1* and *FLT3*-ITD. In the cohort of patients failing to achieve CR the rate of complex karyotypes (≥3 aberrations), del17, del5 or del5p, as well as mutations in *TP53*, *ASXL1*, *RUNX1*, *U2AF1*, *SF3B1* and *IKZF1* was higher than that in patients who achieved CR. Patients who survived longer than 24 months (C) were less likely to harbor del17, del5 or del5q, or have mutations in *TP53*, *SF3B1*, *ASXL1* and *U2AF1* than patients who died within 24 months after initial diagnosis (D).

Significantly lower odds were found for higher age at initial diagnosis, higher white blood cell count, lactate dehydrogenase, and peripheral blood blast count, presence of extramedullary manifestations as well as del(17) or del(17p), del(5) or del(5q) and mutations of *DNMT3A*, *FLT3*-ITD with high (≥0.5) variant allele ratio (again irrespective of *NPM1* status), *SF3B1*, *U2AF1* and *TP53* (Figure 6B). In multivariable analysis including AML status (*de novo* or secondary AML), mutations in *SF3B1* (OR=0.32, 95% CI: 0.14-0.69; *P*=0.004) and *U2AF1* (OR=0.16, 95% CI: 0.06-0.46; *P*=0.001) were independent markers of decreased odds of surviving 2 years after initial diagnosis. In a multivariable model adjusting for double-mutated *CEBPA*, mutations of the bZIP domain of *CEBPA* were still significantly associated with increased odds of 2-year OS (OR=2.36, 95% CI: 1.01-5.23; *P*=0.036). For continuous variables, every 1-year increase in age was associated with a 4.27% decrease in the odds of surviving 2 years or longer after initial diagnosis (*Online Supplementary Figure S4A*). For hemoglobin, every one mmol/L increase until normal values was associated with a 14.08% increase of the odds (*Online Supplementary Figure S4B*). Increases in white blood cell count, peripheral blood blast count and lactate dehydrogenase concentration were also associ-

ated with decreases in the odds of survival, however effect sizes were smaller than those for age or hemoglobin (*Online Supplementary Figure S4C-E*). For molecular genetics associated with 2-year OS, such as *ASXL1*, *DNMT3A*, *SF3B1*, *U2AF1*, and *TP53* mutations, higher variant allele frequency was associated with decreased rates of 2-year OS (*Online Supplementary Figure S5*). For biallelic *CEBPA* mutations and *CEBPA*-bZIP, variant allele frequency was not available for analysis.

