Development of a dynamic counterfactual risk stratification strategy for newly diagnosed acute myeloid leukemia patients treated with venetoclax and azacitidine

Nazmul Islam[1*], PhD; Justin L. Dale[1], BBA; Jamie S. Reuben[1], MPH; Karan Sapiah[1], MS; James W. Coates[1], MS; Frank R. Markson[1], BA; Jingjing Zhang[2], MD, PhD; Lezhou Wu[1], PhD; Maura Gasparetto[3], MD; Brett M. Stevens[1,3], PhD; Sarah E. Staggs[1,3], PhD; William M. Showers[1], MS; Monica R. Ransom[1,3], PhD; Jairav Desai[1], MS; Ujjwal V. Kulkarni[1], MS; Krysta L. Engel[1], PhD; Craig T. Jordan[3], PhD; Michael Boyiadzis[4], MD; Clayton A. Smith[1, 3,4], MD

[1]RefinedScience, Aurora, Colorado

[2]Department of Pathology, University of Colorado Anschutz Campus, Aurora, Colorado

[3]Department of Medicine, University of Colorado Anschutz Campus, Aurora, Colorado

[4]OncoVerity, Aurora, Colorado

***Communicating Author:**

Nazmul Islam, PhD

RefinedScience, 2115 N Scranton Street, Suite 2-70

Aurora Colorado, 80045, USA

+1-919-604-9705

nazmul.islam@refinedscience.com

Key words: Acute Myeloid Leukemia; Venetoclax; Azacitidine; Risk Stratification; Machine Learning

Abstract count: 248

Word count: 4,000 (Tables: 1, Figures: 6)

Article type: Research and Applications

ABSTRACT

Objective

The objective of this study was to develop a flexible risk model (RM) stratification strategy for Acute Myeloid Leukemia (AML) that is specific for the new standard of care venetoclax plus azacitidine (ven/aza), captures disease heterogeneity, and addresses a range of real-world data issues.

Materials and Methods

A series of tunable RMs based on a dynamic counterfactual machine learning (ML) strategy that utilized next generation sequencing, cytogenetics, flow cytometry, and other features of the diagnostic AML samples were developed and tested on a single institutional cohort of 316 newly diagnosed patients treated initially with ven/aza.

Results

Favorable, Intermediate, and Adverse risk groups were identified in a series of novel RMs derived using ML models for overall survival (OS) and event free survival (EFS). Most, but not all models, demonstrated equitable patient distribution into the different risk categories (~20%-40% in each group) with significant separation between categories (Log-Rank based p-values <0.001), and with predictability computed by survival AUC values in the ~0.60-0.70 range.

Discussion and Conclusion

The general strategy employed here is specific for AML patients treated with ven/aza, is based on a wide range of diagnostic AML pathology features, considers feature interactions, addresses data missingness, sparsity, and the confounding effects of allogeneic hematopoietic cell transplant . It is also readily tunable through simple coding and context specific parameter updates, and adaptable to reflect different use case needs. This strategy represents a new approach to developing more effective RMs for AML and possibly other diseases as well.

INTRODUCTION

Acute myeloid leukemia (AML) is an aggressive hematopoietic malignancy that is diagnosed in ~20,000 persons per year in the United States (US) [1]. For young and fit patients, treatment typically involves intensive chemotherapy (IC) initially followed by consolidation with allogeneic hematopoietic cell transplant (allo-HCT) or additional high dose chemotherapy [2,3]. For older and less fit patients, venetoclax (ven) combined with a hypomethylating agent (HMA) such as azacitidine (aza) has become a standard of care [4-6]. For IC therapies, a variety of prognostic models have been developed based on patient- and AML-specific features to stratify patients into subgroups with varying outcomes [7-15]. The European Leukemia Network (ELN) developed a widely used risk categorization approach for AML in 2017 (ELN17) with an update in 2022 (ELN22), which was adopted and modified by the National Comprehensive Cancer Network (NCCN) and others [16,17]. The ELN categories are divided into Favorable, Intermediate, and Adverse groups associated with overall survival (OS) based on AML cytogenetic (CYT), fluorescence *in situ* hybridization (FISH), and next generation sequencing (NGS) features [16,17]. However, it was recently shown that the ELN17 and ELN22 risk categorization approach performed sub-optimally and did not effectively stratify ven/HMA treated patients, likely because the ELN criteria are primarily based on treatment outcomes following IC [18-22]. Here we describe the development of a tunable and flexible AML risk stratification strategy for newly diagnosed AML patients treated specifically with ven/aza based on a wide range of AML diagnostic pathology features while addressing a variety of real-world data (RWD) challenges that can alter the effectiveness of a RM.

METHODS

Patient populations and outcomes definitions

The patient dataset included 316 adult patients from the University of Colorado (CU) with newly diagnosed AML treated front line with ven/aza either as standard of care or in a clinical trial between January 2015 and March 2024. Patients with acute promyelocytic leukemia (APL), bi-phenotypic leukemia, blast crisis chronic myeloid leukemia, or central nervous system involvement with AML at diagnosis were excluded. Details of the patient cohort and subset management are summarized in Figure 1. Patient demographics are summarized in Supplemental Table 1A. Key AML diagnostic pathology features considered in developing the RMs are summarized in Supplemental Tables 1B-F and included NGS, CYT, FISH, polymerase chain reaction (PCR) results, flow cytometric (FC) data, and a category termed "AML mutations" that included the composite of mutations defined by different tests (patients with more than one test positive were counted only once). “p53m” composite mutation, for example, is a combination of p53 abnormalities by NGS and losses or gains of 17p by FISH. Complex cytogenetics was defined as $\geq$3 karyotype abnormalities. Data including NGS and FC results were simplified into binary values as described previously [23]. FISH data were simplified by combining all chromosomal specific probes into one term "Chr Abn". FISH data for DEK and MECOM was not routinely available prior to mid-2019 (0 DEK results, 3/7 positive MECOM results) and so was not utilized prior to that date. Additional feature nomenclature is described in the figure legends and Supplemental Table 7. Responses including OS, BR, and EFS are defined in Figure 1 [17, 23]. This retrospective study was approved by internal review board (IRB) and used a limited dataset with a waiver of consent from the IRB.

Exploratory analyses

Descriptive statistics were provided for baseline covariates and outcomes [24]. Kaplan-Meier (KM) analyses were performed for time-to-event data assuming right censoring. Patients were censored at their last known reported alive date. 95% confidence intervals (CIs) and median survival times were

reported. KM curves were estimated for OS and EFS based on the Full Analytical Set (FAS), a dataset where allo-HCT recipients were censored, and a set excluding allo-HCT recipients. Zero-variation and low-count variables (i.e., ≤ 4 counts in any label of a categorical variable) were not included in analyses. Univariate association filtering by null hypothesis significance testing (NHST) was used to exclude noise variables. The equality of survival curves was evaluated based on Log-rank (LR), Tarone-Ware (TW), Fleming-Harrington (FH) with an emphasis on early and late time periods, weighted multiple direction (mdir) LR, Max-Combo (MC), K-sample omnibus non-proportional hazard (KONP), and restricted mean survival times (RMST) tests [25-31]. The different testing approaches were used to address the complexities from violating proportionality assumptions in survival curves and survival differences over time. Variables with a *P* value < 0.30 based on at least one of the univariate tests in any KM analyses (OS and EFS) were considered as candidates for multivariate models. The resulting combined list of confounding covariates (Supplemental Tables 1B-F) was used for both multivariable OS and EFS models.

Risk model development

Table 1 summarizes the overall RM specifications and Figure 2 summarizes the RM development strategy. Key components and considerations in the RM development include:

Survival regressions

Machine learning (ML) models (multivariable penalized Cox proportional hazard with a ridge penalty (mCOXr) [32,33]) were fitted adjusting for age and gender for OS and EFS. Separate models were fitted adjusting for the confounding variables selected by NHST as described above (see Step 1 of Figure 2) to identify the most influential variables within each group independently (Step 3 of Figure 2). Missing values were treated as a separate category in modeling and as described below (Step 2 of Figure 2). Tuning parameters were selected via cross-validation (CV) method. 95% CIs of adjusted hazard ratios (aHRs) were constructed based on the fractional random weight bootstrap (FRWB) approach of 2,000 runs [34].

Counterfactual effects estimation

Biomarker-specific mean predicted marginal risks (MRs) over time were calculated via "*recycled*" predictions assuming counterfactual arguments with respect to the corresponding mCOXr models (Step 4 of Figure 2) [35,36]. A fixed patient cohort was used for each feature to estimate risk differences (RDs) adjusting for differences in the mix and severity of risk factors among subjects while reducing potential bias stemming from the skewed distributions of patient characteristics. At each time point, MRs of a biomarker being positive or negative were calculated by averaging the predicted marginal probabilities of risk for all patients in the sample as if they had been (hypothetically) (+ve) or (-ve) for genetic features or if they expressed a marker by FC. Bootstrap-based MR profiles were constructed by averaging over FRWB estimates. The corresponding percentile-based 95% CIs were reported. The Kolmogorov-Smirnov (KS) test was used to assess the differences in counterfactual risk distributions [37]. Multivariable penalized (ridge) logistic regressions [33] were used to model BR (i.e., CR/CRh) confirming the consistency of the mCOXr estimates.

Variations in risk models

Several issues were addressed in the model development.

Varying endpoints: RMs with two key outcomes in AML, OS and EFS, were developed separately in contrast to *one-size-fits-all* approach where the learnings based on an OS model are applied to different outcomes such as EFS.

Treatment variation: Allo-HCT is a potentially curative therapy which can profoundly impact outcomes in patients treated with ven/aza, which is largely palliative [38]. To address this potentially confounding factor, RMs were developed by treating allo-HCT patients censored at their first transplant date. In a sample with a high proportion of allo-HCT patients (92/316) as in CU, if not adjusted for allo-HCT, there is a risk of overestimation of survival probabilities leading to improper feature selections; furthermore, the characterization of underlying relationship between response and features will be challenging.

Dynamic behavior: Counterfactual RD were computed dynamically over time for up to thirty-six (36) months (Steps 4-5, Figure 2 and Supplemental Figures 1A-1D). Landmarks of 14.7 and 9.8 months

were chosen to find optimal relative feature-specific RD for OS and EFS respectively as these correspond to the median values in the VIALE-A study, which established ven/aza as a new standard of care [6].

Prevalences: Variables with low and high prevalence for (+ve) cases may lead to unstable parameter estimates and induce imbalances in subject-level stratification. Rule-I (Step 5, Figure 2) is flexible and does not exclude features with respect to prevalences. Three additional rules were considered with varying degrees of stringency – Rule-II excluded any variables containing less than 2% and more than 40% (+ve) cases in calculating risk stratification even if it exhibited more than 5% RD. Rule-III used FRWB based MR profiles to calculate RD, and Rule-IV exploited both Rule-II and Rule-III; see Supplemental Table 4A.

