# 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

**Context Summary:** 100

**Abstract count:** 243

**Word count:** 3292 (Tables: 1, Figures: 6)

**Article type:** Original Report (Research and Applications)

## CONTEXT SUMMARY

### Key Objective

Can an effective machine learning based risk stratification strategy be developed for Acute Myeloid Leukemia (AML) that address real world challenges and varying use cases?

### Knowledge Generated

A RM strategy for AML patients treated with venetoclax/azacitidine was developed based on a dynamic counterfactual machine learning (ML) model. This strategy efficiently generated specific RMs that demonstrated robust performance with varying assumptions, data elements, and endpoints, and was adaptable to an external real world cohort dataset. These findings demonstrate that ML-based RMs are feasible and may have potential advantages over existing AML RM strategies in adapting to context dependent settings.

## ABSTRACT

### Objective

The objective of this study was to develop a flexible risk stratification strategy for Acute Myeloid Leukemia (AML) that is specific for venetoclax plus azacitidine (ven/aza), addresses real-world data (RWD) issues, and is also adaptable to different use cases.

### Materials and Methods

A series of tunable risk models (RMs) were generated from a dynamic counterfactual machine learning (ML) strategy. These utilized a range of features from diagnostic AML samples and were tested using objective metrics on a single institution cohort of 316 newly diagnosed patients treated with ven/aza. RM performance was tested using various model assumptions, data elements, and endpoints, and with applications to an external AML real world cohort (RWC).

### Results

Favorable, Intermediate, and Adverse risk groups were identified in a series of ML-based RMs using different assumptions, for genetic only or genetic-plus-phenotypic data elements and with overall survival and event free survival as endpoints. Most RMs demonstrated equitable patient distribution (~20%-40% in each risk group), significant separation between risk strata (Log-rank based P-values <0.001), and predictability computed by time-dependent survival AUC values of 0.60-0.70. Similar performance was observed when the proposed RM strategy was adapted and compared to the ELN22 using the external RWC.

### Discussion and Conclusion

The proposed ML strategy addresses a variety of RWD considerations and is readily tunable through coding and parameter updates for different contexts and use case needs. This strategy represents a novel approach to developing more effective RMs for AML and possibly other diseases.

## INTRODUCTION

Acute myeloid leukemia (AML) is an aggressive hematopoietic malignancy diagnosed in ~20,000 individuals annually in the United States[1]. For young and fit patients, initial treatment typically involves intensive chemotherapy (IC), followed by consolidation with allogeneic hematopoietic cell transplant (allo-HCT) or high-dose chemotherapy[2,3]. For older and less-fit patients, venetoclax (ven) combined with a hypomethylating agent (HMA) such as azacitidine (aza) is now a standard of care[4-6]. For IC therapies, various 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 risk categorization approach for AML in 2017 (ELN17), with an update in 2022 (ELN22), which has been widely used[16,17]. ELN categories are divided into Favorable, Intermediate, and Adverse risk groups associated with overall survival (OS) based on AML cytogenetic and next-generation sequencing features[16,17]. However, the ELN categorization approaches were recently found to perform sub-optimally with ven/HMA-treated patients, likely because the ELN criteria are primarily based on treatment outcomes following IC[18]. Based on this, a variety of rapidly evolving new AML risk models (RMs) have been developed with differing risk features to better stratify patients treated with lower intensity agents including ven/HMA[19-24]. In addition to risk features in different RMs varying with treatment, they may also vary according to a variety of other contextual considerations including feature prevalence, missingness and sparsity of feature data, varying endpoints, collinearity among features, testing technologies generating the features, and the impact of allo-HCT and other important confounding effects on outcomes interpretation[21-25]. Furthermore, risk features are contextually dependent on the group being studied, which may vary substantially between academic, clinical trial, community, and other settings; additionally, genetic traits can also pose different risks based on a patient's genetic background[19,26]. Lastly, RMs' features and performance vary based on specific use case requirements. For example, up-front patient allocation in clinical trials may prioritize features that can be rapidly resulted, are easy to interpretate, and minimize missingness, while post hoc retrospective analyses can incorporate more complex features that take longer to result.

To address these issues, here we describe the development and initial testing of a flexible AML risk stratification strategy for newly diagnosed AML patients. A machine-learning (ML) based strategy was selected for its efficient adaptability to address various real-world data (RWD) considerations and its readiness to adapt to different datasets and use cases. This strategy was developed and initially tested for performance using multiple objective metrics on a single academic center cohort specifically treated with ven/aza. It was then further evaluated for adaptability and generalizability to other use cases using an external real-world cohort (RWC).

## METHODS

### Internal analytical patient population and outcomes definitions

The internal analytical cohort dataset included 316 adult patients from the University of Colorado (CU) with newly diagnosed AML treated with ven/aza between January 2015 and March 2024. Details of the patient cohort and subset management are shown in Figure 1. Patient demographics are summarized in Supplemental Table 1A. Key AML diagnostic pathology features for developing the RMs are summarized in Supplemental Tables 1B-F, including next generation sequencing (NGS), cytogenetics (CYT), fluorescence in-situ hybridization (FISH), polymerase chain reaction (PCR), flow cytometric (FC) data, and "AML mutations," which included composite mutations from various tests. Further details are provided in Supplemental Methods S1-S2 and Supplemental Table 2. Responses (OS, event-free survival (EFS), best response (BR)) are defined in Figure 1[17,27]. This retrospective study was approved by an institutional review board (IRB) and used a limited dataset with a waiver of consent.

### Exploratory analyses

Standardized mean differences[28] and Kaplan-Meier (KM) analyses for baseline covariates were performed assuming right censoring with respect to OS and EFS based on a Full Analytical Set (FAS), a dataset censoring allo-HCT recipients, and a dataset excluding allo-HCT recipients. The equality of survival curves was evaluated using various testing approaches[29-35] and noise variables were filtered to select a parsimonious feature list; further details are provided in Supplemental Methods S3. The resulting list of confounding variables (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 shows the RM development strategy. Key components include the following.

***Survival regressions***

ML models (i.e., multivariable penalized (ridge) Cox proportional hazard (mCOXr)[36,37] regressions) were fitted adjusting for age and gender for OS and EFS. Separate models were fitted adjusting for confounding variables selected by null hypothesis significance testing (NHST) (Step 1, Figure 2) to identify influential variables within each group (Step 3, Figure 2). Missing values were treated as separate categories (Step 2, Figure 2). Tuning parameters were selected via cross-validation (CV). Corresponding 95% confidence intervals (CIs) of adjusted hazard ratios (aHRs) were constructed using the fractional random weight bootstrap (FRWB)[38] approach of 2,000 runs (Supplemental Table 3-4).

***Counterfactual effects estimation***

Biomarker-specific mean predicted marginal risks (MRs) over time were calculated using "*recycled*" predictions with counterfactual arguments for the corresponding mCOXr models (Step 4, Figure 2)[39,40]; see Supplemental Method S4 for details. Bootstrap-based MR profiles were constructed by averaging the FRWB estimates, and the corresponding percentile-based 95% CIs were reported. The Kolmogorov-Smirnov test assessed the differences in counterfactual risk distributions[41]. Multivariable penalized (ridge) logistic regressions[37] were used to model BR confirming the consistency of the mCOXr estimates.

***Testing variations in risk models***

*Treatment variation:* Allo-HCT is a potentially curative therapy that can profoundly affect ven/aza treated patient outcomes[42]. To address this confounding factor, RMs were developed by censoring patients with allo-HCT.

*Varying endpoints*: RMs for OS and EFS were developed separately.

*Dynamic behavior:* Counterfactual risk differences (RDs) were computed dynamically over 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 feature-specific RD for OS and EFS based on the median time-to-event in the VIALE-A study, which established ven/aza as a new standard of care[6].

*Prevalence*: Variables with low and high prevalence for (+ve) cases may cause unstable parameter estimates and imbalances in subject-level stratification. To address this, rules with varying degrees of stringency were incorporated. Rule-I (Step 5, Figure 2) is flexible and does not exclude features with respect to prevalence. Rule-II excluded any variables containing less than 2% and more than 40% (+ve) cases in risk stratification even if they exhibited more than 5% RD. Rule-III used FRWB based MR profiles to calculate RD, and Rule-IV combined Rule-II and Rule-III (see Supplemental Table 5A).

*Covariate-level classification:* Two levels of classification, covariate and subject, were performed. The magnitude and direction of marginal RDs between (+ve) and (-ve) at a time point of interest were used to classify each feature into either “Adverse” or “Favorable”. Positive RD values of 0-5%, 5-10%, and >10% classified a feature as *slightly*, *moderately*, or *strongly* "Adverse," respectively. Similarly, negative RD values determined *slightly*, *moderately*, or *strongly* "Favorable" features (Step 6, Figure 2). A cut-off value of 5% was assumed to be the clinically relevant minimum RD. Supplemental Tables 5B-5E depict covariate-level classifications under different combinations of parameters.

*Subject-level classification*: Patient-specific risk stratification ("Adverse", "Intermediate", and "Favorable") was determined by combining covariate-level classifications using a 'Boolean' operator (AND/OR). This assigns a patient to the "Favorable" risk group without any "Adverse" features and vice versa. A patient was assigned to the "Intermediate" risk group without any "Favorable" and "Adverse" features or with at least one "Favorable" and one "Adverse" feature (Step 7, Figure 2).

*Model specification*: RMs using genetic and genetic-plus-phenotypic features were developed to reflect different scenarios and their implications on model performance (Table 1).