External validation

We obtained an external independent cohort of 664 pre-



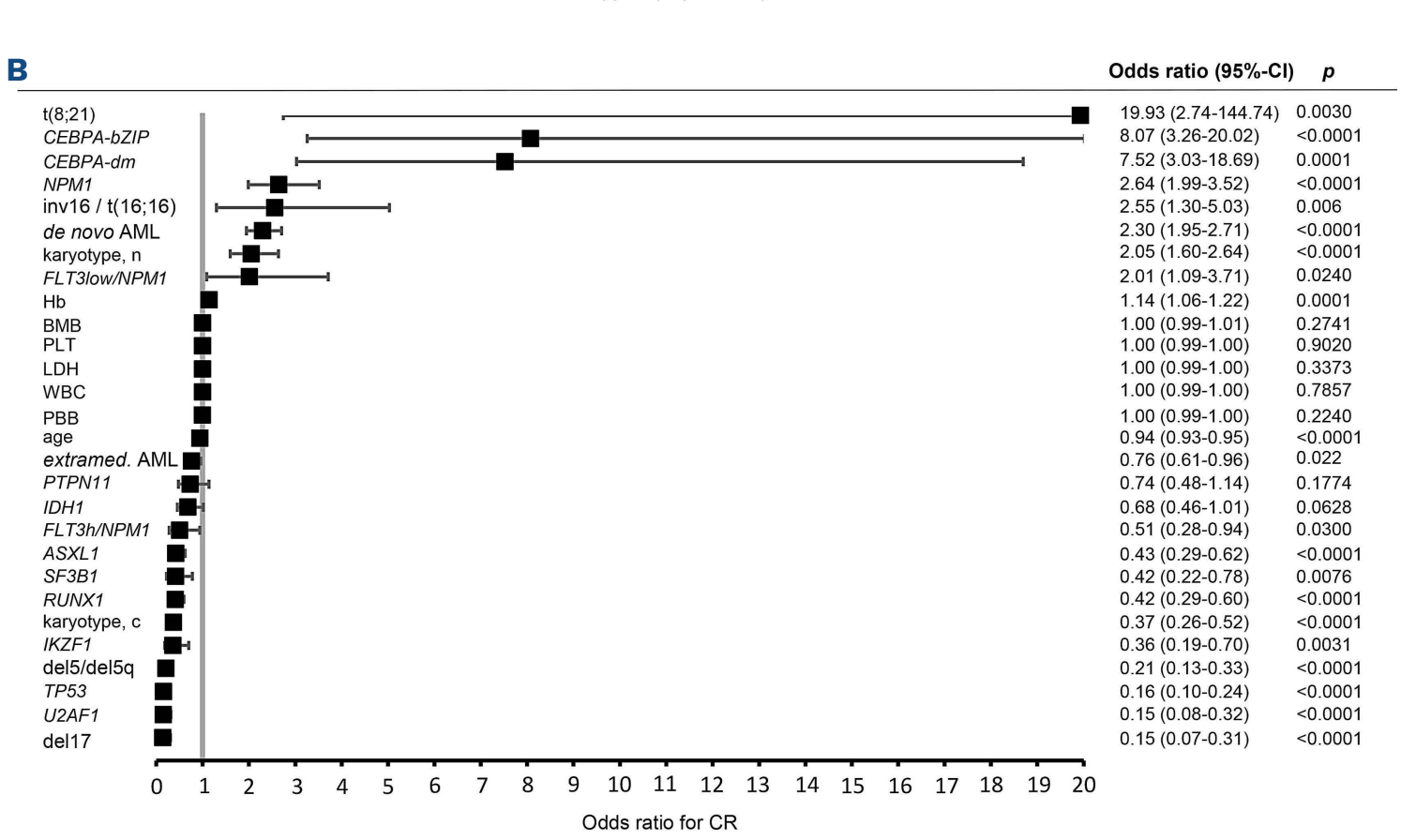


Figure 4. Feature selection for prediction of complete remission. (A) Five-feature selection metrics (linear correlation, chi-square test, recursive feature elimination, lasso regularization and random forest ranking) were implemented to select patient features for the classification algorithms (Figure 1) in order to predict complete remission (CR) after intensive induction therapy. Based on a continuous feature support metric to aggregate to single metrics mentioned above with a predefined cut-off of 0.5 (determined by optimal classification performance), 27 features were automatically selected to be included for prediction of CR. (B) For each of these features predicted by machine learning, odds ratios and 95% confidence intervals (95% CI) were calculated. BMB: bone marrow blast count; FLT3h/low: *FLT3*-ITD ratio, h=high>0.5; Hb: hemoglobin; karyotype, c: complex aberrant karyotype (≥3 aberrations); karyotype, n: normal karyotype (no aberrations); LDH: lactate dehydrogenase; PBB: peripheral blood blast count; PLT: platelet count; WBC: white blood cell count.

Table 2. Performance metrics for prediction of complete remission/complete remission with incomplete hematologic recovery and 2-year overall survival by different machine learning models.

Prediction of CR/CRi after intensive induction therapy								
ML model	**F1-score**		**Precision**		**Recall**		**AUROC**	
	Test	**Val.**	**Test.**	**Val.**	**Test**	**Val.**	**Test**	**Val.**
Random forest	0.75	0.76	0.77	0.77	0.78	0.78	0.86	0.78
Linear SVM	0.75	0.76	0.76	0.77	0.77	0.78	0.84	0.78
Logistic regression	0.75	0.76	0.76	0.77	0.77	0.77	0.84	0.78
Adaptive boosting	0.75	0.75	0.75	0.75	0.76	0.77	0.80	0.74
Gradient boosting	0.74	0.74	0.74	0.74	0.75	0.76	0.79	0.76
Polynomial SVM	0.73	0.72	0.76	0.76	0.77	0.77	0.80	0.77
Artificial neural net	0.73	0.73	0.73	0.73	0.74	0.73	0.77	0.71
RBF-SVM	0.72	0.74	0.75	0.76	0.76	0.77	0.83	0.80
k nearest neighbor	0.72	0.72	0.73	0.72	0.75	0.75	0.82	0.77
Prediction of OS ≥2 years								
Random forest	0.67	0.68	0.67	0.68	0.67	0.68	0.73	0.73
linear SVM	0.70	0.69	0.70	0.69	0.70	0.69	0.74	0.71
Logistic regression	0.70	0.69	0.70	0.69	0.70	0.69	0.74	0.72
Adaptive boosting	0.66	0.66	0.67	0.67	0.66	0.66	0.74	0.65
Gradient boosting	0.65	0.65	0.65	0.65	0.65	0.65	0.72	0.73
RBF-SVM	0.67	0.67	0.67	0.67	0.67	0.67	0.72	0.75
Artificial neural net	0.63	0.63	0.63	0.63	0.63	0.63	0.70	0.68
k nearest neighbor	0.60	0.61	0.60	0.62	0.59	0.61	0.63	0.70
Polynomial SVM	0.60	0.58	0.61	0.60	0.61	0.60	0.70	0.69

Performance of the machine learning models was assessed using the F1-score, precision and recall as well as micro-average area under the receiver operating characteristic curve (see Figures 2 and 5 and *Online Supplementary Figures S6* and *S7*). A comparison between our internal test set (Test) and an external validation cohort (Val.) is shown. Precision (positive predictive value) is the fraction of true positives among all positive predictions. Recall (sensitivity) is the fraction of positive predictions among all true positives. F1-score is the harmonized mean of precision and recall. CR: complete remission; CRi: complete remission with incomplete hematologic recovery; ML: machine learning; AUROC: area under the receiver operating characteristics curve; SVM: support vector machine; RBF: radial basis function kernel. OS: overall survival.

viously untreated AML patients who received intensive induction chemotherapy on two randomized multicenter phase III trials of the German AML Cooperative Group (AMLCG) between 1999 and 2012[19] to validate our trained models for CR and 2-year OS prediction. Detailed patients' characteristics and genetic alterations available for the validation cohort are shown in Table 1 and *Online Supplementary Tables S4* and *S5*, respectively. Both previously trained prediction models including the above-mentioned prognostic variables for CR and 2-year OS prediction were tested on the validation cohort without re-training. It should be noted that not all prognostic variables included in the final prediction models for training and testing were available in the external validation cohort. Mutation status for *FLT3*-TKD and *IKZF1* was missing. For CR prediction, F1 ranged between 0.72 and 0.76 while AUROC ranged between 0.71 and 0.80 (*Online Supplementary Figure S6*). For prediction of 2-year OS, F1 ranged between 0.58 and 0.69 while AUROC ranged between 0.65 and 0.75 (*Online Supplementary Figure S7*). Table 2 provides details of the performance metrics in the internal test set and external validation cohort.