Covariate-level classification by RD: Two levels of classifications, covariate- and subject-level, were performed. First, the magnitude and direction of marginal RDs between (+ve) and (-ve) at a time point of interest (i.e., 14.7 months for OS and 9.8 months for EFS) were used to classify each feature into either "Adverse" or "Favorable". Positive RD values of 0-5%, 5-10%, and >10% classified a feature into either a *slightly*, *moderately*, or *strongly* "Adverse" feature, respectively. Similarly, negative RD values determined *slightly, moderately, or strongly* "Favorable" feature (Step 6, Figure 2). 5% cut-offs were assumed as clinically relevant minimum RD. Supplemental Tables 4B-4E depict covariate-level classifications under different combinations of parameters.

Subject-level classification: Patient-specific risk stratification ("Adverse", "Intermediate", and "Favorable") was determined based on combining covariate-level classifications (*moderately* and *strongly* "Adverse" and "Favorable") using a 'Boolean' operator (i.e., AND/OR). This assigns a patient into "Favorable" risk group in the absence of any "Adverse" features, and vice versa. A patient was assigned to "Intermediate" risk group in the absence of any "Favorable" and "Adverse" features or presence of at-least one "Favorable" and at-least one "Adverse" feature (Step 7, Figure 2).

Model specification: RMs based on both genetic features and genetic-plus-phenotypic features were developed to test the impact of varying feature sets on model performance (Table 1).

Missingness: Primary analyses were conducted treating missing values as a separate category for simplicity, avoiding imputation or exclusion. Reported results were based on this approach unless specified. Additional analyses were performed to check the robustness of different missing data handling techniques at the covariate-level classification step including: (A) *Running weighted models on complete cases (CC)*: Biases stemming from CC analysis was adjusted by incorporating inverse probability weights (IPW) into mCOXr where the subject-specific weights for missing data were estimated from a multivariable logistic regression model [39]. (B) *Imputing via multivariable regression models:* Multiple imputations (5 imputed samples) were generated for each missing data element exploiting multivariate data sequentially (multivariate imputation by chained equations (MICE)) [40,41]. aHRs were averaged over 5 samples and 2,000 FRWB runs; percentile based 95% CIs were reported. Subsequently marginal RDs were calculated as before and averaged over 5 samples. (C) *Synthetic models:* Covariate-level classifications were performed based on different missing methods and the final stratification was obtained by the *majority-votes*.

Evaluation of numerical performance

Multiple evaluation criteria were used to provide a holistic view of numerical performance in terms of *equitability (extent to which different risk group sizes vary)*, *separability (extent to which outcomes vary by different risk strata)*, *conformity (extent to which risk groups overlap between methods)*, and *predictability* (*degree of prognostic relationship with outcomes)*. Internal validation of *predictability* of the proposed and ELN22 for OS and EFS was conducted using 15-fold CV; the corresponding steps are illustrated in Figure 3. Further details of the computation and evaluation criteria are provided in Supplemental Methods.

All statistical tests, except the comparison of MR curves and survival AUC values, were two-sided, with a significance level of 5%. Since the objective of the analyses was exploratory, multiplicity adjustment was not made. All analyses were conducted using R, version 4.2.3.

RESULTS

OS risk models

Using the ML based strategy described above, we first explored whether a novel, genetic features-based RM, termed RM_G, could be developed. Covariate-level classification Rule-II demonstrated superior numerical performance with respect to *equitability* (Figure 4A)*, separability* (Figure 4B-D)*, conformity* (Supplemental Figure 5)*,* and *predictability* (Supplemental Table 5B). Risk features identified using this variant of RM_G, termed $RM_{G,II}$, and the distribution of patients in different risk categories are summarized in Figure 4A. $RM_{G,II}$ demonstrated equitable distributions of patients and effective separation of Adverse, Intermediate, and Favorable subgroups based on OS and BR in both the allo-HCT excluded and FAS cohorts (Figures 4B and 4C). Similar behavior was noted when applied to the patient cohort where allo-HCT recipients were treated as censored (Figure 4D, left panel). In the subset of only the patients who underwent allo-HCT, the RM did not distinguish Favorable and Intermediate risk patients, however it did distinguish the Adverse risk population from the other two categories (Figure 4D, right panel). Similar comparison was performed for RM_G variants using Rule-I ($RM_{G,I}$), Rule-III ($RM_{G,III}$), and Rule-IV ($RM_{G,IV}$), which demonstrated similar performance characteristics except for reduced equitability in the assignment of patients to different risk categories (see Supplemental Figure 2-4 as well as Supplemental Tables 5A, 5C, and 5D). Other OS models summarized in Table 1 provided similar results (Supplemental Table 4B, 4D-4E). To compare the $RM_{G,II}$ to the ELN22, several analyses were performed. The application of the ELN22 model to the FAS demonstrated a relative skewed distribution of patients assigned to the Adverse risk category and less effective separation of OS curves compared to the $RM_{G,II}$ model for both the allo-HCT excluded cohort and the FAS (Supplemental Figure 5A-B). Pairwise comparison in the allo-HCT excluded cohort between $RM_{G,II}$ and ELN22 demonstrated substantial differences between OS in the Adverse risk groups, moderate differences in the Intermediate groups, and relatively similar behavior in the Favorable risk groups (Supplemental Figure 5C). Inter-group conformity analyses by Fleiss kappa demonstrated the highest level of concordance in patient assignment for the Favorable risk category (Supplemental Figure 5D).

Predictive evaluations (time-varying AUCs and iBrier) favored the $RM_{G,II}$ model (Supplemental Table 5B). Next, we explored whether the same RM strategies could be applied to a feature set comprised of genetic-plus-phenotypic features defined by FC termed as RM_{GP}. The impacts of applying Rule-I, II, III, and IV to the RM_{GP} development strategy were explored as above (see Supplemental Figures 6-9 and Supplemental Tables 5A-5D top panels). In general, the overall performance of the RM_{GP} models paralleled those of the RM_{G} models above but with less equitable assignment of patients to the Favorable risk category depending on models, rules, and assumptions. Together these findings demonstrate that the proposed ML based strategy may be effective at generating AML ven/aza specific risk models for OS and highlight the impact of different assumptions and rules on the overall RM performance.

EFS risk models

Next, we explored whether applying the general RM development strategy to EFS was feasible as this is another key metric in AML analysis [42]. RMs with features based on the genetics and genetics-plus-phenotypic features set were developed and tested as above. The covariate-level classification Rule-II, EFS specific model using genetic features only (i.e., $RM_{G,II\text{-}EFS,}$ Figure 5A) demonstrated the best performance (Figure 5A-5C and Supplemental Table 6). Comparison between $RM_{G,II\text{-}EFS}$ and ELN22 for EFS was similar to that of the OS comparison above (Supplemental Figure 13). Rule-I, III, and IV were also explored in the context of the genetic features-based EFS RM. $RM_{G,I\text{-}EFS}$ resulted in skewed distributions among risk strata with few Favorable and Adverse risk patients leading to suboptimal numerical performance (Supplemental Figure 10 and Supplemental Table 6A bottom panel). In contrast, $RM_{G,EFS}$ models utilizing Rule-III and Rule-IV behaved more similarly to that of Rule-II (Supplemental Figures 11-12 and Supplemental Table 6C-6D bottom panels). When utilizing the genetics-plus-phenotypic features, the RM_{GP} strategy applied to EFS roughly paralleled the behavior of the RM_{G} strategy (Supplemental Figures 14-17 and Supplemental Tables 6A-6D top panels). Overall,

the findings indicate that EFS can be effectively stratified using the appropriate EFS specific RM_G and RM_{GP} models but again highlight the impact of different assumptions and rules on model performance.

A summary of heatmaps of patient-specific risk classifications of the 316 patients for both OS and EFS based on the various RMs described above are depicted in Figure 6 and Supplemental Figure 18 that further highlight the performance and variations of other models and assumptions.

DISCUSSION

The purpose of the study was to develop a flexible risk stratification strategy reflecting practical RWD challenges which could be readily tunable for specific use cases, objectives, and resources without loss of generality. This approach is inherently more flexible and adaptive than the ELN17 and ELN22 risk categories where the risk stratification remains relatively fixed until accumulated published data leads to a new update after several years. Specifically, this traditional risk categorization approach can lag rapidly evolving practice changes in diagnoses and treatments such as the advent of ven/HMAs as a new standard of care treatment in patients unfit for IC. Recently, lower intensity specific RMs have been developed to address some of these concerns for patients with AML unfit for IC[18-21,43]. However, several of these RMs were derived from patients treated with a range of agents and schedules, and were based on a limited number of NGS features that may not fully reflect AML heterogeneity and biology. In addition, these newer RMs may not have captured important RWD challenges including missingness, sparsity, scarcity, type of end-points, collinearity, and the impact of allo-HCT on OS [19-21,43,44]. Lastly, as with the ELN17 and ELN22, many of these models were defined by relatively fixed rules and assumptions that are not readily adaptable as technologies and findings evolve. In contrast, the proposed ven/aza specific RMs 1) were based on probabilistic RDs, 2) dealt with feature interactions (using a logic based approach at the level of two-factor interactions), 3) addressed dynamic effects of features on outcomes, 4) were developed separately for specific end-points (OS and EFS), 5) were treatment specific, 6) accounted for the confounding effects of allo-HCT on therapeutic outcomes, and 7) included a range of features (i.e., genetic only and genetic-plus-phenotypic) adjusting for collinearity. The proposed RMs were assessed by *equitability*, *separability*, *conformity*, and *predictability* demonstrating both clinical and numerical effectiveness while highlighting that different underlying modeling assumptions, adjustments, and rules can have influences on both the covariate- and subject-level classifications. In initial analyses, model $RM_{G,II}$ appeared to have better performance with these metrics than the ELN22; see Supplemental Figure 5 and 13 for *equitability, conformity* and *separability* and Supplemental Table 5 for *predictability*. If further validated, this could have important

consequences for assessing the impact of novel AML therapies and diagnostics in clinical trials and real-world settings.

One important step in the RM development strategy was adopting counterfactually estimated RD over time in quantifying feature importance. Rather than solely relying on the traditional NHST approach, which is subject to underlying statistical power considerations, we assumed and employed a clinically relevant RD threshold (e.g., 5%) for selecting important risk features. The RMs were also developed accounting for variables with low and high frequencies along with bootstrapping that minimizes sampling variations inducing more robust parameter estimates. Our models and results were derived from a dataset comprised in part from real-world AML data and adjusted for practical RWD issues. Lastly, these RM models are tunable through relatively simple coding, parameters, and features updates so they can readily accommodate changes in diagnostics and treatments as well as in the assumptions, definitions, and rules to reflect the evolving changes in information, technology, and use cases. This is particularly important in AML as rapid advancements in clinical care, treatment, insights related to disease biology, and new testing modalities lead to increasingly dynamic and complex datasets; and all these developments collectively introduce rapid changes into RM performance and utility. The development, external validation, and further comparison to the ELN22 of a single, simple, and practical version of the ML based models described here is underway and the current status of this work is described in a preprint [45].