*Missingness:* Primary analyses treated missing values as a separate category for simplicity, avoiding imputation or exclusion. Reported results were based on this approach unless otherwise specified. Additional analyses were performed to check the robustness of different missing data handling techniques at the covariate-level classification step including: (A) Biases stemming from complete cases (CC) analyses were adjusted by incorporating inverse probability weights (IPW) into mCOXr, where subject-specific weights signaling the propensity of having any missing data were estimated from

multivariable logistic regression models[43]. (B) Multiple imputations (five) were generated for each missing data element by exploiting multivariate data sequentially (multivariate imputation by chained equations (MICE))[44,45]. The corresponding aHRs were averaged over five samples and 2,000 FRWB runs, and percentile-based 95% CIs were reported. Subsequently, the marginal RDs were calculated as before and averaged over five samples. (C) Covariate-level classifications were performed based on different missing methods, with final stratification obtained by synthesizing with respect to *majority-votes*.

**Evaluation of numerical performance**

Multiple evaluation criteria were used to provide a holistic view of numerical performance in terms of *equitability* (extent of risk group size variation), *separability* (extent of outcome variation by risk strata), *conformity* (extent of risk group overlap between methods), and *predictability* (degree of prognostic relationship with outcomes). Internal validation of *predictability* of the RMs 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 the Supplemental Methods S5-S6.

**Application of risk models on an external analytical dataset**

A set of specific RMs for OS was tested on an independent and heterogeneous RWC comprised of 971 AML patients treated with ven/aza at 87 unique sites of care. The RWC was derived from the nationwide Flatiron Health electronic health record-derived, de-identified, longitudinal database consisting of patient-level structured and unstructured de-identified data curated via technology-enabled abstraction[46,47]. Missing values in the RWC were generated by *imputation-by-mode* approach; further evaluation with respect to MICE was performed for sensitivity. Additional details are discussed in Supplementary Methods S7-S10.

All statistical tests, except for the comparison of MR curves, predicted curves, and survival AUC values, were two-sided with 5% significance level. As the analyses were exploratory, no multiplicity adjustments were made. All analyses were performed using R, version 4.2.3.

## RESULTS

The overall strategy for the ML-based AML RMs developed and tested here is summarized in Figure 2. By applying this strategy, we first assessed how the varying prevalence rules detailed in Table 1 influenced model performance. Rule I was identified as the least stringent, whereas Rule IV was determined to be the most stringent (Supplemental Figure 5A). The initial RMs were populated with genetic features only (termed $RM_G$) and derived from the CU analytical cohort. Performance was assessed with respect to *equitability, separability, conformity,* and *predictability.* The $RM_G$ that used covariate-level classification Rule-II (termed $RM_{G,II}$) demonstrated superior performance in both *equitable* distribution of patients and OS *separability* compared to Rules I, III, and IV (Figures 4A-4F, Supplemental Figure 5 and Supplemental Table 6B). $RM_{G,II}$ demonstrated similar performance characteristics in the allo-HCT excluded, FAS, and allo-HCT recipients censored cohorts for OS as well as BR (Figures 4B-4D). In the cohort including only allo-HCT recipients, 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). Covariate-level classifications of the $RM_G$ variants using Rule-I ($RM_{G,I}$), Rule-III ($RM_{G,III}$), and Rule-IV ($RM_{G,IV}$) are presented in Supplemental Tables 5B-5E. In general, these varying models demonstrated competitive performance in terms of *separability* and *conformity*. However, they showed reduced *equitability* in assigning patients to different risk categories, which further led to lower *predictability* in risk assessment (Supplemental Figures 2-4, Supplemental Tables 6A, 6C, and 6D). An additional evaluation of the performance of the ML-based RM strategy was conducted by comparison to ELN22. The latter demonstrated a relatively skewed distribution of patients assigned to the Adverse risk category and differences in separation of OS curves for both the allo-HCT excluded cohort and the FAS (Supplemental Figure 5A-C). Inter-group conformity analyses using Fleiss kappa demonstrated the highest level of agreement in patient assignment for the Favorable risk category and predictive evaluations (time-varying AUCs and iBrier) favored the $RM_{G,II}$ model over ELN22 (Supplemental Figure 5E-F, Supplemental Table 6B) which is consistent with other recent reports demonstrating the limitations of ELN22 for non-IC treated patients[18,19,21-24].

To test the adaptability of this RM strategy to a dataset with more expanded data elements, we explored the impact of including both genetic and phenotypic (flow cytometry) data on RM performance (termed $RM_{GP}$). The impacts of applying Rule-I, II, III, and IV and the various Models in Table 1 to $RM_{GP}$ were tested as described above. The specific features in the Adverse, Intermediate, and Favorable groups defined by the ML-based strategy varied from the $RM_G$ models (Supplemental Figures 6-9, Supplemental Tables 6A-6D). In general, the overall performance of the $RM_{GP}$ models with these different features paralleled those of the $RM_G$ models, but in some cases with less *equitable* assignment of patients to the Favorable risk category depending on models, rules, and assumptions.

To test whether a new RM could be developed using the ML-based strategy for an endpoint other than OS, we developed and explored the RMs using the EFS endpoint[48]. Again, the specific risk features for Adverse, Intermediate, and Favorable risk populations varied from the OS endpoint-based models (Figure 5A) further confirming the possibility that RMs would be context dependent. Covariate-level classification Rule-II specific RM for EFS using genetic features only (i.e., $RM_{G,II\text{-}EFS}$) demonstrated the best performance (Figure 5B-5C) and superior *predictability* (Supplementary Table 7B). In general, other EFS RMs based on the various Rules as described above demonstrated performance roughly similar to that of OS (Supplemental Tables 7A-7D, Supplemental Figures 10-17). Supplemental Figure 18 presents heatmaps that illustrate the variations in risk stratification across the different rules, endpoints, assumptions, and models, underscoring the necessity for context-specific RMs.

Lastly, we assessed the adaptability of the RM strategy to a different use case than retrospective analysis of an academic center cohort. The strategy was adapted to retrospective risk stratification of patients in a different external dataset (the “RWC”) derived from patients treated primarily in community centers. This RWC had differences in populations and data availability as described in the main and Supplementary Methods. These RWC specific RMs, termed RWC-RMs were tested across the different Rules for performance characteristics as above. A subset of the original $RM_{G,II}$ risk features for Adverse,

Intermediate, and Favorable risk populations was identified based on available features in the RWC (Figure 6A). Supplementary Figures 20-21 illustrate the distribution of patients with observed outcomes across risk groups for the RWC-$RM_{G,II}$ and ELN22. The RWC-$RM_{G,II}$ demonstrated somewhat less *equitable* distribution than the parental $RM_{G,II}$ using the academic center cohort but with strong *separation* of OS in both the RWC FAS and RWC allo-HCT excluded recipients (Figure 6A-6B). The RWC-$RM_{G,II}$ showed lower empirical survival probabilities for the Adverse risk group than ELN22 while achieving significant conformity with ELN22 in assigning patients to this group (Figure 6C-6D). For OS risk *prediction*, it demonstrated a higher cAUC and iAUC than an ELN22-based approach, with median cAUC values of 0.62 (95% CI: 0.60, 0.75) versus 0.52 (95% CI: 0.50, 0.59); see Figure 6E-6F. The performance for RM incorporating Rules I, III, and IV for the RWC as well as a comparison to ELN22 roughly paralleled those for the academic cohort albeit again with different specific risk features identified for each RWC-RM (Supplemental Figures 22-25). This further demonstrates the context-specific adaptability and transportability of the proposed strategy.

## DISCUSSION

We developed a novel ML-based risk stratification strategy for AML patients that was designed to address a variety of RWD challenges and be flexible and adaptable to varying use cases. Having a common strategy is important for RM development to provide a consistent, objective framework for developing RMs across different populations, data elements, and applications. Unique features of the ML strategy include: (1) it was based on probabilistic RDs, (2) dealt with feature interactions using a logic-based approach, (3) addressed dynamic effects of features, (4) supported development of different specific RMs for different endpoints, (5) could account for important confounding effects (e.g., allo-HCT), and 6) included various features adjusting for collinearity. We deployed multiple objective metrics to assess internal performance and cross model comparisons to define the performance of RMs derived from this strategy. These metrics provide insights into model-specific effectiveness, while demonstrating the impact of modeling assumptions and rules on risk classifications. Testing this strategy showed: 1) proof of principle for using counterfactual dynamic estimation to develop RMs in AML, 2) confirmation that risk features depend on context, 3) adaptability of RMs to different modeling assumptions and rules, 4) adaptability to varying data types, endpoints, and use cases, and 5) ability to tune and generate specific RMs through coding, parameter, and feature modifications to accommodate changes in evolving diagnostics, assumptions, and rules. This capability is particularly important in AML, where advances in clinical care, treatment, disease biology, and testing modalities have yielded rapidly changing datasets. At a population level, unlike ELN22, the ML-based strategy effectively generated RMs tailored to specific use cases that demonstrated robust performance across the range of metrics and comparators used in both single academic center and RWC for retrospective analysis. Figure 6G summarizes how this strategy may be deployed to generate RMs for a variety of other potential use cases occurring before, during, and after treatment, including rapid pre-therapy patient allocation and more deliberate post-therapy retrospective analysis for discovery work (see S11 for additional details). To support testing and development, the code for applying this strategy to various datasets and applications has been made publicly accessible; refer to the Code Sharing section.

Please note that this code is intended for research and exploratory purposes only; extensive testing and validation on specific RMs is essential before applying them in practice.