Discussion

Based on genetic and clinical data from a large multicenter cohort of patients we implemented ML models to derive prognostic parameters and subsequently predict CR and 2-year OS in AML patients who received intensive induction therapy. Our ML models were completely agnostic of any pre-existing models or risk scores such as

ELN 2017.[17] Nevertheless, among the selected features for both CR and OS we found many established markers of good or poor prognosis. Regarding mutational status, established markers for AML risk stratification[17] such as *TP53*, *ASXL1*, *RUNX1*, *FLT3*-ITD, *NPM1*, and double-mutated C*EBPA* were selected. Mutations of *TP53* are known to be associated with higher age, complex karyotypes and lower response rates to chemotherapy, yielding poor outcomes.[22,23] Accordingly, mutations of *RUNX1*[24] and *ASXL1*[25] have been reported to be associated with lower CR rates as well as poor survival and AML with mutated *RUNX1* is considered a provisional entity in the 2016 WHO classification.[26] In contrast, AML with mutations of *NPM1*[27–29] or AML with biallelic *CEBPA* mutations[30] were reported to be associated with improved outcomes and distinct co-mutational phenotypes, and also constitute distinct entities in the 2016 WHO classification.[26] The prognostic role of *FLT3*-ITD mutations largely depends on the allelic ratio and concurrent mutations of *NPM1*.[31,32] Additionally, in our CR model *U2AF1*, *IKZF1*, and *SF3B1* mutations were identified as predictive markers for decreased odds of achieving CR while mutations in *U2AF1*, *SF3B1*, as well as *DNMT3A* were also predictive for decreased 2-year OS. In a multivariable model adjusting for AML status (*de novo*/secondary AML) independent prognostic value was confirmed. Mutations of *U2AF1* and *SF3B1* affect RNA splicing and are frequent in myelodysplastic syndromes[33] while in AML they are more commonly found in secondary rather than *de novo* AML and previous studies reported poor outcomes.[34] *IKZF1* is a well-established marker of adverse risk in acute lymphoblastic leukemia,[35] however, its role in AML is still controversial. Previous studies have shown frequent co-mutational patterns in AML suggesting antecedent myeloproliferative neoplasms,[21,21] nevertheless their prognostic impact is unclear. In AML with mutated *DNMT3A*, prognostication is controversial: various studies found inferior survival, but these results have been questioned by other analyses that either found no differences in outcomes or improved survival.[36–38] Additionally, mutations of the bZIP domain of *CEBPA* were significantly associated with increased odds of achieving CR and 2-year OS irrespective