Different risk models have been proposed to emphasize the importance of different variations in real-world and clinical trial data, objectives, assumptions, and requirements. The model performing the "best" typically is subjective and depends on the context and dataset where the model is applied to. The proposed approach is inherently different from the conventional *one-size-fits-all* approach which assumes that the model works well for all scenarios, irrespective of differences in the underlying datasets. In contrast, the proposed strategy asks an investigator to consider the nuances in the

respective analytical dataset and then choose the model parameters accordingly. For example, the proportion of allo-HCT varies significantly among different sites and has a profound confounding effect on overall or event-free survival. Therefore, for a site such as the CU where the proportion of allo-HCT is high, adjustment needs to be made to control this confounding effect. However, this may be redundant for sites where the proportion of allo-HCT is low or negligible. In addition, one needs to choose how rigorously to account for missingness in the dataset which primarily depends on the magnitude of missingness. Moreover, depending on the objectives of a study, a user may prefer phenotypic-and-genetic models over genetic models alone. Therefore, our approach is philosophically different in that, rather than prescribing a static final model, it enables potential clinical implementation of a risk stratification methodology that is adapted to site-specific preferences and the population of patients treated at the site.

There are several limitations to this study. This is an exploratory retrospective study, so the strategy and findings need to be validated in confirmatory prospective studies. While genetic, FC, FISH, and composite AML mutation features were simplified into binary variables for numerical convenience and ease of interpretability in counterfactual estimates, these simplification steps may result in critical information loss and miss important biological aspects of AML. Similarly, the composite mutation category for RUNX and other features (see Supplemental Table 1E) treated all abnormalities as potentially biologically relevant which may also be an oversimplification. Reintroducing the underlying complexity of the constituents of these features may improve RM stratification ability providing important clinical insights; however future studies with larger datasets will be required for validations. While we were able to develop RMs combining both genetic and phenotypic features, FC remains unstandardized between many centers, so the proposed models may be best suited currently for a single-center or small consortium setting with consistent FC datasets. There is also a risk of underreporting samples with dimly expressing FC features which may occur with markers such as CD45. Some RMs exhibited sub-optimal stratification ability for allo-HCT recipients which may be

improved by developing RMs utilizing features relevant only to this patient cohort. The interplay between a variety of genetic and AML features was accounted for based on a logical rather than parametric basis. An alternative modeling approach with parametric-based interaction effects or non-parametric tree-based approaches trained on a larger sample size could alleviate this concern. Due to using the same analytical dataset for different tasks such as variable selection, covariate- and subject-level risk classifications, and internal validation, there is a risk of “data-leakage” leading to over-fitting. Additionally, all the models, particularly those with higher stringency, run a risk of missing potentially important features due to the underestimation of effects, small sample size, and prespecified cut-offs for prevalence rates and RDs. In contrast, the less stringent RMs may incorporate noisy features with spurious associations. Therefore, to understand the generalizability of the proposed RMs, it is critical to validate the findings using additional external, independent, and diverse data from multiple institutions. Further simplifications of the RMs are also warranted to facilitate more seamless adoption and application in AML management and research. Finally, these findings should be further validated through comparison to published findings and newly emerging RMs developed with other datasets specific for ven/aza and other low intensity AML treatments [20,43,44]. A simplified and practical version of the risk models described in this report has been developed and validated with a separate external dataset; details are provided in Islam et al [45].

CONCLUSION

In summary, we developed and internally validated a risk stratification strategy and series of RMs for AML patients treated specifically with ven/aza which attempt to address a variety of limitations inherent in previous AML risk classification models. The simplified practical version of these novel models could be useful for upfront patient allocation to the different risk groups in clinical trials as well as the post hoc analysis of novel treatment approaches on different AML risk groups. The proposed overall strategies are disease-agnostic and may be adaptable to other diseases as well.

ACKNOWLEDGMENTS

The authors gratefully thank Austin E. Gillen from the Department of Medicine, University of Colorado Anschutz Campus, Aurora, Colorado; in addition, Bin Yao and Adam George from the OncoVerity, Aurora, Colorado; and Grant Weller from the RefinedScience, Aurora, Colorado for their feedback and reviews during this project.

CONFLICT OF INTEREST

Both CAS and MB are employees of and hold equity in OncoVerity. In addition, CAS is a consultant to RefinedScience. All other authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

CAS and NI designed the study and drafted the manuscript. JLD, JSR, KS, JWC, FRM, and UVK processed and pulled the structured analytical datasets. CAS, NI, JSR, JLD, JZ, KS, JWC, and LW assessed the validity and quality of data. NI performed numerical analyses. CAS, NI, CTJ, JSR, JLD, JZ, MB, SES, MG, BMS, WMS, MRR, and KLE interpreted the results of the analyses. All authors reviewed, provided constructive comments, and agreed to its publication.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

SUPPLEMENTARY MATERIAL

Supplementary material is available along with the submission detailing additional numerical results pertinent to the study.

DATA AVAILABILITY

The raw, individual patient data are protected and not available due to data privacy laws. The processed data are available at reasonable request to the corresponding author.

CODE SHARING

Controlled access to the code may be provided to qualified researchers upon request under a confidentiality and data use agreement, subject to institutional approval. For request, please email at nazmul.islam@refinedscience.com.

ABBREVIATIONS

Acute myeloid leukemia	AML
Intensive chemotherapy	IC
Allogeneic hematopoietic cell transplant	allo-HCT
European Leukemia Network	ELN
Complete response	CR
Best response	BR
CR with incomplete hematologic recovery	CRi
CR with partial hematologic recovery	CRh
Morphologic leukemia free state	MLFS
Partial remission	PR
Cytogenetic	CYT
Flow cytometric	FC
Fluorescence in situ hybridization	FISH
Next generation sequencing	NGS
Polymerase chain reaction	PCR
Venetoclax plus azacytidine	ven/aza
Full analytical set	FAS
Event-free survival	EFS
Overall survival	OS
Institutional review board	IRB
University of Colorado	UC
Kaplan-Meier	KM
Confidence interval	CI
Null hypothesis significance testing	NHST
Log-rank	LR
Tarone-Ware	TW
Fleming-Harrington	FH
Weighted multiple direction	mdir

Max-Combo	MC
K-sample omnibus non-proportional hazard	KONP
Restricted mean survival times	RMST
Machine learning	ML
Multivariable Cox proportional hazard models with ridge (L2-norm) penalty	mCOXr
Proportional hazard	PH
Fractional random weight bootstrap	FRWB
Adjusted hazard ratio	aHR
Cross-validation	CV
Kolmogorov-Smirnov	KS
Area under the curve	AUC
Receiver operative characteristics	ROC
Cumulative case dynamic control ROC	cAUC
Incident case dynamic control ROC	iAUC
Integrated Brier scores	IBS
Complete cases	CC
Inverse probability weights	IPW
Multivariate imputation by chained equations	MICE
Marginal risk	MR
Risk model	RM
Risk difference	RD
RM based on genetic-plus-phenotypic features with respect to OS	RM_{GP}
RM based on genetic features with respect to OS	RM_{G}
RM based on genetic-plus-phenotypic features with respect to EFS	$RM_{GP\text{-}EFS}$
RM based on genetic features with respect to EFS	$RM_{G\text{-}EFS}$
Real world data	RWD
Rule-I	I
Rule-II	II
Rule-III	III
Rule-IV	IV

REFERENCES

1 Shallis, R. M., Wang, R., Davidoff, A., Ma, X. & Zeidan, A. M. Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges. *Blood Rev* **36**, 70-87 (2019). https://doi.org/10.1016/j.blre.2019.04.005

2 Bittencourt, M. C. B. & Ciurea, S. O. Recent Advances in Allogeneic Hematopoietic Stem Cell Transplantation for Acute Myeloid Leukemia. *Biol Blood Marrow Transplant* **26**, e215-e221 (2020). https://doi.org/10.1016/j.bbmt.2020.06.007

3 Blum, W. G. & Mims, A. S. Treating acute myeloid leukemia in the modern era: A primer. *Cancer* **126**, 4668-4677 (2020). https://doi.org/10.1002/cncr.32904

4 Chua, C. C. *et al.* Chemotherapy and Venetoclax in Elderly Acute Myeloid Leukemia Trial (CAVEAT): A Phase Ib Dose-Escalation Study of Venetoclax Combined With Modified Intensive Chemotherapy. *J Clin Oncol*, JCO2000572 (2020). https://doi.org/10.1200/JCO.20.00572

5 DA, P., M, A., P, S. & MY, K. Venetoclax for AML: changing the treatment paradigm. Blood Adv. 2019;3(24):4326-4335. *Blood Adv* **4**, 1020 (2020). https://doi.org/10.1182/bloodadvances.2020001769

6 DiNardo, C. D. *et al.* Azacitidine and Venetoclax in Previously Untreated Acute Myeloid Leukemia. *N Engl J Med* **383**, 617-629 (2020). https://doi.org/10.1056/NEJMoa2012971

7 Kantarjian, H. *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* **106**, 1090-1098 (2006). https://doi.org/10.1002/cncr.21723

8 Malfuson, J. V. *et al.* Risk factors and decision criteria for intensive chemotherapy in older patients with acute myeloid leukemia. *Haematologica* **93**, 1806-1813 (2008). https://doi.org/10.3324/haematol.13309

9 Kantarjian, H. *et al.* Intensive chemotherapy does not benefit most older patients (age 70 years or older) with acute myeloid leukemia. *Blood* **116**, 4422-4429 (2010). https://doi.org/10.1182/blood-2010-03-276485

10 Pastore, F. *et al.* Combined molecular and clinical prognostic index for relapse and survival in cytogenetically normal acute myeloid leukemia. *J Clin Oncol* **32**, 1586-1594 (2014). https://doi.org/10.1200/JCO.2013.52.3480

11 DiNardo, C. D. & Cortes, J. E. Mutations in AML: prognostic and therapeutic implications. *Hematology Am Soc Hematol Educ Program* **2016**, 348-355 (2016). https://doi.org/10.1182/asheducation-2016.1.348

12 Daneshbod, Y., Kohan, L., Taghadosi, V., Weinberg, O. K. & Arber, D. A. Prognostic Significance of Complex Karyotypes in Acute Myeloid Leukemia. *Curr Treat Options Oncol* **20**, 15 (2019). https://doi.org/10.1007/s11864-019-0612-y

13 Estey, E. H. Acute myeloid leukemia: 2021 update on risk-stratification and management. *Am J Hematol* **95**, 1368-1398 (2020). https://doi.org/10.1002/ajh.25975

14 Sasaki, K. *et al.* Prediction of survival with intensive chemotherapy in acute myeloid leukemia. *Am J Hematol* **97**, 865-876 (2022). https://doi.org/10.1002/ajh.26557

15 Eckardt, J. N. *et al.* Prediction of complete remission and survival in acute myeloid leukemia using supervised machine learning. *Haematologica* **108**, 690-704 (2023). https://doi.org/10.3324/haematol.2021.280027

16 Dohner, H. *et al.* Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* **129**, 424-447 (2017). https://doi.org/10.1182/blood-2016-08-733196