There are several limitations to this study. The simplification of genetic, FC, FISH, and composite AML mutation features into binary variables for numerical convenience and interpretability may lead to information loss and overlook critical biological aspects of AML. Similarly, treating all composite mutation categories for RUNX and other features (Supplemental Table 1E) as equally biologically relevant may lead to an oversimplification. Reintroducing the complexity of the constituents of these features may enhance RM stratification ability, offering additional clinical insights; however, future studies with larger datasets are needed for validation of such insights. Although RMs integrating genetic-and-phenotypic features were developed, lack of FC standardization across centers indicates that the models may be most applicable in single-center or small consortium settings with consistent FC datasets. There is also a risk of underreporting samples with dimly expressed FC features (e.g., CD45). Some RMs showed suboptimal stratification for allo-HCT recipients, which may be improved by developing RMs with features specific to this cohort. Using the same dataset for variable selection, risk classifications, and internal validation may cause overfitting. Additionally, all models, particularly those with higher stringency, risk omitting important features due to underestimated effects, small samples, and prespecified cut-offs for prevalence rates and RDs. In contrast, less stringent RMs may incorporate noisy features with spurious associations. To optimize RM's performance, it is essential to not only address the above limitations but also to employ a flexible non-parametric approach, such as tree- or deep-learning-based survival models, coupled with a robust feature selection scheme. Lastly, this model was solely based on ven/aza treated patients so its further transportability to other treatments, including different lower intensity therapies or ven/aza based triplets, will need to be independently tested. Additional testing and validation with external, diverse datasets from multiple institutions, through prospective studies, and via comparison with published findings and emerging RMs are critical next steps to further evolve this strategy and the RMs generated from it[21,23,25].

## CONCLUSION

In summary, we developed a novel ML-based risk stratification strategy for AML patients treated with ven/aza. We generated and tested a range of specific RMs that exhibited robust performance across various data elements and endpoints with respect to important metrics, demonstrating adaptability to a distinct external RWC use case. This indicates the proposed strategy's effectiveness and the RMs' ability to generalize across different contexts. This ML-based strategy may also be applicable to other AML use cases and diseases beyond AML.

## 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 OncoVerity, Aurora, Colorado; and Grant Weller from 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 methods and 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

We simulated a synthetic survival dataset consisting of 200 subjects, distinct from the original analytical set, to illustrate the proposed strategy for developing risk models. This simulated dataset was used to implement the methodology, and the implementation details are provided in an HTML file included as a supplementary document. The corresponding R codes are incorporated within the HTML file and are also available in a public GitHub repository https://github.com/RefinedScience/Risk_Stratification.

## 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 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).
19 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** (2024). https://doi.org/10.1016/j.bneo.2024.100017
20 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* **8**, 5297-5305 (2024). https://doi.org/10.1182/bloodadvances.2024013685
21 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
22 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
23 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
24 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
25 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
26 Stiff, A. *et al.* Multiomic profiling identifies predictors of survival in African American patients with acute myeloid leukemia. *Nat Genet* (2024). https://doi.org/10.1038/s41588-024-01929-x
27 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
28 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
29 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>.
30 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
31 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
32 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

33 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
34 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
35 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
36 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
37 Gareth, J., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning*. Vol. 112 (Springer, 2013).
38 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
39 Stoltz, D. & Williams, R. (London, 2020).
40 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
41 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
42 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
43 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
44 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
45 Rubin, D. *Multiple imputation for nonresponse in surveys*. (John Wiley & Sons, Inc., 1987).
46 Ma, X., Long, L., Moon, S., Adamson, B. J. S. & Baxi, S. S. Comparison of Population Characteristics in Real-World Clinical Oncology Databases in the US: Flatiron Health, SEER, and NPCR. *medRxiv*, 2020.2003.2016.20037143 (2023). https://doi.org/10.1101/2020.03.16.20037143
47 Birnbaum, B. *et al.* Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. *arXiv preprint arXiv:2001.09765* (2020).
48 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

## 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 + endpoints*

| | 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.

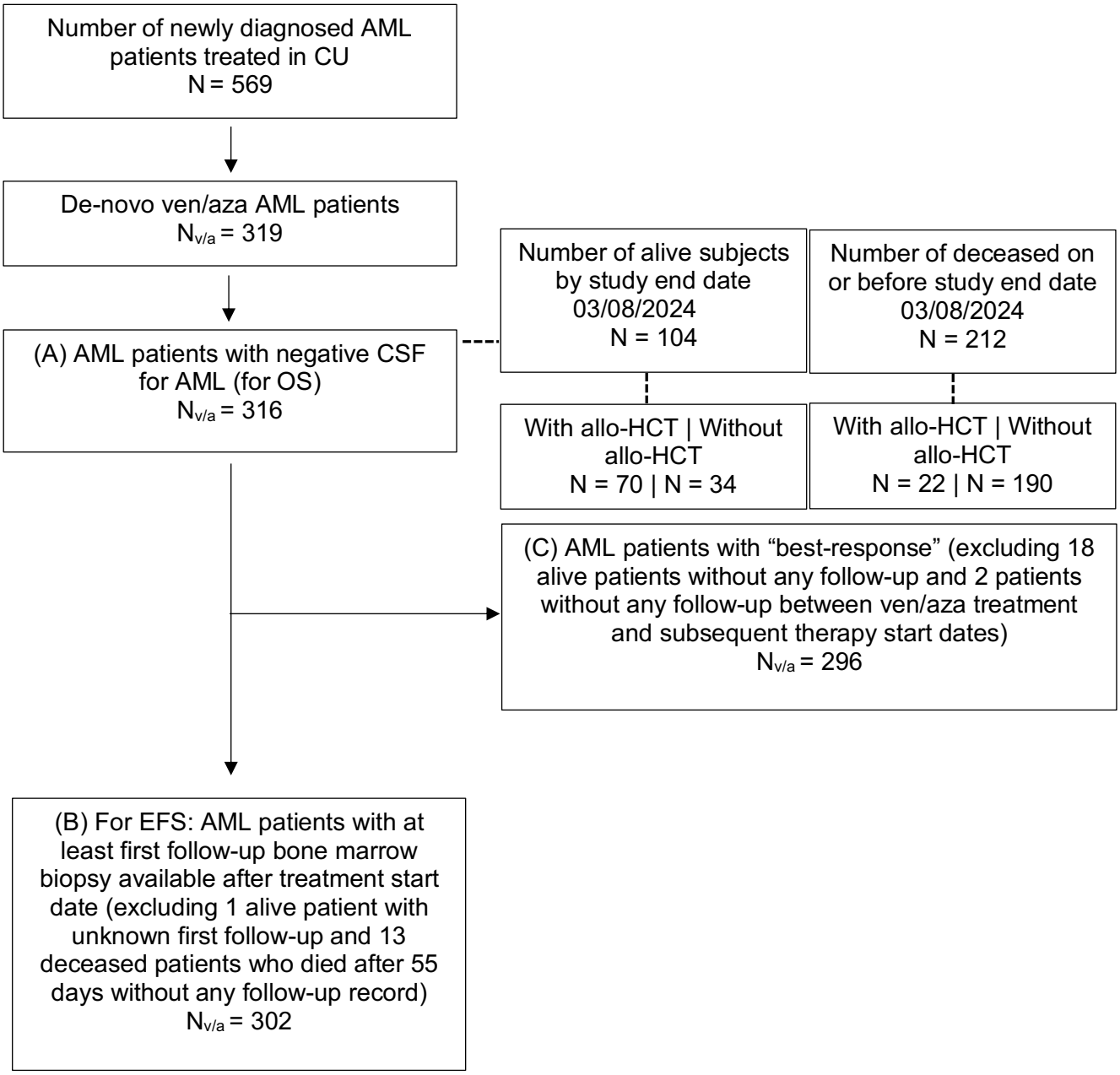
**Figure 4. Overall survival (OS) following treatment based on the genetic features-based risk model.** The genetic features-based covariate-level classification Rule-II risk model ($RM_{G,II}$) is depicted. A) Definitions of the feature identification strategy and genetic features in $RM_{G,II}$. The category column includes color legends for reference; 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.

**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) Definitions of features associated with the $RM_{G,II\text{-}EFS}$. The category column includes color legends for reference; B) EFS based on the $RM_{G,II\text{-}EFS}$ applied to the full analytical set (FAS) (left panel) and set excluding (right panel) allo-HCT recipients; and C) EFS based

on the $RM_{G,II\text{-}EFS}$ treating allo-HCT recipients as censored (left panel) and only within the cohort comprised of allo-HCT recipients.

**Figure 6. External validation for the $RM_{G,II}$ on the real-world cohort (RWC) with respect to overall survival (OS).** A) Definitions of features for the RWC-$RM_{G,II}$. The category column includes color legends for reference. Biomarkers such as *BCOR, SF3B1, SMC1A, KMT2A, PTPN11, Chr 8 abn:loss, DEK:loss, MECOM:gain,* and *RARA:loss* were unavailable in the RWC. *Chromosome 7 deletion* (i.e., *Minus 7*) was used as a proxy of *Chromosome 7q deletion* (i.e., *Del7Q*); B) OS based on the RWC-$RM_{G,II}$ with respect to the full analytical set (FAS) (left) and the set excluding allo-HCT (right) patients; C) Pairwise comparison between the ELN 2012 and RWC-$RM_{G,II}$ for the Adverse, Intermediate, and Favorable risk groups in the cohort excluding allo-HCT patients; D) Comparison of inter-risk group conformity between the RWC-$RM_{G,II}$ and ELN22 based on the FAS; E) Predictability of OS based on the RWC-$RM_{G,II}$ and ELN22 over follow-up time up to four years. Composite molecular definitions used in classifying risk groups in the RWC are defined in Supplemental Methods section; F) Pairwise comparisons of predicted survival probability curves between the ELN 2012 and RWC-$RM_{G,II}$ for the Adverse, Intermediate, and Favorable risk groups; G) A flowchart illustrating the practical application of the ML strategy to different potential use cases, note the availability and requirements of risk stratification features may vary greatly across the different use cases leading to the need for context dependent specific RMs.