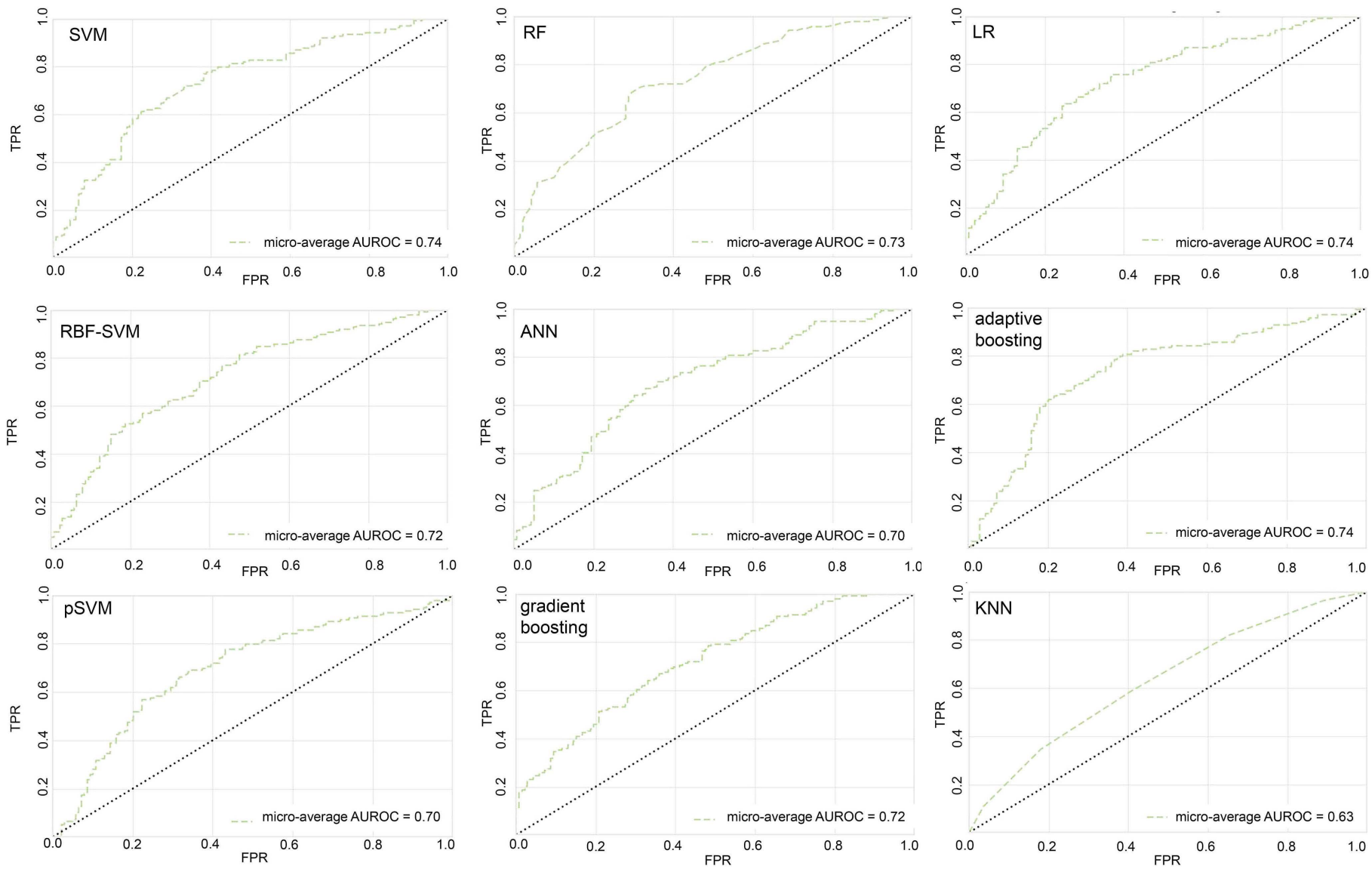


Figure 5. Performance of machine learning algorithms for prediction of overall survival ≥2 years. As for the prediction of complete remission (Figure 1), machine learning algorithms were also implemented for prediction of overall survival. Micro-average area under the receiver operating characteristic curve (AUROC) was used to evaluate performance. 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; OS: overall survival; 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.

of biallelic status in multivariable models which is in accordance with recent reports.[39,40] Regarding cytogenetic features, we initially one-hot encoded every cytogenetic aberration found in the entire cohort; however, to reduce data dimensionality[41] those that were present in less than 1% of the cohort were automatically excluded (as was the case for rare molecular genetic features). Cytogenetic features selected by our algorithm were inv(16)/t(16;16), t(8;21), del(5)/del(5q), and del(17) or del(17p), which are established markers for outcome prediction.[17] Strikingly, t(8;21) was associated with the largest increase in odds for both achievement of CR as well as 2-year OS (only 1/52 patients with t(8;21) did not achieve CR) which is in line with previous reports.[42,43] With respect to baseline