17 Dohner, H. *et al.* Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. *Blood* **140**, 1345-1377 (2022). https://doi.org/10.1182/blood.2022016867
18 Hoff, F. W. *et al.* Beat-AML 2024 ELN-Refined Risk Stratification for Older Adults with Newly Diagnosed AML Given Lower-Intensity Therapy. *Blood Adv* (2024). https://doi.org/10.1182/bloodadvances.2024013685
19 Dohner, H. *et al.* Genetic Risk Stratification and Outcomes Among Treatment-Naive Patients With AML Treated With Venetoclax and Azacitidine. *Blood* (2024). https://doi.org/10.1182/blood.2024024944
20 Dohner, H. *et al.* Genetic risk classification for adults with AML receiving less-intensive therapies: the 2024 ELN recommendations. *Blood* (2024). https://doi.org/10.1182/blood.2024025409
21 Bataller, A. *et al.* Prognostic risk signature in patients with acute myeloid leukemia treated with hypomethylating agents and venetoclax. *Blood Adv* **8**, 927-935 (2024). https://doi.org/10.1182/bloodadvances.2023011757
22 Dohner, H. *et al.* ELN Risk Stratification Is Not Predictive of Outcomes for Treatment-Naïve Patients with Acute Myeloid Leukemia Treated with Venetoclax and Azacitidine. *Blood* **140 (Supplement 1)**, 1441-1444 (2022).
23 Islam, N. *et al.* Machine Learning-Based Exploratory Clinical Decision Support for Newly Diagnosed Patients With Acute Myeloid Leukemia Treated With 7 + 3 Type Chemotherapy or Venetoclax/Azacitidine. *JCO Clin Cancer Inform* **6**, e2200030 (2022). https://doi.org/10.1200/CCI.22.00030
24 Austin, P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res* **46**, 399-424 (2011). https://doi.org/10.1080/00273171.2011.568786
25 Ditzhaus, M. & Friedrich, S. More powerful logrank permutation tests for two-sample survival data. arXiv:1807.05504 (2018). <https://ui.adsabs.harvard.edu/abs/2018arXiv180705504D>.
26 Fleming, T. R., Harrington, D. P. & O'Sullivan, M. Supremum Versions of the Log-Rank and Generalized Wilcoxon Statistics. *Journal of the American Statistical Association* **82**, 312-320 (1987). https://doi.org/10.1080/01621459.1987.10478435
27 Gorfine, M., Schlesinger, M. & Hsu, L. K-sample omnibus non-proportional hazards tests based on right-censored data. *Stat Methods Med Res* **29**, 2830-2850 (2020). https://doi.org/10.1177/0962280220907355
28 TARONE, R. E. & WARE, J. On distribution-free tests for equality of survival distributions. *Biometrika* **64**, 156-160 (1977). https://doi.org/10.1093/biomet/64.1.156
29 Tian, L., Zhao, L. & Wei, L. J. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics* **15**, 222-233 (2014). https://doi.org/10.1093/biostatistics/kxt050
30 Uno, H. *et al.* Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol* **32**, 2380-2385 (2014). https://doi.org/10.1200/JCO.2014.55.2208
31 Horiguchi, M. & Uno, H. On permutation tests for comparing restricted mean survival time with small sample from randomized trials. *Stat Med* **39**, 2655-2670 (2020). https://doi.org/10.1002/sim.8565
32 Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw* **39**, 1-13 (2011). https://doi.org/10.18637/jss.v039.i05

33 Gareth, J., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning*. Vol. 112 (Springer, 2013).
34 Xu, L., Gotwalt, C., Hong, Y., King, C. B. & Meeker, W. Q. Applications of the Fractional-Random-Weight Bootstrap. *The American Statistician* **74**, 345-358 (2020). https://doi.org/10.1080/00031305.2020.1731599
35 Stoltz, D. & Williams, R. (London, 2020).
36 Zhang, Z. *et al.* Estimate risk difference and number needed to treat in survival analysis. *Ann Transl Med* **6**, 120 (2018). https://doi.org/10.21037/atm.2018.01.36
37 Marsaglia, G., Tsang, W. W. & Wang, J. Evaluating Kolmogorov's Distribution. *Journal of Statistical Software* **8**, 1 - 4 (2003). https://doi.org/10.18637/jss.v008.i18
38 Pollyea, D. A. *et al.* Venetoclax and azacitidine followed by allogeneic transplant results in excellent outcomes and may improve outcomes versus maintenance therapy among newly diagnosed AML patients older than 60. *Bone Marrow Transplant* **57**, 160-166 (2022). https://doi.org/10.1038/s41409-021-01476-7
39 Seaman, S. R. & White, I. R. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res* **22**, 278-295 (2013). https://doi.org/10.1177/0962280210395740
40 van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45**, 1 - 67 (2011). https://doi.org/10.18637/jss.v045.i03
41 Rubin, D. *Multiple imputation for nonresponse in surveys*. (John Wiley & Sons, Inc., 1987).
42 Mustafa Ali, M. K. *et al.* Outcomes of Newly Diagnosed Acute Myeloid Leukemia Patients Treated With Hypomethylating Agents With or Without Venetoclax: A Propensity Score-Adjusted Cohort Study. *Front Oncol* **12**, 858202 (2022). https://doi.org/10.3389/fonc.2022.858202
43 Gangat, N. *et al.* Venetoclax and hypomethylating agent combination therapy in newly diagnosed acute myeloid leukemia: Genotype signatures for response and survival among 301 consecutive patients. *Am J Hematol* **99**, 193-202 (2024). https://doi.org/10.1002/ajh.27138
44 Othman, J. *et al.* Real-world outcomes of newly diagnosed AML treated with venetoclax and azacitidine or low-dose cytarabine in the UK NHS. *Blood Neoplasia* **1**, 100017 (2024). https://doi.org/https://doi.org/10.1016/j.bneo.2024.100017
45 Islam, N. *et al.* A simplified risk model for pretreatment stratification of newly diagnosed acute myeloid leukemia patients treated with venetoclax and azacitidine. *medRxiv*, 2024.2012.2002.24318344 (2024). https://doi.org/10.1101/2024.12.02.24318344

TABLE LEGEND

Table 1. Model specifications for risk stratification models with respect to censoring assumptions, methodologies to deal with missing data, end points, variable selection rules, and feature sets. Primary models were denoted by (M1) – (M16) and reported results were based on (M2) and (M8) (highlighted in red).

Table 1. Specifications of risk stratification models.

Variations in specifications due to missing data + censoring + end-points

	Missing data adjustment	Censoring	Survival
(M1)	Category	FAS	OS
(M2)	Category	allo-HCT as censored	OS
(M3)	CC with IPW	FAS	OS
(M4)	CC with IPW	allo-HCT as censored	OS
(M5)	Imputation by MICE	FAS	OS
(M6)	Imputation by MICE	allo-HCT as censored	OS
(M7)	Category	FAS	EFS
(M8)	Category	allo-HCT as censored	EFS
(M9)	CC with IPW	FAS	EFS
(M10)	CC with IPW	allo-HCT as censored	EFS
(M11)	Imputation by MICE	FAS	EFS
(M12)	Imputation by MICE	allo-HCT as censored	EFS
(M13)	Ensemble of (M1, M3, M5)	FAS	OS
(M14)	Ensemble of (M2, M4, M6)	allo-HCT as censored	OS
(M15)	Ensemble of (M7, M9, M11)	FAS	EFS
(M16)	Ensemble of (M8, M10, M12)	allo-HCT as censored	EFS

Covariate-level classification rules

[**Rule-I**] Use point-estimates to compute mean marginal risks for each variable at t

[**Rule-II**] Use point-estimates to compute mean marginal risks for each variable at t. We excluded variables that are sparsely (< 2% (+ve) cases) or overly (> 40% (+ve) cases) represented.

[**Rule-III**] Use bootstrap-based point-estimates averaged over 2000 runs to compute mean marginal risks for each variable at t.

[**Rule-IV**] Combining Rule-II and Rule-III

Type of markers

[**Type-1**] Genetic

[**Type-2**] Genetic + Phenotypic

Total risk stratification models (covariate and subject level)

16 X 4 X 2 = 128

Remarks:
CC – Complete case analysis
IPW – Inverse probability of weights
MICE – Multiple imputation by chained equations
FAS – Full analytical dataset
allo-HCT – allogeneic hematopoietic cell transplant

FIGURE LEGENDS

Figure 1. Summary of the de-novo AML ven/aza treated patient cohort. The flow diagram summarizes the patient populations in the CU dataset and how they were utilized in generating the risk stratification models for OS/BR and EFS.

Figure 2. Graphical representation of the machine-learning methodology used to develop risk models with respect to overall survival and event free survival. Details of each step 1-7 are described in the Methods section.

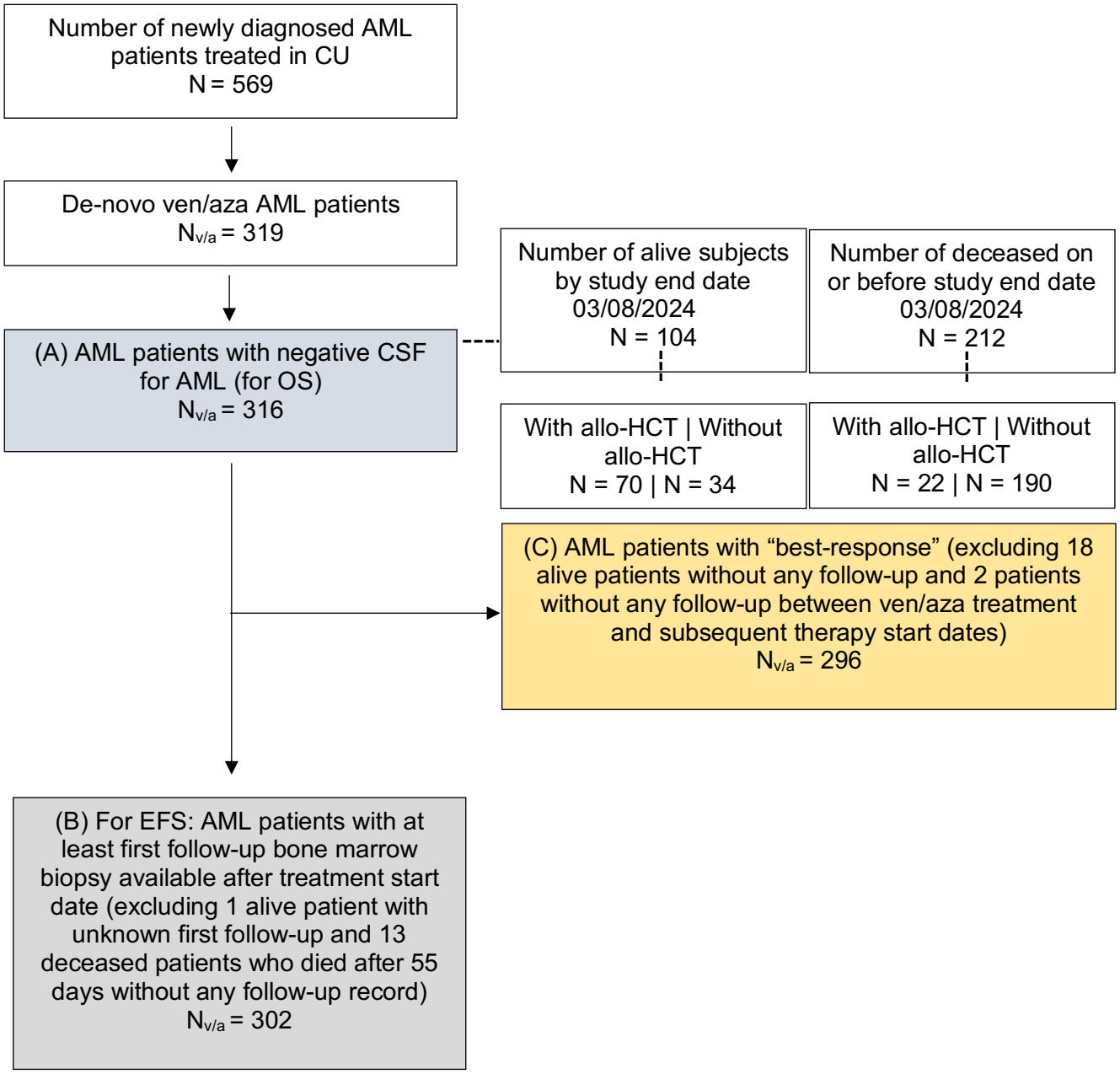
Figure 3. Internal validation of predictive performance for proposed risk stratification variables. Architecture of internal validation steps assessing predictive performance for proposed risk stratification variables.