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

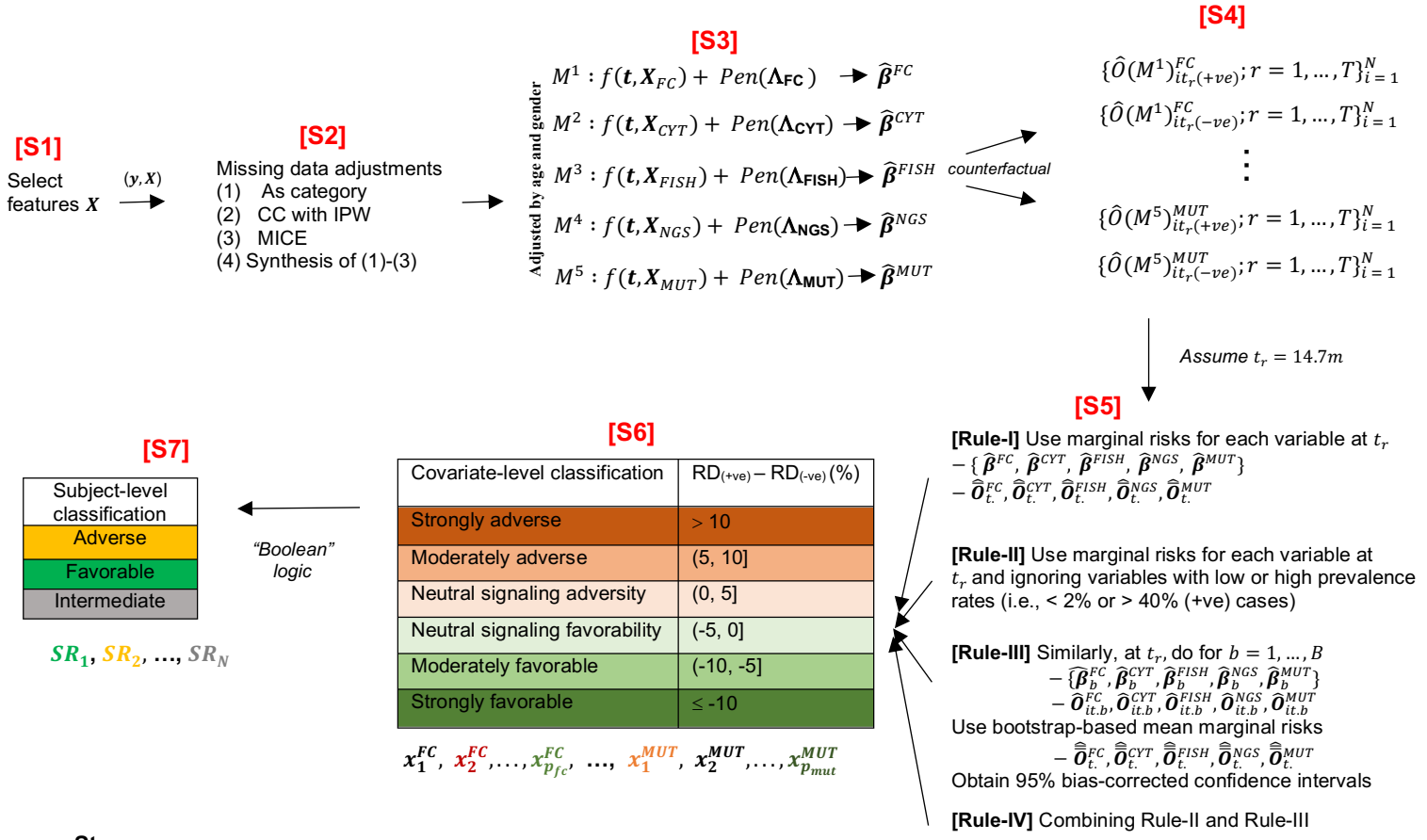



(A) (Uni)-Multivariate hazard and counterfactual risk-stratification model for overall survival (OS)
(B) (Uni)-Multivariate hazard and counterfactual risk-stratification model for event free survival (EFS)
(C) Validation analyses by transporting risk models on best responses.

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 survival (OS) and event-free survival (EFS).**

$i$ : 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 $\boldsymbol{\mathcal{X}}$
$t_r$ : Grid of time-points of interest; $r = 1, \ldots, T$
$M^j$: $j$-th hazard function for Cox-PH model with ridge (i.e., $L2$ norm) penalty
$\beta^j$: 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 $\boldsymbol{\mathcal{X}}$.

**[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).

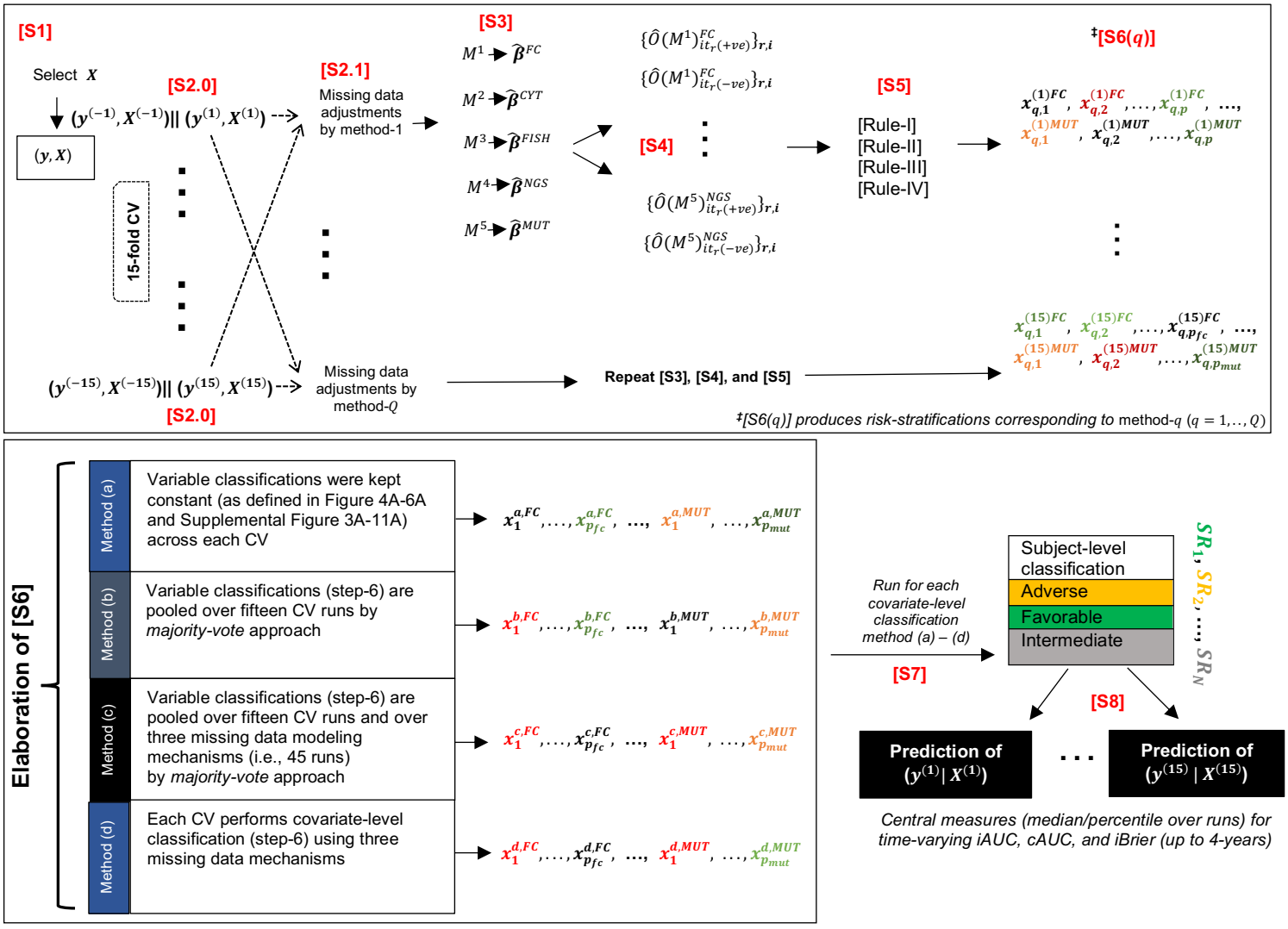
**[S5]** Obtain bootstrap-based 95% CI for estimates, marginal risks, and risk differences based on different rules.