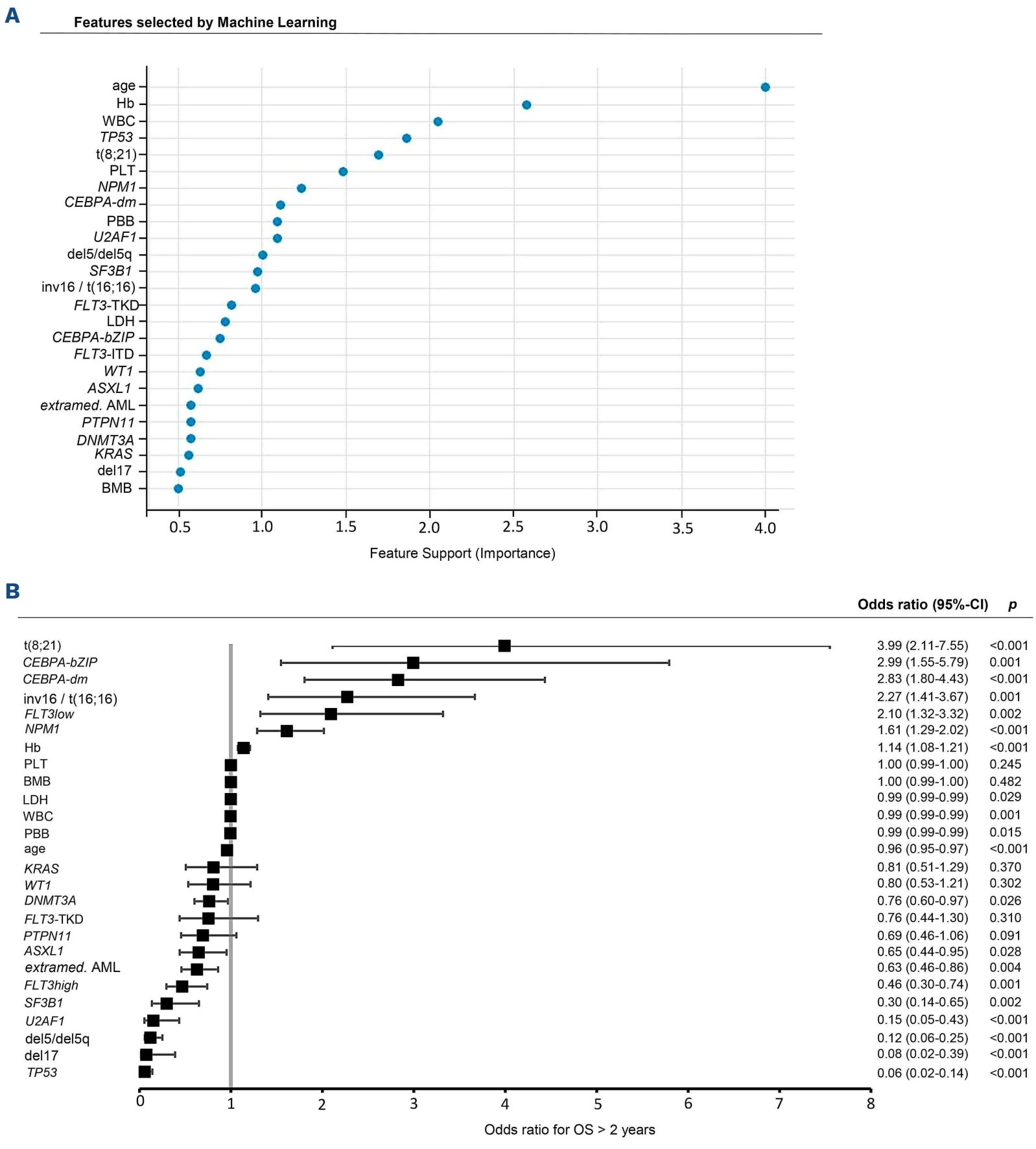


Figure 6. Feature selection for prediction of overall survival ≥2 years. (A) As for the prediction of complete remission (CR)/complete remission with incomplete hematologic recovery (CRi) (Figure 3), the feature selection algorithms were implemented to determine predictive features for overall survival (OS). Based on a continuous feature support metric with the same predefined cut-off that was used for CR/CRi prediction, 20 features were selected to predict OS ≥2 years. (B) For each of these features, odds ratios and 95% confidence intervals (95% CI) were calculated. BMB: bone marrow blast count; FLT3h/low: *FLT3*-ITD ratio, h=high>0.5; Hb: hemoglobin; karyotype, c: complex aberrant karyotype (≥3 aberrations); karyotype, n: normal karyotype (no aberrations); LDH: lactate dehydrogenase; PBB: peripheral blood blast count; PLT: platelet count; WBC: white blood cell count.

clinical and laboratory parameters, our analysis showed both CR rates and 2-year OS were significantly associated with age and hemoglobin level at initial diagnosis. Increasing age was associated with progressively lower odds of achieving CR and surviving for 2 years or longer despite the fact that all patients in our cohort received intensive induction therapy. Age is associated with high-risk molecular and cytogenetic features, lower frequencies of favorable markers and poor CR rates and survival.[5,44] Correspondingly, decreasing hemoglobin levels at initial diagnosis were associated with decreased odds of achieving CR and OS. Using these pre-trained ML prediction models, we validated our findings in an external multicenter cohort of 664 AML patients. Model performance remained stable in the external validation despite the fact that two important prognostic variables – *FLT3*-TKD and *IKZF1* - were missing in the external validation cohort, thus demonstrating adequate model transferability both for CR and 2-year OS predictions. Smaller discrepancies in performance between the internal test set and the external validation cohort may stem from missingness of these prognostic variables and/or random fluctuations. Potentially, an inclusion of more external data into the models' training may further boost performance and even out smaller discrepancies in performance metrics.

The performance of previous efforts at CR prediction in AML using conventional statistical approaches was reportedly moderate. In an analysis of over 4,500 intensively treated adult patients including commonly available clinical characteristics as well as *FLT3* and *NPM1* mutation status, Walter *et al.*[7] reported an AUROC between 0.71 and 0.78 while Krug *et al.*[45] similarly reported an AUROC of 0.72 in a cohort of more than 2,000 patients aged ≥60 years with newly diagnosed and intensively treated AML. These moderate accuracies even in large data sets incentivize the implementation of new approaches for data processing in risk evaluation. So far, only a few studies have used ML to predict CR in AML. Gal *et al.*[46] reported a k-nearest neighbor classifier evaluating bone marrow specimens from 473 AML patients between 8 days and 28 years old with an AUROC of 0.81 in their test set. The recent Dialogue for Reverse Engineering Assessment and Methods (DREAM) Acute Myeloid Leukemia Outcome Prediction Challenge was a crowdsourcing effort of 270 registered participants and 79 contributing teams developing over 60 algorithms on proteomic data from a training set of 191 and a test set of 100 AML patients with response to therapy being the primary clinical endpoint in sub-challenge one.[47] A final AUROC of 0.796 and a balanced accuracy of 0.779 were reported for the best performing model in the sub-challenge using a random forest model with an evolutionary weighting approach to feature selection.[47] Arguably, recent ML efforts in risk stratification, including our study, demonstrate the feasibility of ML technology to identify patients at high risk of treatment failure even considering that most of these recent studies using ML had far smaller data sets than the previously reported models using conventional statistical approaches. In order to implement these models meaningfully into clinical practice, they should not only include genetic alterations, but also acknowledge clinical patients' characteristics. While genetic alterations are undoubtedly powerful predictors of disease progression, a third of observed variation in survival still stems from demographic and clinical data.[48] We believe that the combination of both clinical and genetic data is essential for ML approaches to be beneficial for clinical practice in terms of treatment decision support, possibly in the form of knowledge banks, as recently reported by Gerstung *et al.*[49] They used a data-mining approach comparing different statistical models for outcome prediction with respect to matched genomic and clinical data of 1,540 patients. Gerstung *et al.*[49] reported that models including a larger variety of relevant data are able to predict patients' outcome more precisely than done so by restricted models such as the ELN 2017 classification.[17] We concur that predictive models incorporating a wide variety of available data from multiple sources for an individual patient may potentially provide a more detailed outlook on the outcome of that particular patient. However, a lack of clinical variables reduces the transferability of ML models based solely on genomic data sets to everyday clinical use as in-depth genetic sequencing is often either not available or not implemented in routine diagnostics. Our approach, however, utilizes both commonly available clinical variables as well as genetic events that can easily be extracted by commercial next-generation sequencing panels encompassing the most commonly mutated genes in AML. Furthermore, our approach was trained and tested on a large multicenter data set and validated on multicenter external data showing high accuracy in identifying patients at risk of primary treatment failure after intensive induction regimens. In such patients, in whom intensive therapy likely does more harm than good, novel regimens with less toxicity can be used, such as the combination of venetoclax and azacitidine for older patients with newly diagnosed AML.[50]