Figure 4. Overall survival (OS) following treatment based on the genetic features-based risk model. The genetic features-based covariate-level classification Rule-II risk model ($RM_{G,II}$) is depicted. A) Definitions of the feature identification strategy and genetic features in $RM_{G,II}$; B) OS (left panel) and best response (BR) (right panel) based on $RM_{G,II}$ applied to the cohort excluding allo-HCT recipients; C) OS (left panel) and BR (right panel) based on $RM_{G,II}$ applied to the full analytical set (FAS); and D) $RM_{G,II}$ based on OS treating allo-HCT recipients as censored (left panel) and within the cohort comprised solely of allo-HCT recipients. Note BCOR, DNMT3A, DDX41, SF3B1, KRAS, KMT2A, PTPN11, SMC1A, JAK2 are mutations defined by NGS. IDH1, IDH2, NPM1 and p53m are defined by composite molecular tests as defined in Methods and Supplemental Table 1E for “composite mutation features”. Chr 7 abn:loss, Chr 8 abn:loss, DEK:loss, MECOM:gain, RARA:loss are FISH features. The remainder including Del7q, T(9;11), Minus5, Minus17, Complex cytogenetics are karyotype features.

Figure 5. Event free survival (EFS) based on the genetic model using covariate-level classification. The covariate-level classification Rule-II Risk Model based on genetic features and specific to EFS ($RM_{G,II\text{-}EFS}$) is depicted**.** A) Feature identification strategy and definitions of the features of $RM_{GP,II\text{-}EFS}$; B) EFS based on $RM_{GP,II\text{-}EFS}$ applied to the full analytical set (FAS) (left panel) and set excluding (right panel) allo-HCT recipients; and C) EFS based on $RM_{GP,II\text{-}EFS}$ treating allo-HCT recipients as censored (left panel) and only within the cohort comprised of allo-HCT recipients. Note BCOR, KRAS, KMT2A, PTPN11, JAK2, were mutations defined by NGS. IDH1, IDH2, NPM1 and p53m are defined by composite molecular tests as defined in Methods and Supplemental Table 1E for "composite mutation features", Chr 7 abn:loss, Chr 8 abn:loss, RARA:loss, MECOM:rearrangement, KMT2A:rearrangement are FISH features. The remainder including Del 5q Del7q, T(9;11), Minus5, Minus17, Complex cytogenetics, Indeterminate cytogenetics and Normal cytogenetics are karyotype features.

Figure 6. Heatmap of risk stratification. Subject-specific risk stratification by various risk models (16 models) described in this report are summarized in the two panels and are denoted as Adverse risk (black), Intermediate risk (red), and Favorable risk (green).

Figure 1. Summary of the de-novo AML ven/aza treated patient cohort.

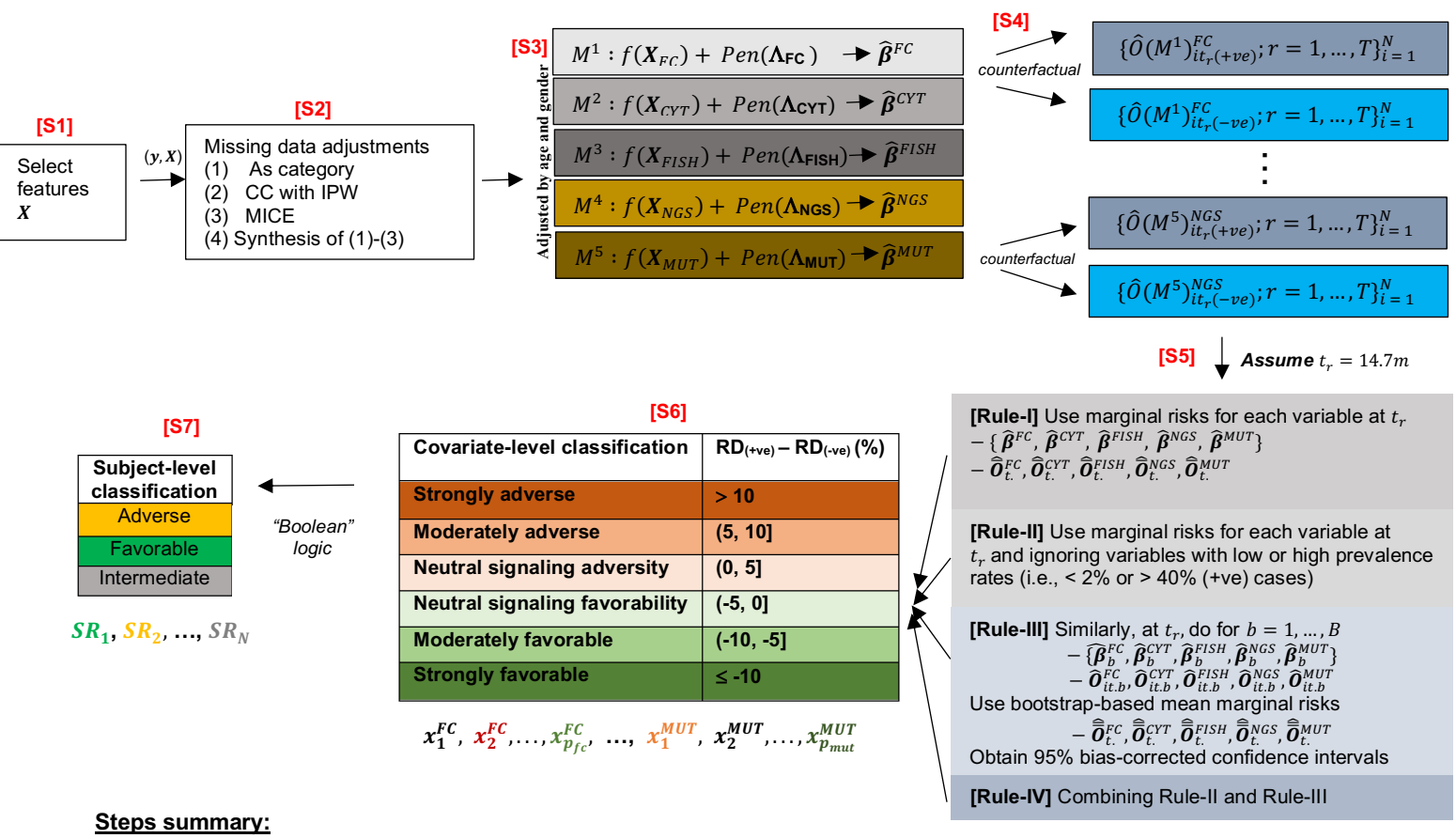


(A) (Uni)-Multivariate hazard and counterfactual risk-stratification model for overall survival (OS)
(B) (Uni)-Multivariate hazard and counterfactual risk-stratification model for event free survival (EFS)
(C) Validation analyses by transporting risk models on best responses.

i. **Overall survival** (OS): Time from ven/aza treatment initiation to all-cause mortality
ii. **Event free survival** (EFS): Time from ven/aza treatment initiation to the first documented "relapse from CR/CRi/MLFS", "no response", "partial remission", and all-cause mortality whichever occurs first
iii. **Best response** (BR): Defined as the best response recorded within the cycles containing the first treatment course of ven/aza.

Figure 2. Strategy and methodologies used to develop risk models with respect to overall and event-free survival.

i : Subscript defines a subject; $i = 1, \ldots, N$
j : Superscript defines type of OMICs data (i.e., FC, CYT, FISH, NGS, AML mutation)
b : Subscript denotes a bootstrap sample; $b = 1, \ldots, B$
$(\boldsymbol{c}, \boldsymbol{t}, \boldsymbol{X})$: Observed data where $\boldsymbol{y} = (\boldsymbol{c}, \boldsymbol{t})$ refers to survival data and $\boldsymbol{X} \subseteq \mathcal{X}$: Set of features obtained from feature space $\mathcal{X}$
t_r : Grid of time-points of interest; $r = 1, \ldots, T$
M^j: j-th Cox-PH model with ridge (i.e., $L2$ norm) penalty
β^j: Parameter estimates
$\boldsymbol{\Lambda}_j$: Set of tuning parameters corresponding to M^j-th model
$\hat{O}(M)^j_{it(+ve)}$: Marginal estimation of relative risks for i-th subject at t-th time based on M^j-th model for (+ve)
$\hat{\bar{O}}(M)^j_{t(+ve)}$: Mean marginal estimation of relative risks at t-th time based on M^j-th model for (+ve)
$\hat{\bar{\bar{O}}}(M)^j_{t(+ve)}$: Mean marginal estimation of relative risks at t-th time averaged over B bootstrap runs based on M^j-th model for (+ve)



Steps summary:

[S1] Select features (i.e., FC, CYT, FISH, NGS, AML mutation) based on univariate KMs with respect to OS, EFS, MRD, and RFS under three censoring mechanisms (for allo-HCT) and get $\boldsymbol{X} = \{\boldsymbol{X}_{FC}, \boldsymbol{X}_{CYT}, \boldsymbol{X}_{FISH}, \boldsymbol{X}_{RD}, \boldsymbol{X}_{MUT}\}$ from $\mathcal{X}$.

[S2] Adjust for missing data via three methods. IPWs were estimated based on GLM for complete case analyses. Missing elements were imputed five times via MICE exploiting $\boldsymbol{X}$ under "*impute then-select*" approach for each biomarker.

[S3] Fit mCOXr for each type of biomarkers separately. Corresponding tuning parameters are selected via cross-validation.

[S4] Mean marginal probabilities of risk over time were obtained utilizing "*counterfactual*" arguments for each subject hypothetically had each biomarker been (+ve) and (-ve).