**[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
$\beta^j$: 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)
$\hat{\hat{O}}(M)^j_{t(+ve)}$: Mean marginal estimation of relative risks at $t$-th time based on $M^j$-th model with respect to (+ve)
$\hat{\hat{\hat{O}}}(M)^j_{t(+ve)}$: 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 ($RM_{G,II}$).**

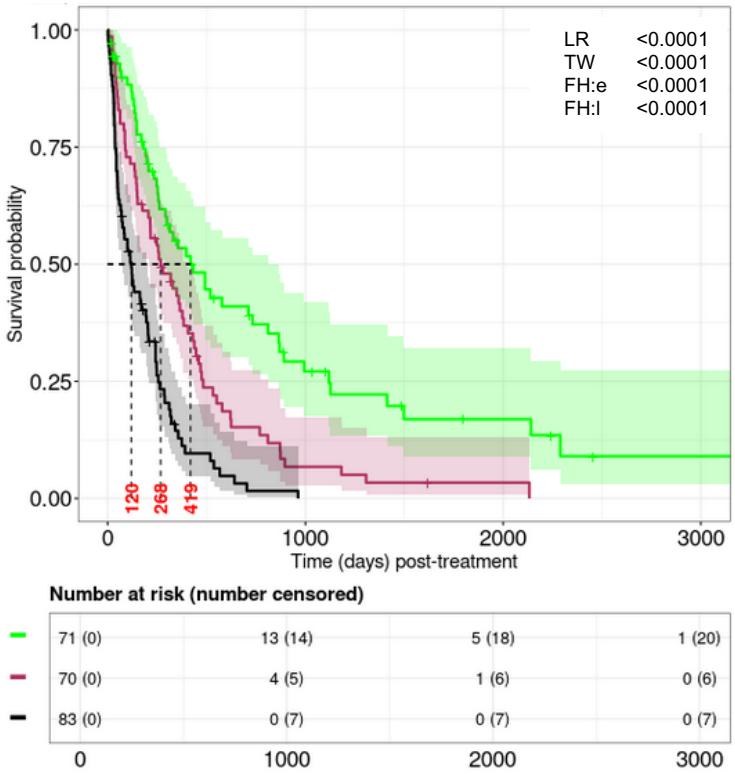
**A. Feature identification strategy and $RM_{G,II}$ features.**

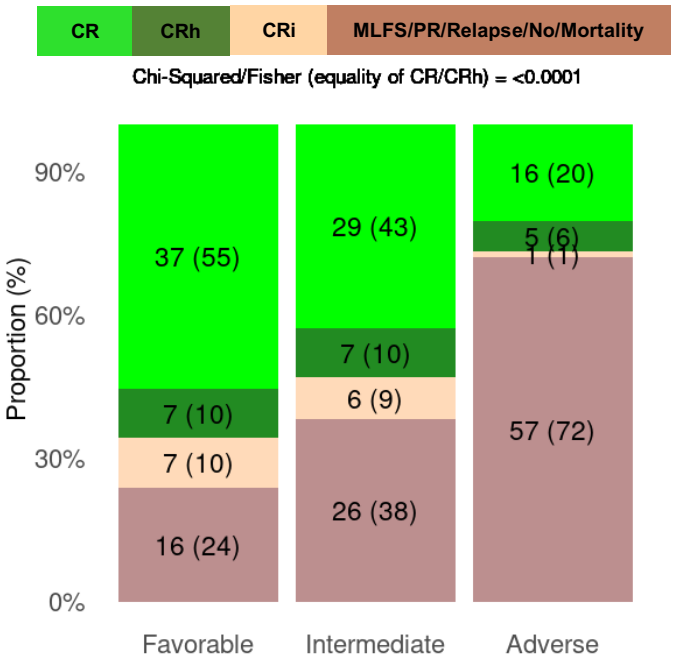
| Category | Genetic features | Frequency† | Frequency¶ |
|---|---|---|---|
| **Favorable** | Any of ***BCOR, DNMT3A, DDX41, SF3B1, IDH1, IDH2, NPM1*** (in the absence of any Adverse features) | **71/224 (32%)** | **101/316 (32%)** |
| **Intermediate** | No Favorable or Adverse features, or ≥1 Favorable 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

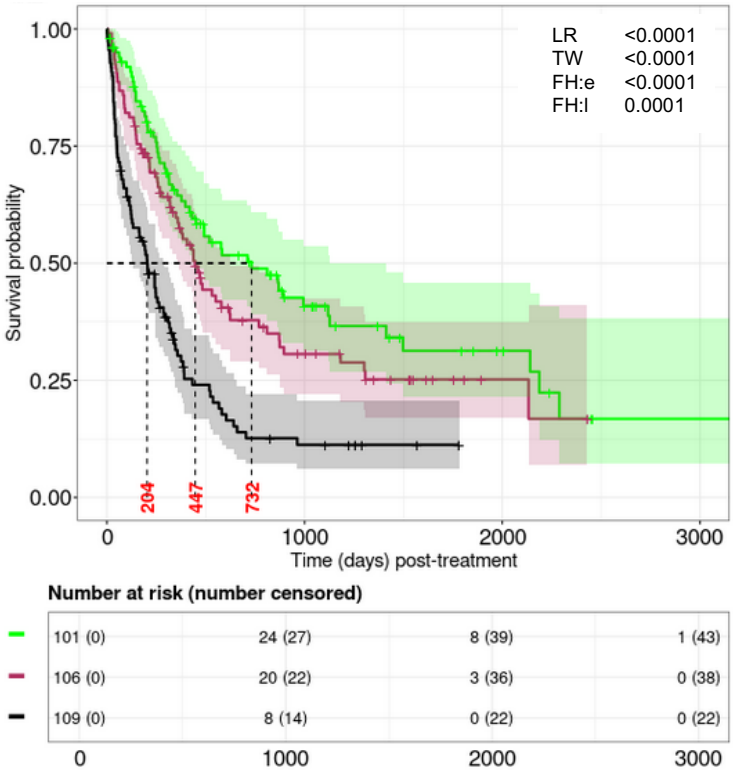
**Figure 4, cont'd**

**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.**



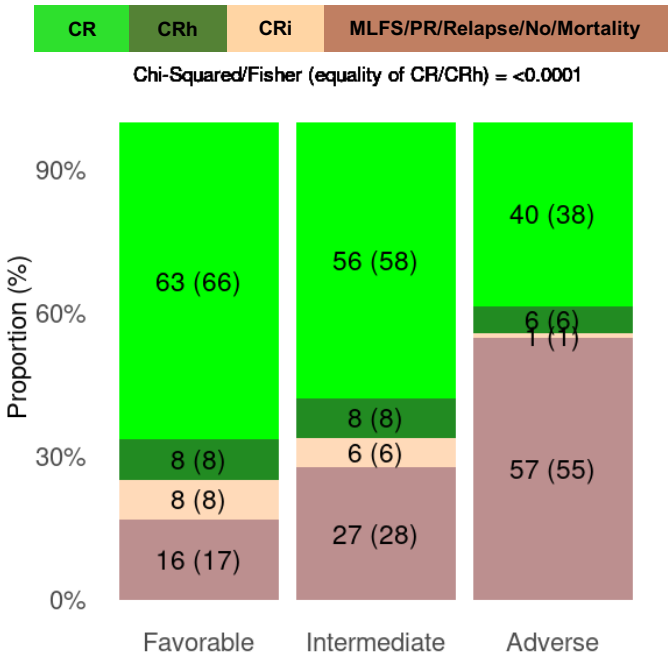



**Figure 4, cont'd**

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

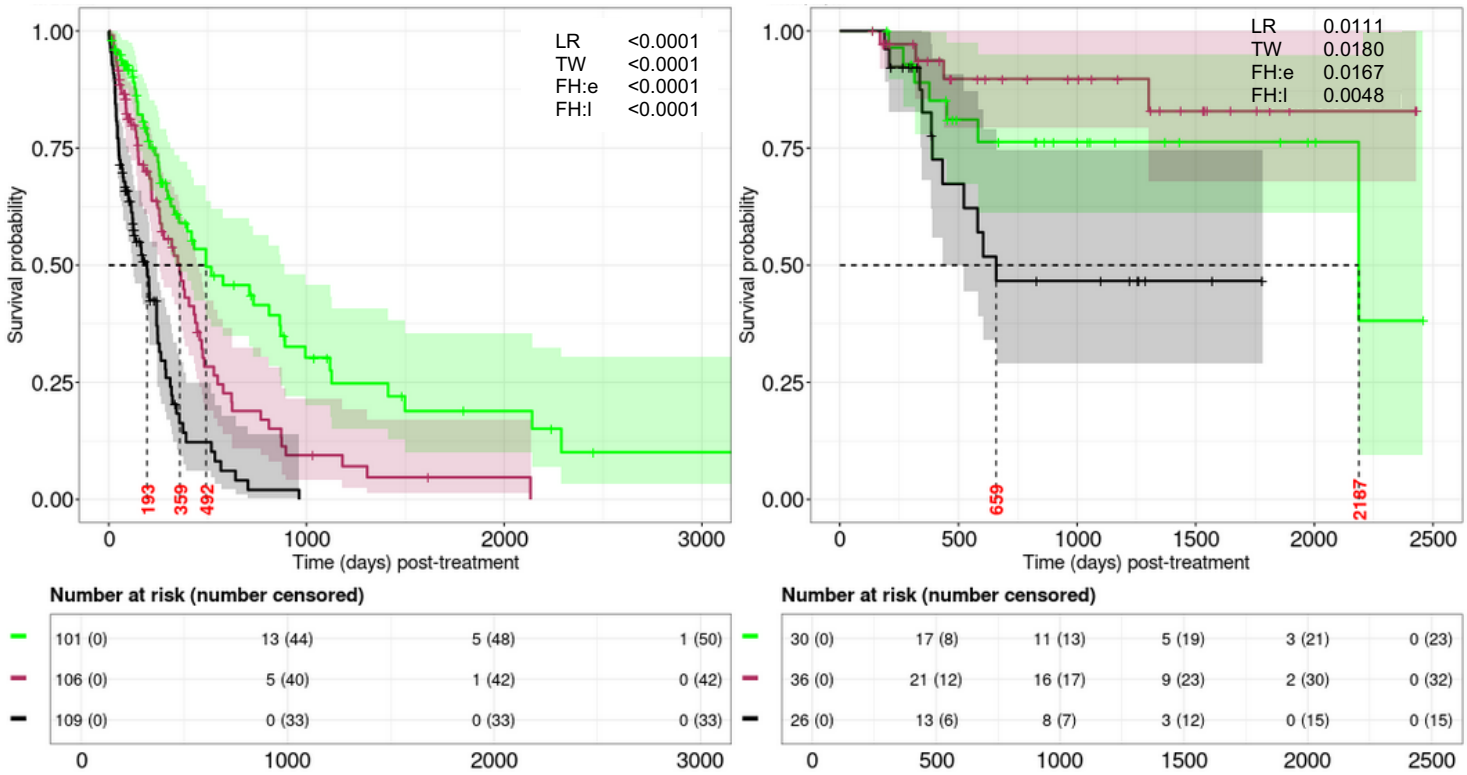


**Figure 5. Event free survival (EFS) based on the genetic features specific risk model using Rule-II ($RM_{G,II\text{-}EFS}$).**