A limitation of our approach, however, is its retrospective nature. Many recent efforts of ML in hematology, including our study, are based on historic data sets.[11] Another limitation of our study is the unavailability of data on measurable residual disease. Assessment of measurable residual disease has become increasingly important in treatment surveillance in AML.[51] All of the patients in our study were treated with conventional chemotherapy regimens, except a minority of patients from the SORAML

study who were additionally treated with sorafenib. However, according to the original report, sorafenib did not affect CR rate or OS.[15] Future work will address the ability of ML to predict response to novel treatment regimens, measurable residual disease as well as prospective validation and the implementation of CR prediction for the individual patient at initial diagnosis, ideally including data for a variety of targeted therapies. ML performance is known to scale with sample size and a challenge will be the transfer to smaller data sets as data from trials with targeted therapies emerge. As another limitation of our approach, estimation of OS was confined to a binary classification after dichotomization of the cohort of patients into those who survived longer than 2 years and those who died within 2 years after initial diagnosis. The F1 scores for OS prediction were lower than those for CR prediction. This is arguably a result of the dichotomization of OS and consequent loss of longitudinal information regarding different survival times. Future work will focus on the implementation of longitudinal ML regression models for a more precise estimation of survival times.

In order to be implemented into clinical practice, such ML models must be easily accessible by practicing clinicians, build on commonly available data and should be cost-effective while providing accurate and robust prediction results to guide therapeutic strategies. An important goal of our work from a technological perspective was the transferability of our ML pipeline to other cases as most parts of the pipeline are automated and can, potentially, be used for other use cases after adequate data pre-processing, as demonstrated in external validation. Therefore, future work will also focus on transferability of our methodology to other cancer entities which is advantageous over more static conventional statistical approaches that are designed for a specific data set. Incorporating nine ML classifiers instead of one into the pipeline acknowledges that one classifier may be better suited for one use case while another may be superior in a different use case. This is especially evident in the direct comparison of performance between the internal test set and external validation cohort. for example in CR prediction for which the best performing algorithms in internal testing were random forest, logistic regression and linear SVM while in external validation RBF-SVM was superior to random forest, logistic regression and linear SVM, thereby demonstrating the relevance of including more than one ML algorithm in cancer data analysis.

In summary, we evaluated nine ML models on a large multicenter data set of 1,383 intensively treated AML patients and demonstrated high accuracy for CR and OS prediction in both internal testing and external validation. We provide a method to automatically select predictive features from different data types, cope with gaps and redundancies, apply and optimize different ML models and evaluate optimal configurations in a scalable and reusable ML platform. In a proof-of-concept manner, our algorithms utilize both established markers of favorable or adverse risk and also provide further evidence for the roles of *U2AF1*, *IKZF1*, *SF3B1*, *DNMT3A* and bZIP mutations of *CEBPA* in AML risk prediction. Our study serves as a fundament for prospective validation and data-driven ML-guided risk assessment in AML at initial diagnosis for the individual patient.

Disclosures

CT is chief executive officer and co-owner of AgenDix GmbH, a company that performs molecular diagnostics. The other authors declare that they have no competing financial interests.

Contributions

J-NE, KW and JMM designed the study. SS, J-AG, and CT performed molecular analyses. J-NE, CR, KM, MK, KS, UK, JB, DG, CMS, BW, TH, WB, WH, FK, JS, UP, CM-T, TS, HS, CB, KS-E, MK, SK, MHänel, CS, MHanoun, CT, MB, and JMM provided patients' samples. J-NE, PH, and KW developed and implemented the machine learning framework. All authors analyzed and interpreted the data. J-NE wrote the draft. All authors provided important scientific insights, critically revised and edited the manuscript. All authors approved the final version of the manuscript.