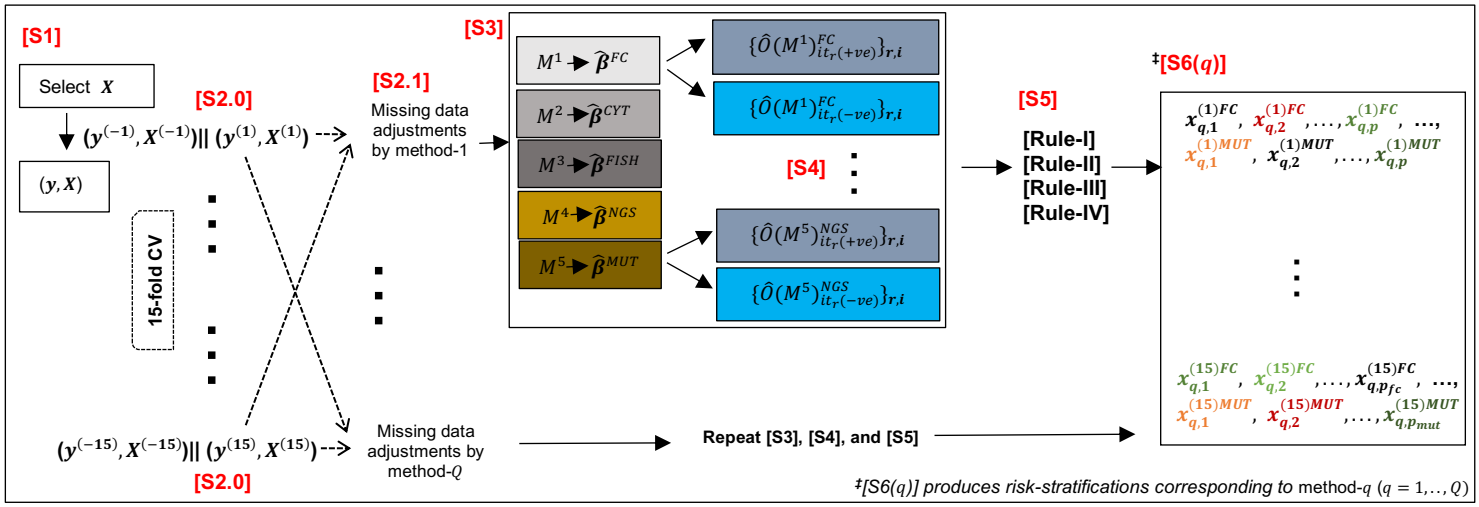
[S5] Obtain bootstrap-based 95% CI for estimates, marginal risks, and risk differences.

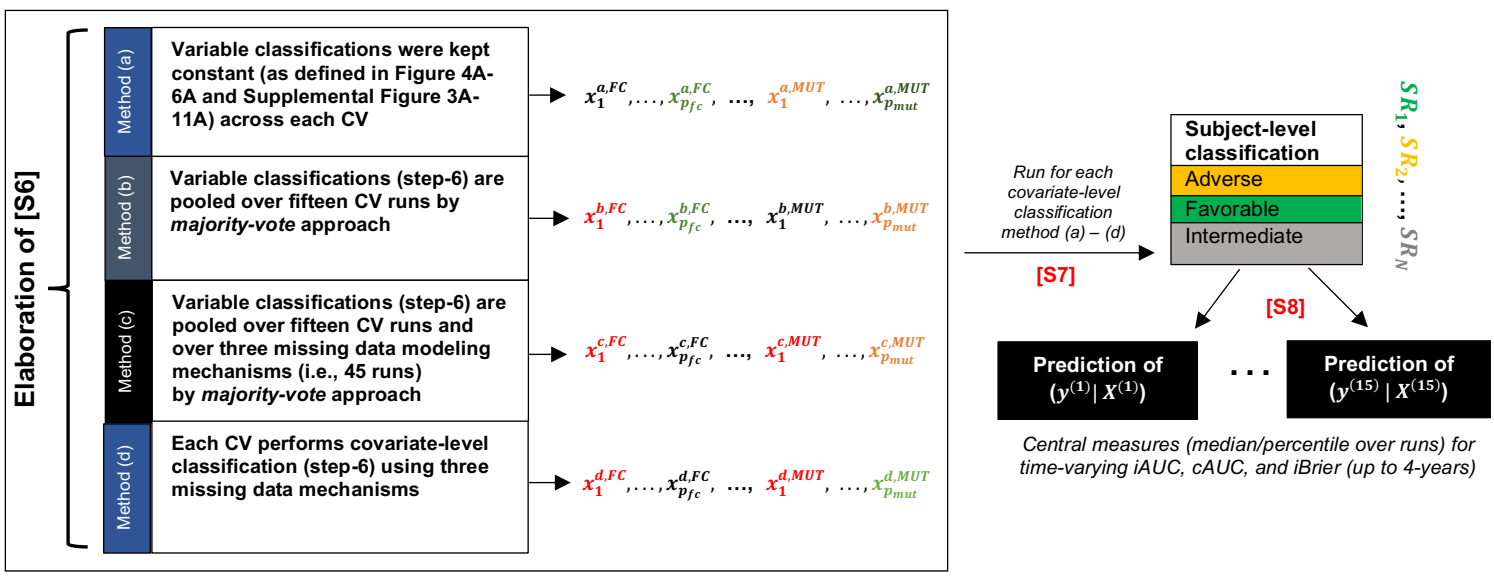
[S6] Calculate risk differences $((1/N)\sum_i\{\hat{O}(M^1)^{FC}_{it_{14.7}(+ve)} - \hat{O}(M^1)^{FC}_{it_{14.7}(-ve)}\}$ were calculated at t_r. Next obtain covariate-level stratification using 5% risk difference rules.

[S7] Apply logic to determine subject specific risk stratification.

Figure 3. Internal validation of predictive performance of risk models.

i : Subscript defines a subject; $i = 1, \ldots, N$
j : Superscript defines type of OMICs data (i.e., FC, CYT, FISH, NGS, AML mutation)
b : Subscript denotes a bootstrap sample; $b = 1, \ldots, B$
$(\boldsymbol{c}, \boldsymbol{t}, \boldsymbol{\mathcal{X}})$: Observed data where $\boldsymbol{y} = (\boldsymbol{c}, \boldsymbol{t})$ refers to survival data and $\boldsymbol{X} \subseteq \boldsymbol{\mathcal{X}}$: Set of features obtained from feature space $\boldsymbol{\mathcal{X}}$
v : Subscript denotes a cross-validation sample; $v = 1, \ldots, V$
t_r : Grid of time-points of interest; $r = 1, \ldots, T$
$(\boldsymbol{y}^{(v)}, \boldsymbol{X}^{(v)})$: Observed data leaving v-th CV fold out
M^j: j-th Cox-PH model with ridge (i.e., $L2$ norm) penalty
β^j: Parameter estimates
$\boldsymbol{\Lambda_j}$: Set of tuning parameters corresponding to M^j-th model
$\hat{O}(M)^j_{it(+ve)}$: Marginal estimation of relative risks for i-th subject at t-th time based on M^j-th model with respect to (+ve)
$\hat{\bar{O}}(M)^j_{t(+ve)}$: Mean marginal estimation of relative risks at t-th time based on M^j-th model with respect to (+ve)
$\hat{\bar{\bar{O}}}(M)^j_{t(+ve)}$: Mean marginal estimation of relative risks at t-th time averaged over B bootstrap based on M^j-th model with respect to (+ve)
SR_i : Subject-level risk stratification for i-th subject





Steps summary:

[S1-S7] Steps 1-7 are as same as before. Note 15-fold CV is performed by splitting datasets into mutually exclusive training and validation sets [S2.0].

[S8] Step 8 evaluates predictive performance over fifteen CV runs by different covariate-level classification approaches.

Figure 4. Overall survival (OS) based on the genetic features specific risk model using Rule-II ($RM_{G,II}$).

A. Feature identification strategy and $RM_{G,II}$ features.

Covariate-level risk classification

$RD_{(+ve)} - RD_{(-ve)}$	%
Strongly adverse	> 10
Moderately adverse	(5, 10]
Slightly adverse	(0, 5]
Slightly favorable	(-5, 0]
Moderately favorable	(-10, -5]
Strongly favorable	≤ -10

Subject-level risk stratification

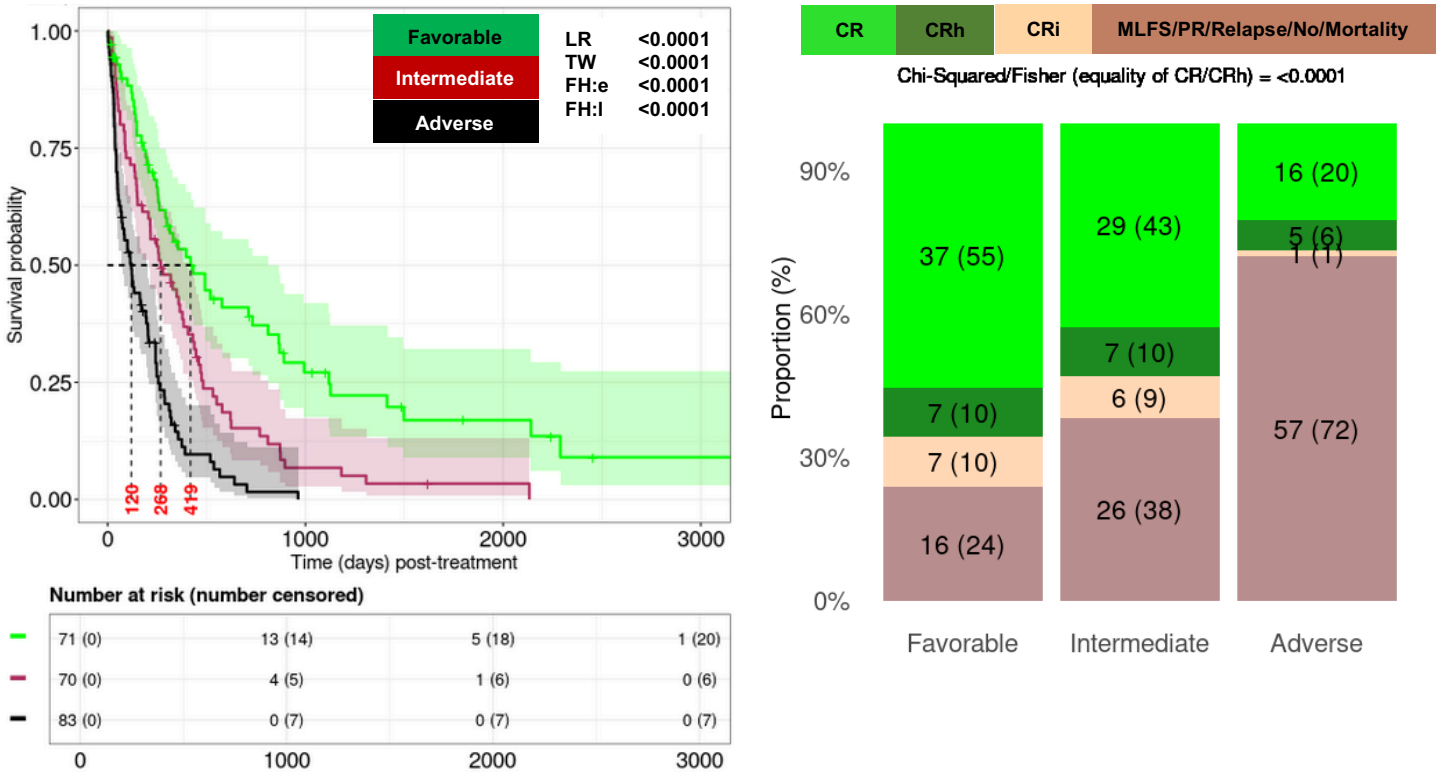
Groups	Definitions
Adverse	Having at-least one 'Strongly adverse' and/or 'Moderately adverse' risk factor in the absence of any 'Strongly favorable' and/or 'Moderately favorable' factor
Favorable	Having at-least one 'Strongly favorable' and/or 'Moderately favorable' factor in the absence of any 'Strongly adverse' and 'Moderately adverse' risk factor.
Intermediate	Otherwise