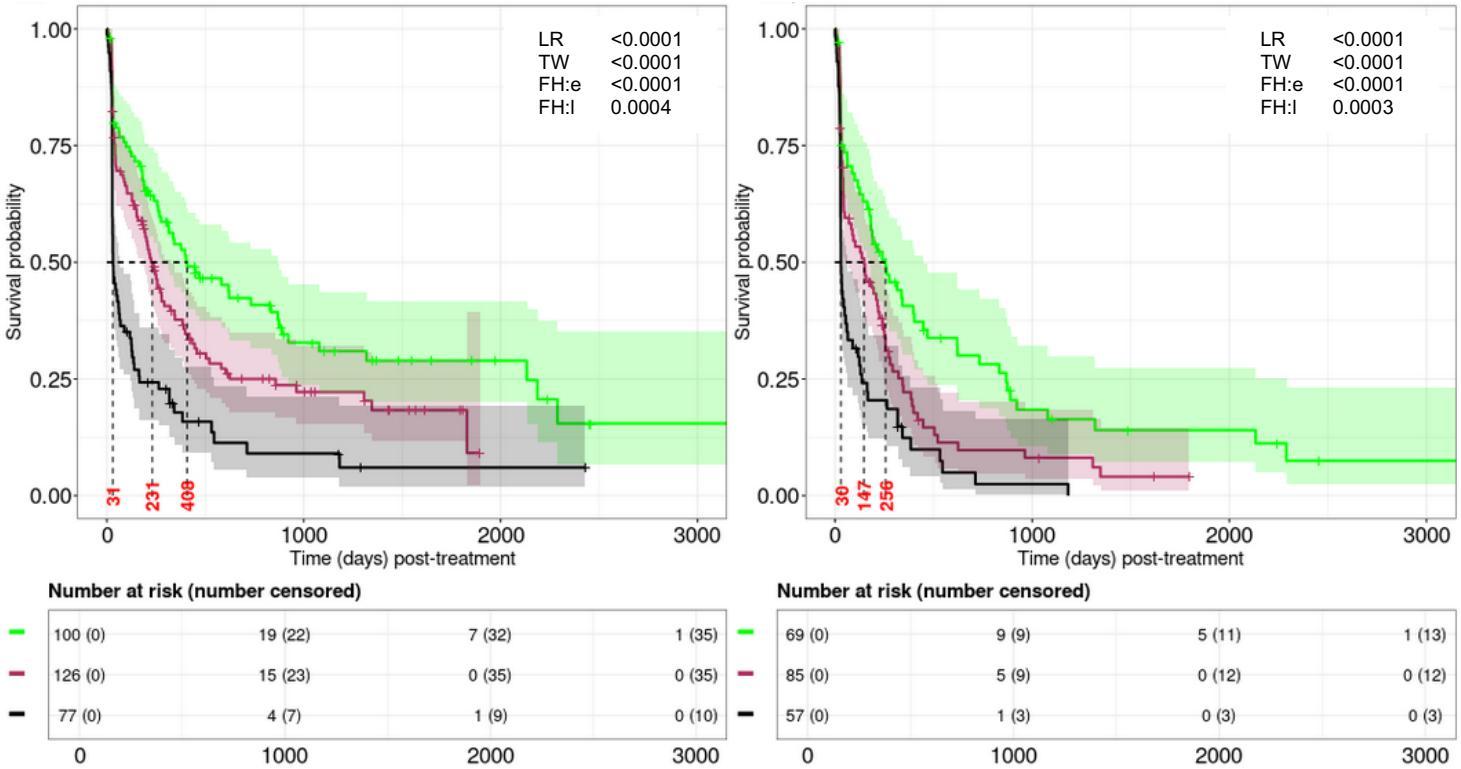
**A. Feature identification strategy and $RM_{G,II\text{-}EFS}$ features.**

| 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 set excluding allo-HCT patients (right).**




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

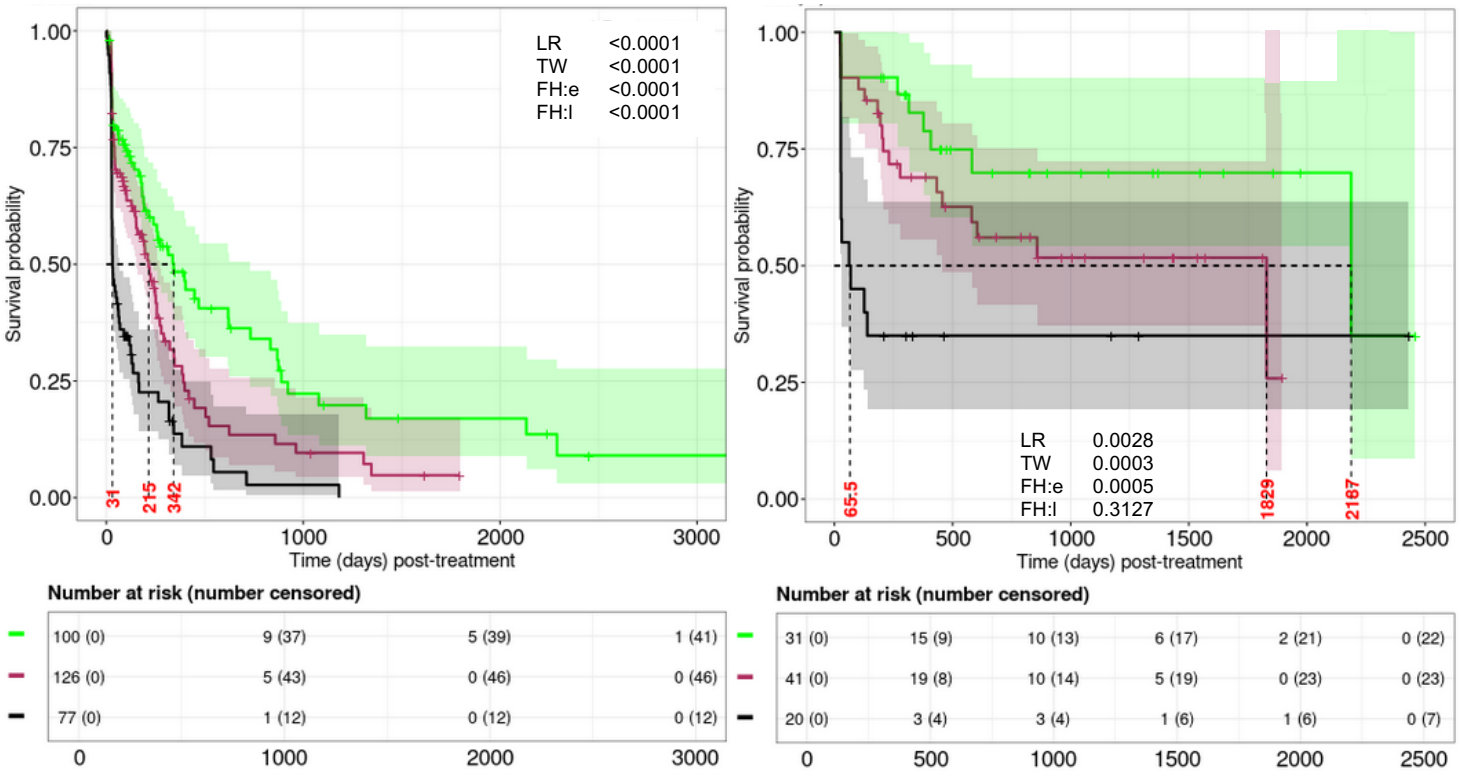


**Figure 6. Adaptation of the ML strategy to an external real-world cohort (RWC) using genetic features only and Rule II (RWC-$RM_{G,II}$).**

**A. Feature set for the RWC-$RM_{G,II}$ for OS.**

| Category | RWC-$RM_{G,II}$ classification | Frequency† | Frequency¶ |
|---|---|---|---|
| **Favorable** | Any of ***DNMT3A, DDX41, IDH1, IDH2, NPM1*** (in the absence of any Adverse features) | **169/911 (18%)** | **180/971 (19%)** |
| **Intermediate** | No Favorable or Adverse features, or ≥1 Favorable features with ≥1 Adverse features | **471/911 (52%)** | **506/971 (52%)** |
| **Adverse** | Any of ***KRAS, JAK2, p53m, Minus7, t(9;11), Minus5, Minus17, Complex cytogenetics, Chr 7 abn:loss*** (in the absence of any Favorable features) | **271/971 (30%)** | **285/971 (29%)** |