Acknowledgments

The authors would like to thank all contributing physicians, laboratories and nurses associated with the German Study Alliance Leukemia and especially participating patients for their valuable contributions. The Else-Kroener-Fresenius Center for Digital Health (EKFZ) is acknowledged for supporting the AI initiative at the Medical Faculty of the Technical University Dresden.

Funding

This work was supported by a MeDDrive grant, number 60499 'Machine learning for advanced integrated diagnostics in hematological malignancies' to JMM from the Technical University Dresden. J-NE is grateful for research support via a scholarship from the Mildred-Scheel-Nachwuchszentrum (German Cancer Aid).

Data-sharing statement

Data are available from the corresponding author upon reasonable request.

References

1. Shallis RM, Wang R, Davidoff A, Ma X, Zeidan AM. Epidemiology of acute myeloid leukemia: recent progress and enduring challenges. Blood Rev. 2019;36:70-87.
2. Walter RB, Kantarjian HM, Huang X, et al. Effect of complete remission and responses less than complete remission on survival in acute myeloid leukemia: a combined Eastern Cooperative Oncology Group, Southwest Oncology Group, and M. D. Anderson Cancer Center study. J Clin Oncol. 2010;28(10):1766-1771.
3. Koreth J, Schlenk R, Kopecky KJ, et al. Allogeneic stem cell transplantation for acute myeloid leukemia in first complete remission: systematic review and meta-analysis of prospective clinical trials. JAMA. 2009;301(22):2349-2361.
4. Bose P, Vachhani P, Cortes JE. Treatment of relapsed/refractory acute myeloid leukemia. Curr Treat Options Oncol 2017;18(3):17.
5. Appelbaum FR, Gundacker H, Head DR, et al. Age and acute myeloid leukemia. Blood. 2006;107(9):3481-3485.
6. Farag SS, Archer KJ, Mrózek K, et al. Pretreatment cytogenetics add to other prognostic factors predicting complete remission and long-term outcome in patients 60 years of age or older with acute myeloid leukemia: results from Cancer and Leukemia Group B 8461. Blood. 2006;108(1):63-73.
7. Walter RB, Othus M, Burnett AK, et al. Resistance prediction in AML: analysis of 4,601 patients from MRC/NCRI, HOVON/SAKK, SWOG, and MD Anderson Cancer Center. Leukemia. 2015;29(2):312-320.
8. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-387.
9. Alpaydin E. Introduction to Machine Learning. MIT Press; 2020. 709 p.
10. Bishop C. Pattern Recognition and Machine Learning. New York: Springer-Verlag.
11. Eckardt J-N, Bornhäuser M, Wendt K, Middeke JM. Application of machine learning in the management of acute myeloid leukemia: current practice and future prospects. Blood Adv. 2020;4(23):6077-6085.
12. Röllig C, Thiede C, Gramatzki M, et al. A novel prognostic model in elderly patients with acute myeloid leukemia: results of 909 patients entered into the prospective AML96 trial. Blood. 2010;116(6):971-978.
13. Schaich M, Parmentier S, Kramer M, et al. High-dose cytarabine consolidation with or without additional amsacrine and mitoxantrone in acute myeloid leukemia: results of the prospective randomized AML2003 trial. J Clin Oncol. 2013;31(17):2094-2102.
14. Röllig C, Kramer M, Gabrecht M, et al. Intermediate-dose cytarabine plus mitoxantrone versus standard-dose cytarabine plus daunorubicin for acute myeloid leukemia in elderly patients. Ann Oncol. 2018;29(4):973-978.
15. Röllig C, Serve H, Hüttmann A, et al. Addition of sorafenib versus placebo to standard therapy in patients aged 60 years or younger with newly diagnosed acute myeloid leukaemia (SORAML): a multicentre, phase 2, randomised controlled trial. Lancet Oncol. 2015;16(16):1691-1699.
16. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood. 2016;127(20):2391-2405.
17. Döhner H, Estey E, Grimwade D, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood. 2017;129(4):424-447.
18. Stasik S, Schuster C, Ortlepp C, et al. An optimized targeted next-generation sequencing approach for sensitive detection of single nucleotide variants. Biomol Detect Quantif. 2018;15:6-12.
19. Metzeler KH, Herold T, Rothenberg-Thurley M, et al. Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. Blood. 2016;128(5):686-698.
20. Montalban-Bravo G, Kanagal-Shamanna R, Class CA, et al. Outcomes of acute myeloid leukemia with myelodysplasia related changes depend on diagnostic criteria and therapy. Am J Hematol. 2020;95(6):612-622.
21. Zhang X, Zhang X, Li X, et al. The specific distribution pattern of IKZF1 mutation in acute myeloid leukemia. J Hematol Oncol. 2020;13(1):140.
22. Hunter AM, Sallman DA. Current status and new treatment approaches in TP53 mutated AML. Best Pract Res Clin Haematol. 2019;32(2):134-144.
23. Middeke JM, Herold S, Rücker-Braun E, et al. TP53 mutation in patients with high-risk acute myeloid leukaemia treated with allogeneic haematopoietic stem cell transplantation. Br J Haematol. 2016;172(6):914-922.
24. Gaidzik VI, Bullinger L, Schlenk RF, et al. RUNX1 mutations in acute myeloid leukemia: results from a comprehensive genetic and clinical analysis from the AML Study Group. J Clin Oncol. 2011;29(10):1364-1372.
25. Pratcorona M, Abbas S, Sanders MA, et al. Acquired mutations in ASXL1 in acute myeloid leukemia: prevalence and prognostic value. Haematologica. 2012;97(3):388-392.
26. Swerdlow SH, Campo E, Pileri SA, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood. 2016;127(20):2375-2390.
27. Falini B, Brunetti L, Sportoletti P, Martelli MP. NPM1-mutated acute myeloid leukemia: from bench to bedside. Blood. 2020;136(15):1707-1721.
28. Falini B, Martelli MP, Bolli N, et al. Acute myeloid leukemia with mutated nucleophosmin (NPM1): is it a distinct entity? Blood. 2011;117(4):1109-1120.
29. Thiede C, Koch S, Creutzig E, et al. Prevalence and prognostic impact of NPM1 mutations in 1485 adult patients with acute myeloid leukemia (AML). Blood. 2006;107(10):4011-4020.
30. Taskesen E, Bullinger L, Corbacioglu A, et al. Prognostic impact, concurrent genetic mutations, and gene expression features of AML with CEBPA mutations in a cohort of 1182 cytogenetically normal AML patients: further evidence for CEBPA double mutant AML as a distinctive disease entity. Blood. 2011;117(8):2469-2475.
31. Gale RE, Green C, Allen C, et al. The impact of FLT3 internal tandem duplication mutant level, number, size, and interaction with NPM1 mutations in a large cohort of young adult patients with acute myeloid leukemia. Blood. 2008;111(5):2776-2784.
32. Thiede C, Steudel C, Mohr B, et al. Analysis of FLT3-activating mutations in 979 patients with acute myelogenous leukemia: association with FAB subtypes and identification of subgroups with poor prognosis. Blood. 2002;99(12):4326-4335.
33. Cazzola M. Myelodysplastic syndromes. N Engl J Med. 2020;383(14):1358-1374.
34. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic classification and prognosis in acute myeloid leukemia. N Engl J Med. 2016;374(23):2209-2221.
35. Vairy S, Tran TH. IKZF1 alterations in acute lymphoblastic leukemia: the good, the bad and the ugly. Blood Rev. 2020;44:100677.