Category	Genetic features	Frequency†	Frequency¶
Favorable	Any of ***BCOR, DNMT3A, DDX41, SF3B1, IDH1, IDH2, NPM1*** (in the absence of any Adverse features)	**71/224 (32%)**	**101/316 (32%)**
Intermediate	No Favorable or Adverse features, or ≥1 Favorable features with ≥1 Adverse features	**70/224 (31%)**	**106/316 (34%)**
Adverse	Any of ***KRAS, KMT2A, PTPN11, SMC1A, JAK2, p53m, Del7q, t(9;11), Minus5, Minus17, Complex cytogenetics, Chr 7 abn:loss, Chr 8 abn:loss, DEK:loss, MECOM:gain, RARA:loss*** (in the absence of any Favorable features)	**83/224 (37%)**	**109/316 (34%)**

† Counts (proportions) of patients after excluding allo-HCT
¶ Counts (proportions) of patients without excluding allo-HCT

Figure 4, cont'd

B. $RM_{G,II}$ model for OS (left) and best response (right), excluding allo-HCT recipients.



C. $RM_{G,II}$ model for OS (left) and best response (right), with respect to the FAS.

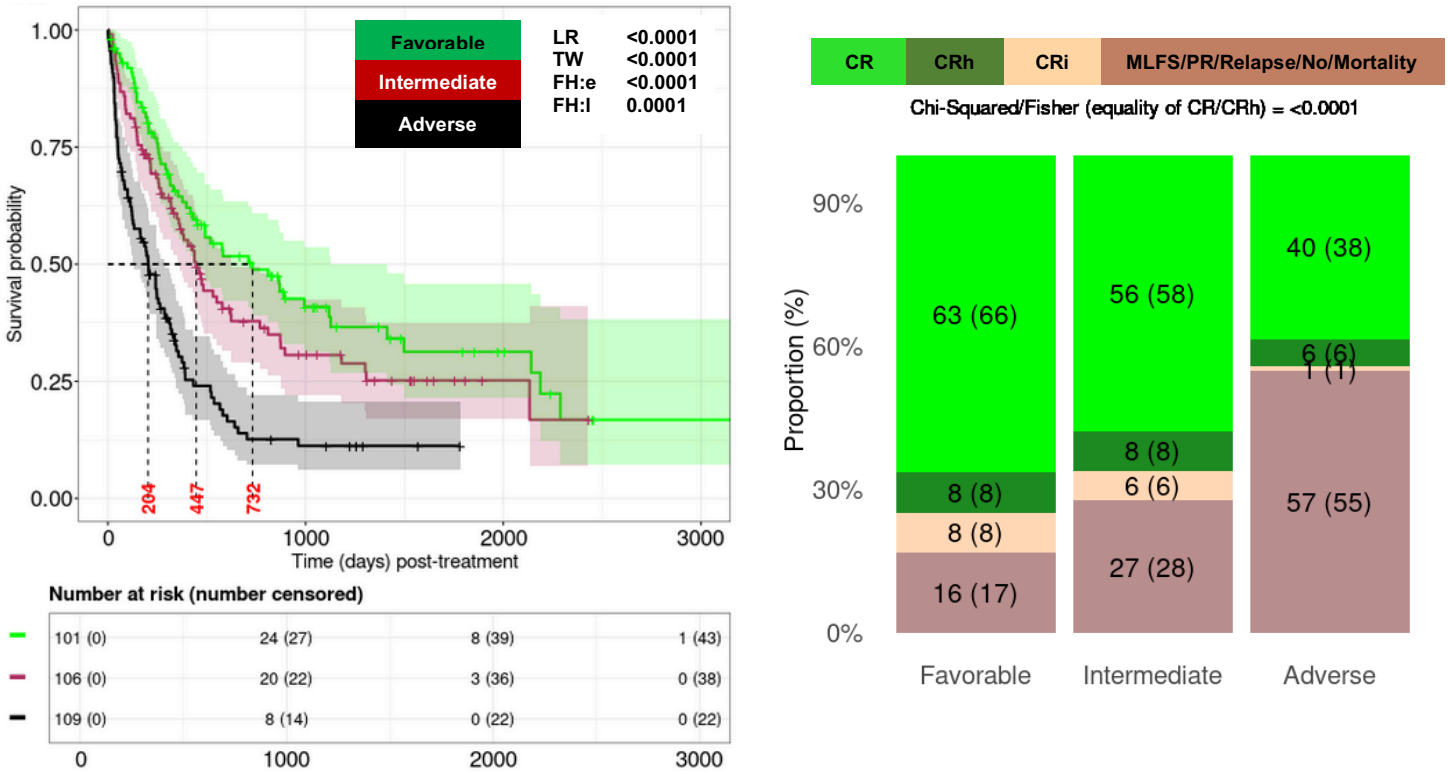


Figure 4, cont'd

D. $RM_{G,II}$ on OS censoring allo-HCT recipients (left) and only within allo-HCT cohort. (right).

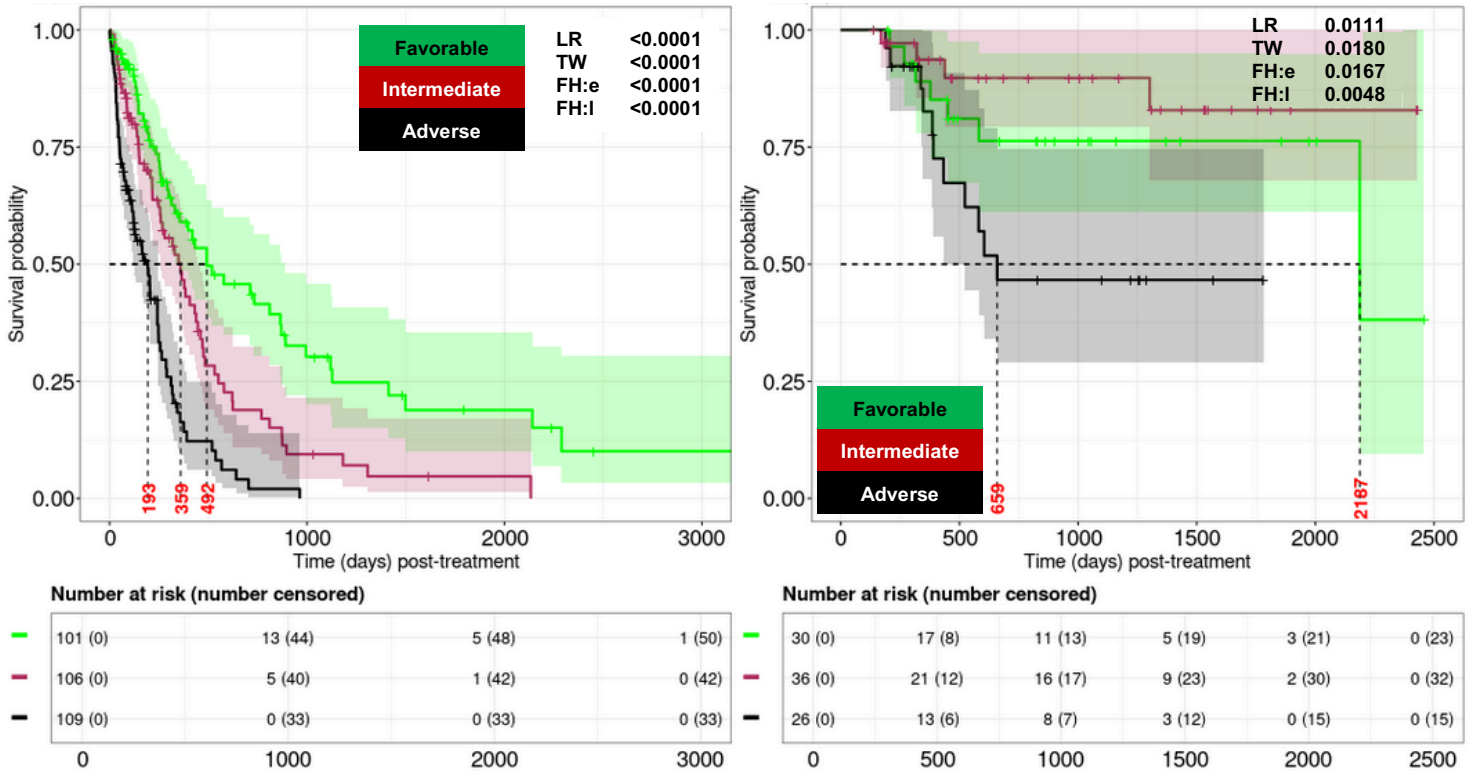


Figure 5. Event free survival (EFS) based on the genetic features specific risk model using Rule-II ($RM_{G,II\text{-}EFS}$).