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

**B. RWC-$RM_{G,II}$ for FAS (left) and set excluding allo-HCT (right) patients.**

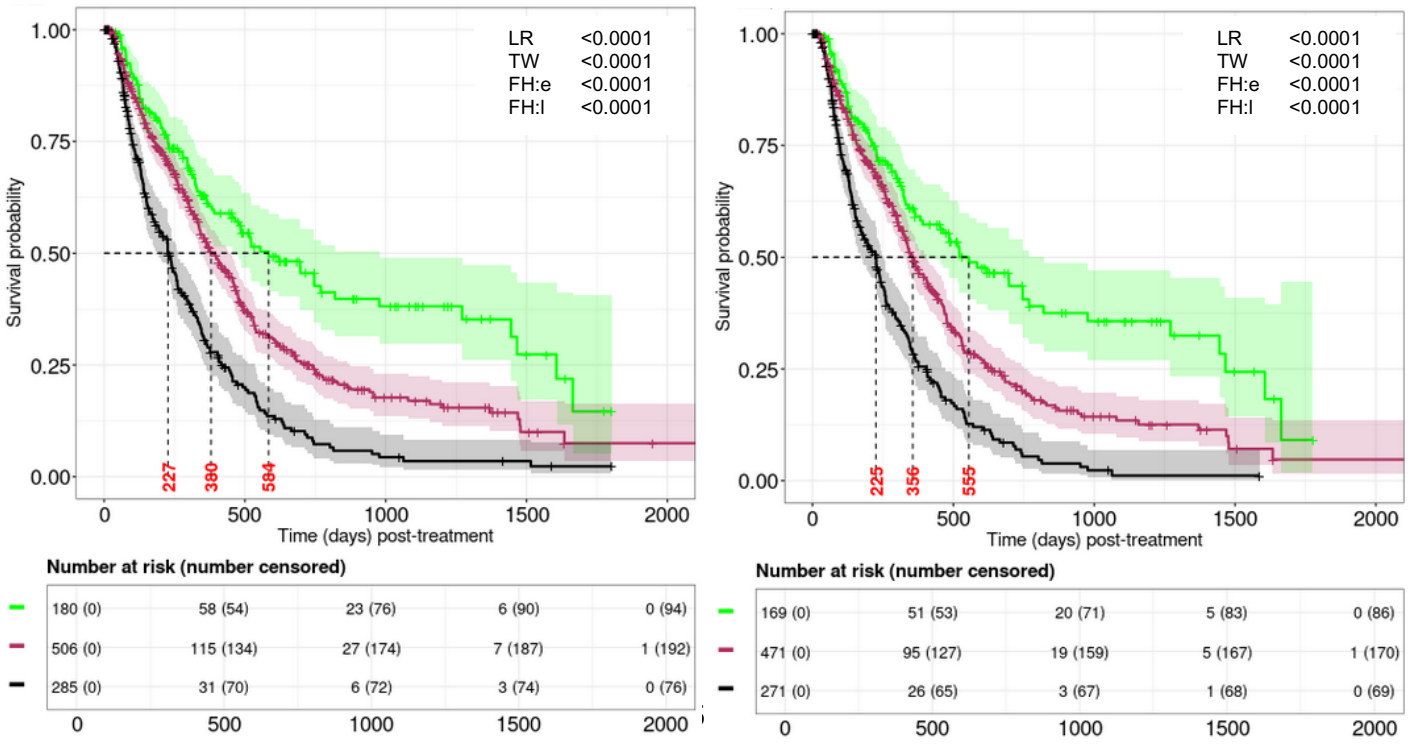


**Figure 6, cont'd**

**C. Pairwise comparison between RWC-RM$_{G,II}$ and ELN22.**

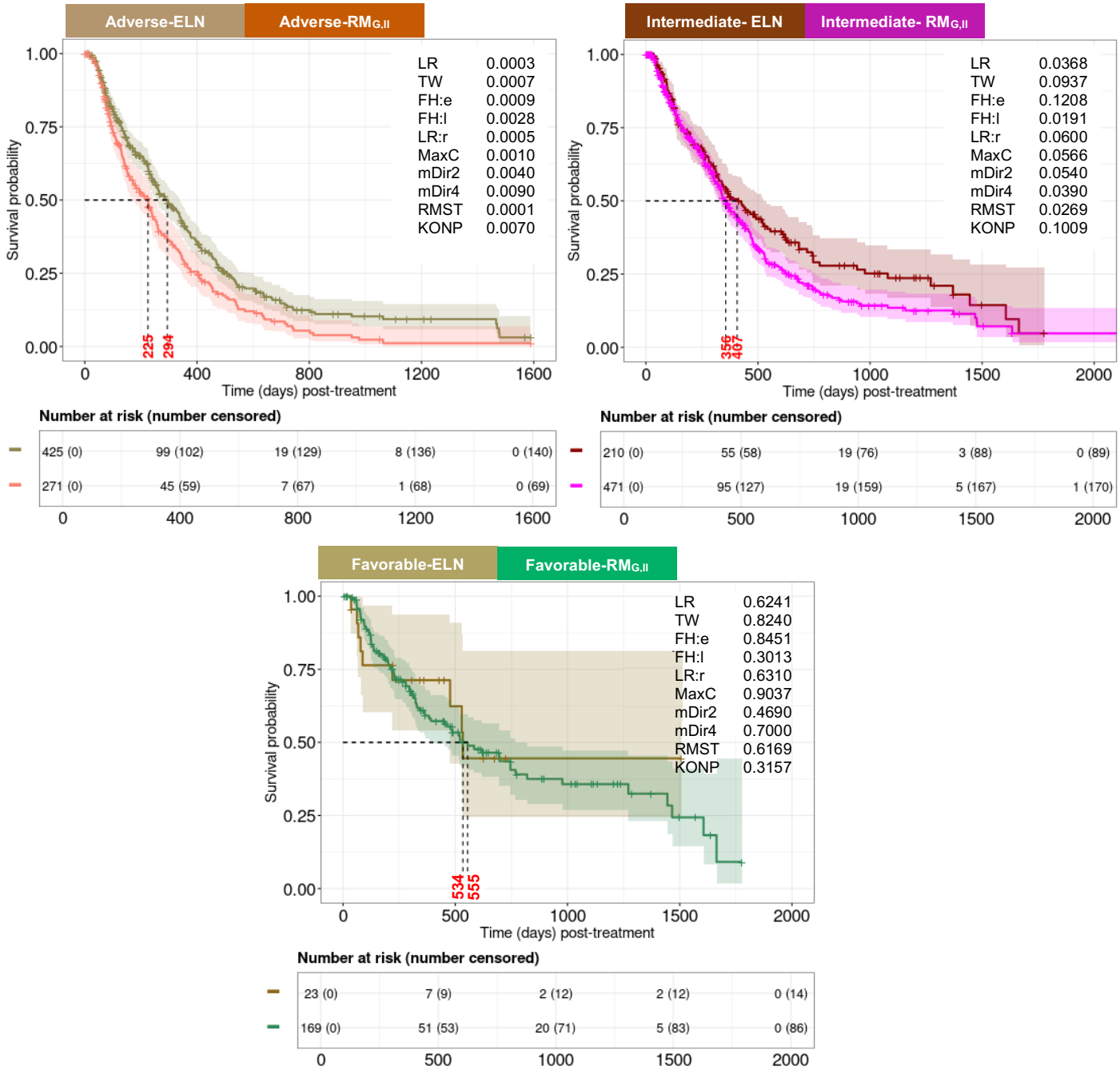
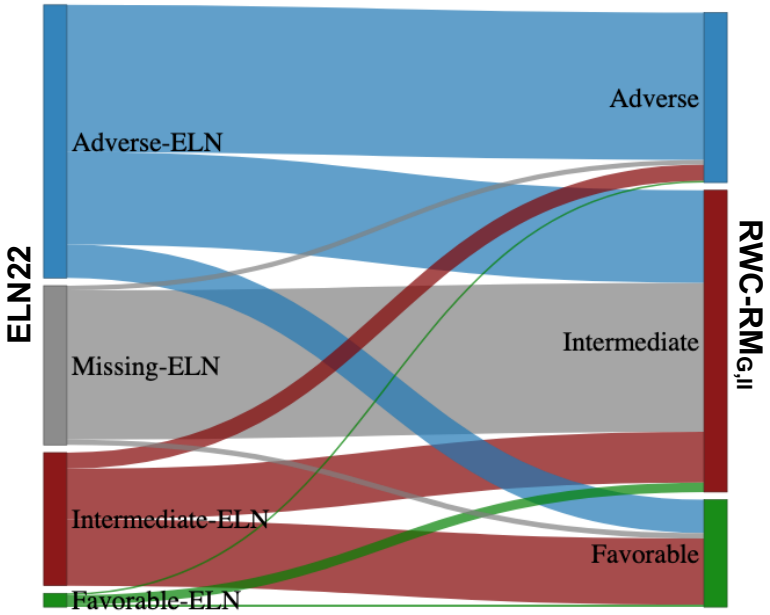


**Figure 6, cont'd**

**D. Agreement between RWC-RM$_{G,II}$ and ELN22.**




| | Favorable | Intermediate | Adverse |
|---|---|---|---|
| [a]Fleiss kappa (*P* values) | -0.11 (0.003) | 0.02 (0.554) | 0.31 (<0.001) |

**Remarks:**

| |
|---|
| The higher positive value (i.e., close to 1) means more agreement |
| The lower negative value (i.e., close to -1) means less agreement |
| Value close to 0 means agreement is no better than obtained by chance |

[a]Analytical set is comprised of the patients for who ELN22 was observed in the CU dataset.