36. Ley TJ, Ding L, Walter MJ, et al. DNMT3A mutations in acute myeloid leukemia. N Engl J Med. 2010;363(25):2424-2433.
37. Patel JP, Gönen M, Figueroa ME, et al. Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. N Engl J Med. 2012;366(12):1079-1089.
38. Yang L, Rau R, Goodell MA. DNMT3A in haematological malignancies. Nat Rev Cancer. 2015;15(3):152-165.
39. Tarlock K, Lamble A, Wang J, et al. CEBPA bZip mutations are associated with favorable prognosis in de novo AML: a report from the Children's Oncology Group. Blood. 2021;138(13):1137-1147.
40. Taube F, Georgi JA, Kramer M, et al. CEBPA mutations in 4708 patients with acute myeloid leukemia - differential impact of bZIP and TAD mutations on outcome. Blood. 2022;139(1):87-103.
41. Marimont RB, Shapiro MB. Nearest neighbour searches and the curse of dimensionality. IMA J Appl Math. 1979;24(1):59-70.
42. Schiffer CA, Lee EJ, Tomiyasu T, Wiernik PH, Testa JR. Prognostic impact of cytogenetic abnormalities in patients with de novo acute nonlymphocytic leukemia. Blood. 1989;73(1):263-270.
43. Dastugue N, Payen C, Lafage-Pochitaloff M, et al. Prognostic significance of karyotype in de novo adult acute myeloid leukemia. The BGMT group. Leukemia. 1995;9(9):1491-1498.
44. Kantarjian H, O'Brien S, Cortes J, et al. Results of intensive chemotherapy in 998 patients age 65 years or older with acute myeloid leukemia or high-risk myelodysplastic syndrome: predictive prognostic models for outcome. Cancer. 2006;106(5):1090-1098.
45. Krug U, Röllig C, Koschmieder A, et al. Complete remission and early death after intensive chemotherapy in patients aged 60 years or older with acute myeloid leukaemia: a web-based application for prediction of outcomes. Lancet. 2010;376(9757):2000-2008.
46. Gal O, Auslander N, Fan Y, Meerzaman D. Predicting complete remission of acute myeloid leukemia: machine learning applied to gene expression. Cancer Inform. 2019;18:1176935119835544.
47. Noren DP, Long BL, Norel R, et al. A crowdsourcing approach to developing and assessing prediction algorithms for AML prognosis. PLOS Comput Biol. 2016;12(6):e1004890.
48. Walter RB, Estey EH. Selection of initial therapy for newly-diagnosed adult acute myeloid leukemia: limitations of predictive models. Blood Rev. 2020;44:100679.
49. Gerstung M, Papaemmanuil E, Martincorena I, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. Nat Genet. 2017;49(3):332-340.
50. DiNardo CD, Jonas BA, Pullarkat V, et al. Azacitidine and venetoclax in previously untreated acute myeloid leukemia. N Engl J Med. 2020;383(7):617-629.
51. Voso MT, Ottone T, Lavorgna S, et al. MRD in AML: the role of new techniques. Front Oncol. 2019;9:655.