A. Feature identification strategy and $RM_{G,II\text{-}EFS}$ features.

Covariate-level risk classification

$RD_{(+ve)} - RD_{(-ve)}$	%
Strongly adverse	> 10
Moderately adverse	(5, 10]
Slightly adverse	(0, 5]
Slightly favorable	(-5, 0]
Moderately favorable	(-10, -5]
Strongly favorable	≤ -10

Subject-level risk stratification

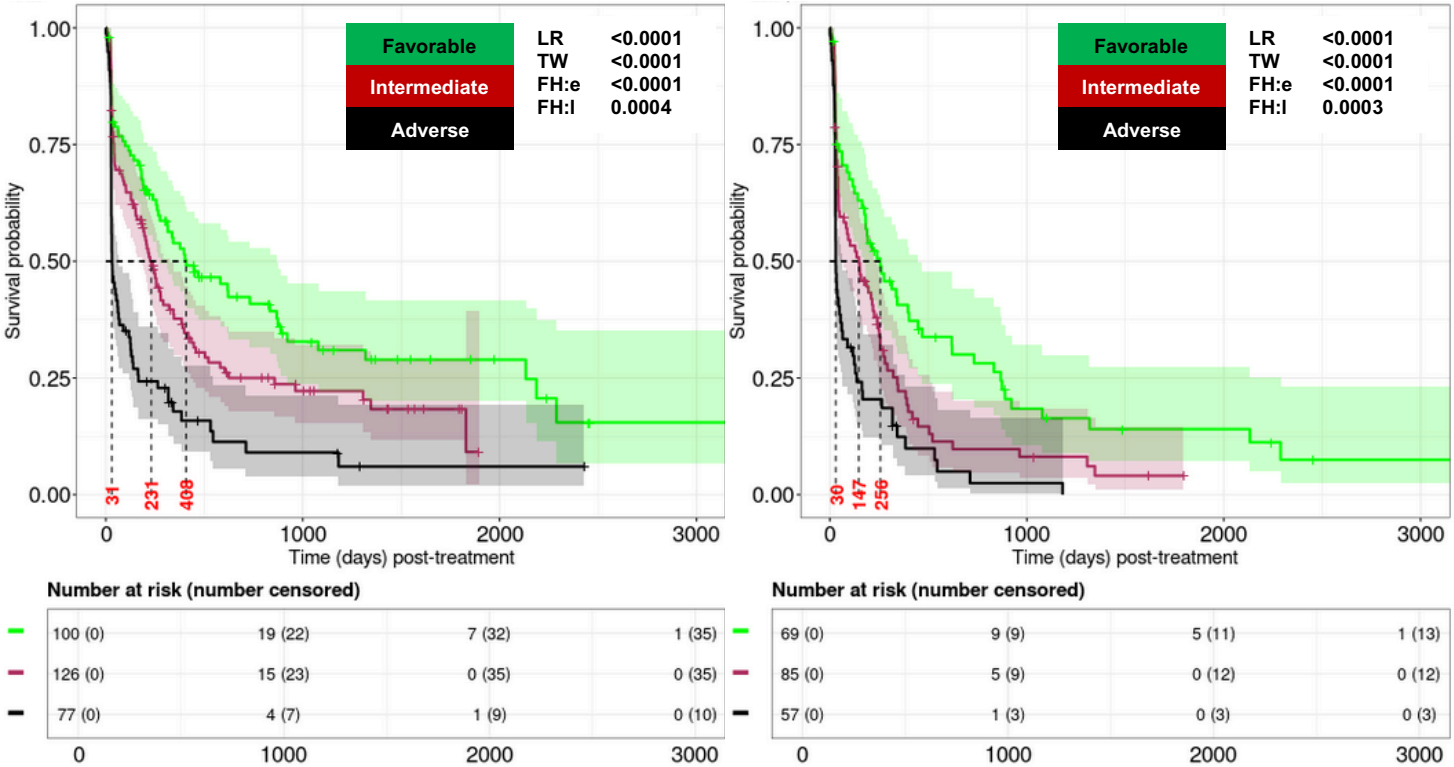
Groups	Definitions
Adverse	Having at-least one 'Strongly adverse' and/or 'Moderately adverse' risk factor in the absence of any 'Strongly favorable' and/or 'Moderately favorable' factor
Favorable	Having at-least one 'Strongly favorable' and/or 'Moderately favorable' factor in the absence of any 'Strongly adverse' and 'Moderately adverse' risk factor.
Intermediate	Otherwise

Category	Genetic features	Frequency†	Frequency¶
Favorable	Any of ***BCOR, Normal cytogenetics, Monosomal karyotype, Del5q, IDH1, IDH2, NPM1*** (in the absence of any Adverse features)	**69/211 (33%)**	**100/303 (33%)**
Intermediate	No Favorable or Adverse features, or ≥1 Favorable features with ≥1 Adverse features	**85/211 (40%)**	**126/303 (42%)**
Adverse	Any of ***KRAS, KMT2A, PTPN11, JAK2, p53m, Del7q, t(9;11), Minus5, Minus17, Complex cytogenetics, Indeterminant cytogenetics, Chr 7 abn:loss, Chr 8 abn:loss, RARA:loss MECOM:rearrangement, KMT2A:rearrangement*** (in the absence of any Favorable features)	**57/211 (27%)**	**77/303 (25%)**

† Counts (proportions) of patients after excluding allo-HCT
¶ Counts (proportions) of patients without excluding allo-HCT

Figure 5, cont'd

B. $RM_{G,II\text{-}EFS}$ on EFS, with respect to the FAS (left) and dataset excluding allo-HCT patients (right).



C. $RM_{G,II\text{-}EFS}$ on EFS censoring allo-HCT (left) and only within allo-HCT recipients (right).

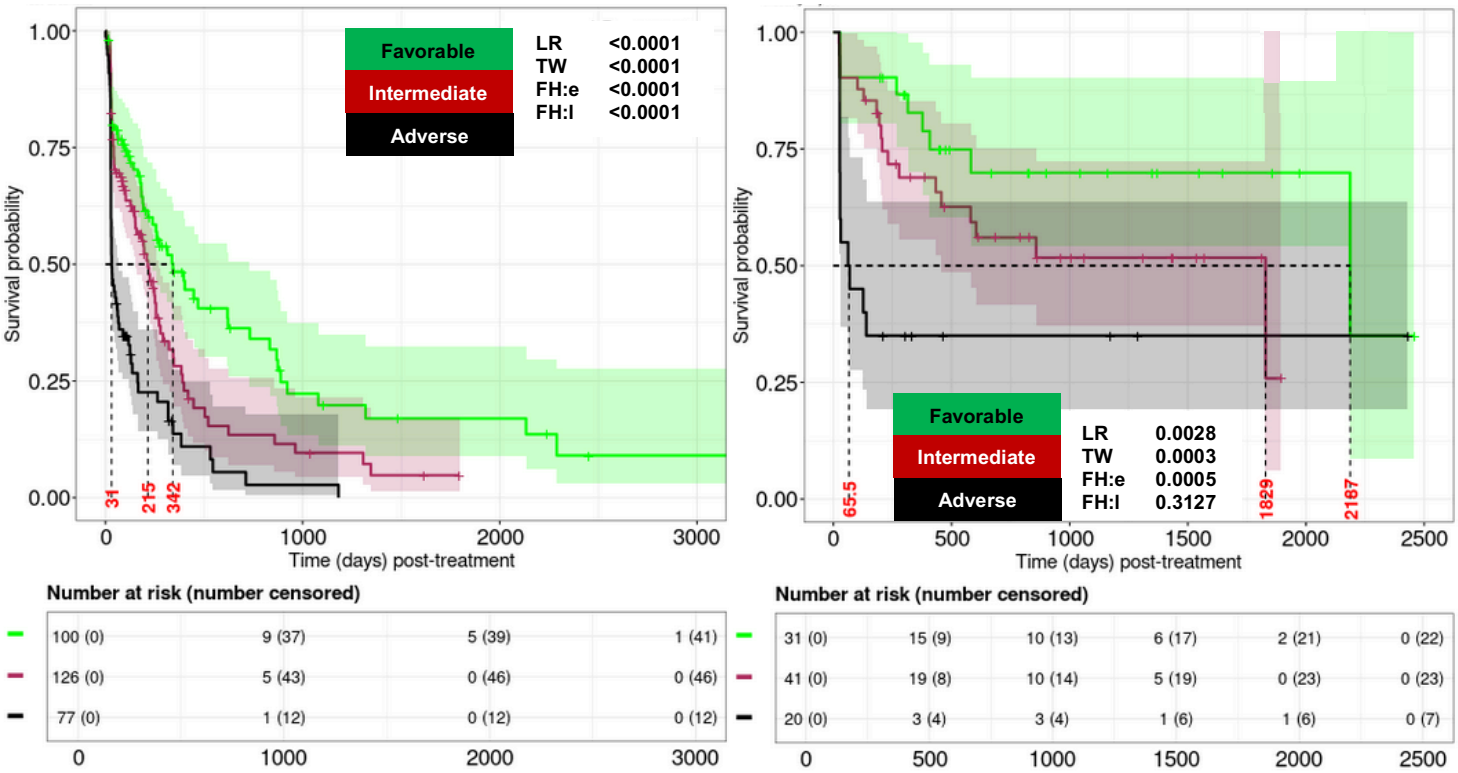
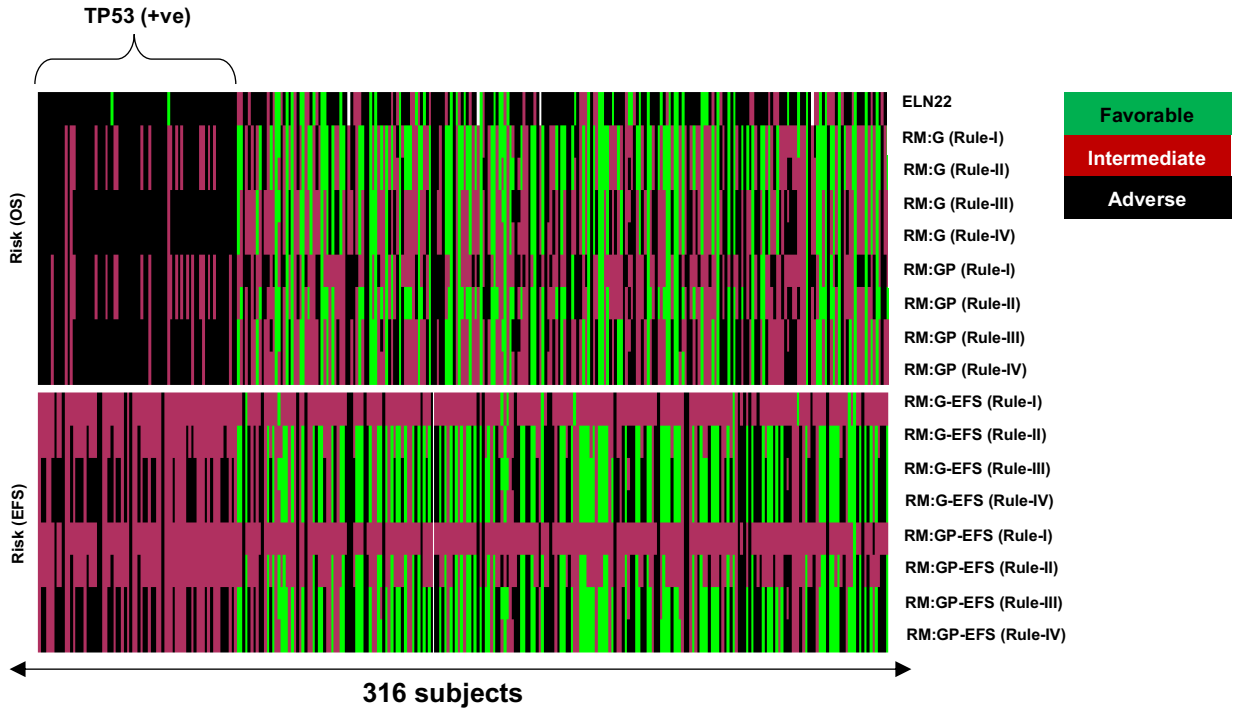



Figure 6. Subject-specific risk stratification of AML patients by risk models.