**E. Predictability comparison between the ELN22 and RWC-RM$_{G,II}$ over follow-up time.**

| | Cumulative/dynamic time-dependent AUC | | | Incident/dynamic time-dependent AUC | | |
|---|---|---|---|---|---|---|
| [a]Model | cAUC (2.5$^{th}$,97.5$^{th}$) | (25$^{th}$, 75$^{th}$) | cAUC$_{15}$ | iAUC (2.5$^{th}$,97.5$^{th}$) | (25$^{th}$, 75$^{th}$) | iAUC$_{15}$ |
| ELN22 | 0.52 (0.50, 0.59) | (0.51, 0.53) | 0.52 | 0.51 (0.51, 0.51) | (0.51, 0.51) | 0.51 |
| RWC-RM$_{G,II}$ | 0.62 (0.60, 0.75) | (0.61, 0.66) | 0.64 | 0.59 (0.57, 0.60) | (0.59, 0.60) | 0.59 |

[a]Penalized CoxPH model was trained on the CU dataset adjusting for age, gender, and by either ELN22 or original RM$_{G,II}$. For the test set covariate, RWC-RM$_{G,II}$ was used to determine risk classification in the RWC set; see supplementary section S9 for details.
*One-sided Mann-Whitney U-test based P-values comparing cAUC and iAUC are <0.0001 and <0.0001, respectively between ELN22 and RWC-RM$_{G,II}$.

**Figure 6, cont'd**

**F. Pairwise comparisons of predicted survival probability curves.**

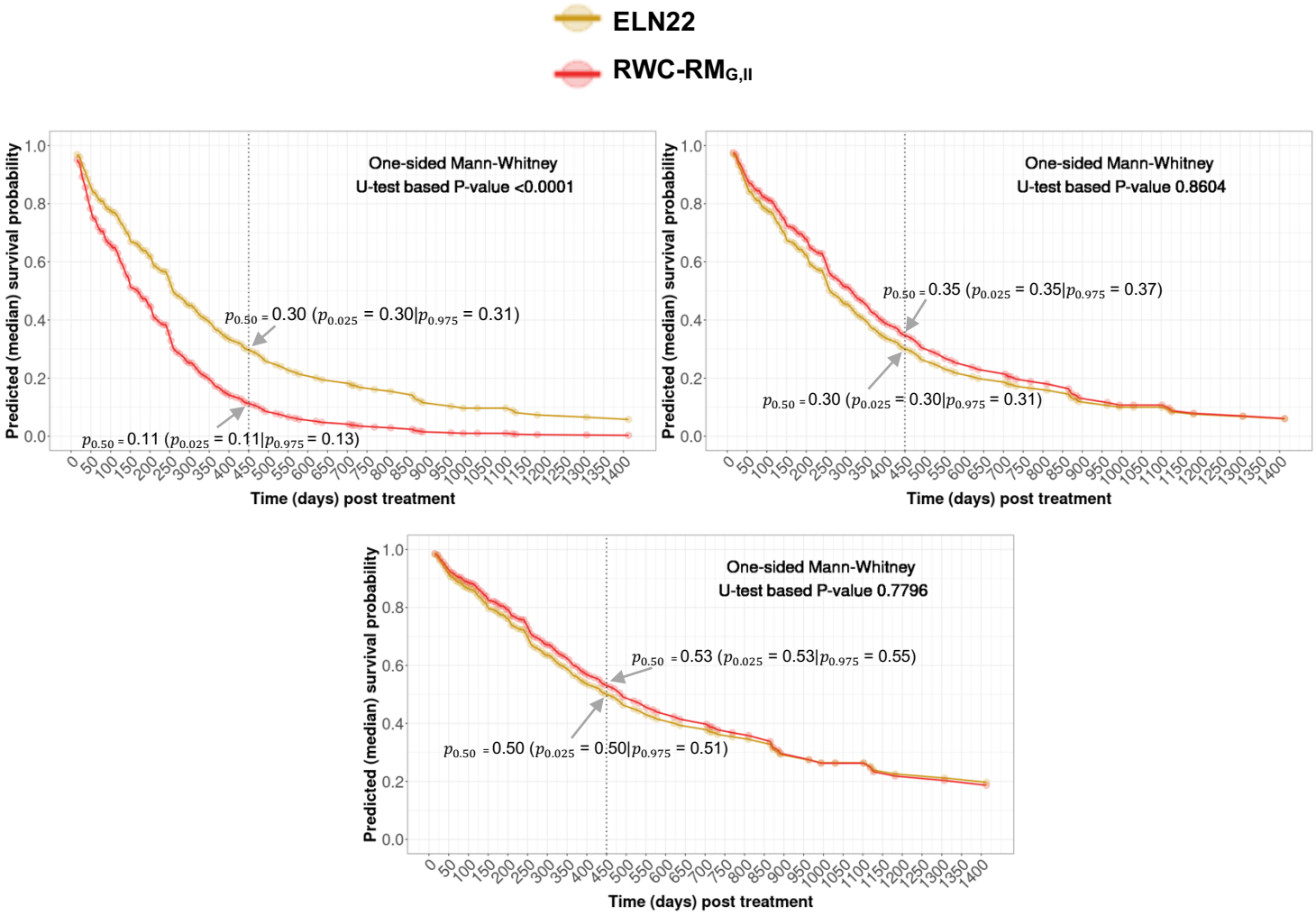
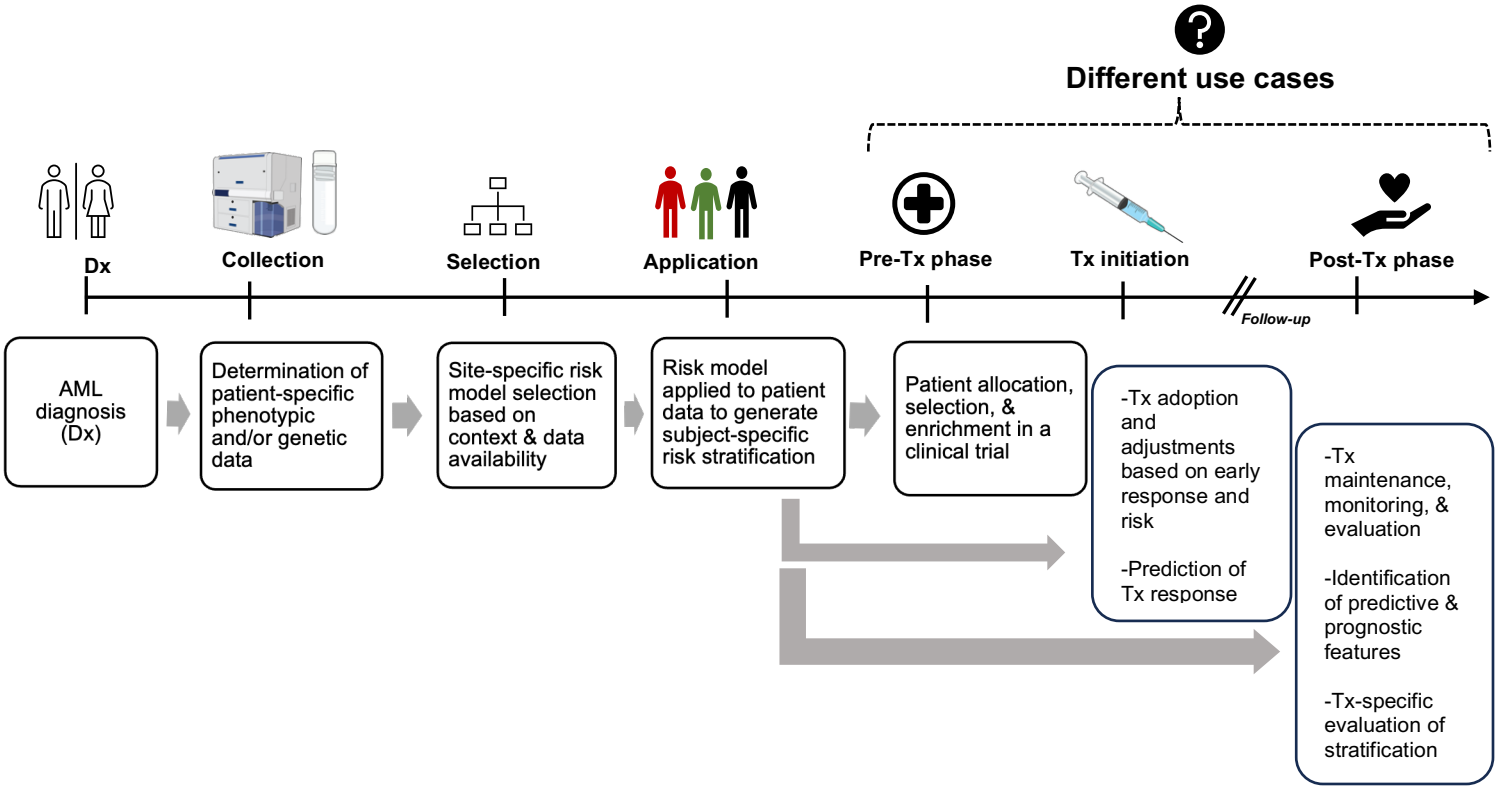


**Figure 6, cont'd**

**G. Summary of RM procedures and potential range of specific use cases**