QRT Data Challenge 2025 : Overall Survival Prediction of Patients with Myeloid Leukemia

1st place out of 634 participants

Arthur DE ROUCK, Ruben BARATA, Rémy SIAHAAN-GENSOLLEN

`https://github.com/arthurdrk/QRT-Challenge-2025`

Table of Contents

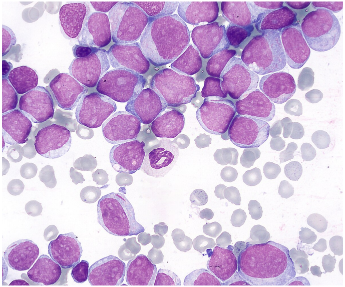
1. Introduction
   - Data overview
   - Target & evaluation metric
   - Data preprocessing
2. Feature Engineering
   - Determinants of AML prognosis
   - Cytogenetic feature extraction
   - Mutation feature engineering
3. Modeling Strategy
   - Model Architecture Overview
   - XGBoost with Accelerated Failure Time (XGB AFT)
   - Neural Multi-Task Logistic Regression (N-MTLR)
4. Results & Ensembling Analysis
   - Ensembling Strategy
   - Results

1. Introduction

This year's data challenge focuses on a subtype of blood cancer called **Acute Myeloid Leukemia (AML)**.

Characteristics of the disease :

- Rapid accumulation of abnormal immature myeloid cells (blasts).
- The bone marrow produces dysfunctional blood cells instead of healthy ones.



Microscopic view showing the accumulation of large immature leukemic blasts, in purple (Credit : Jarun Ontakrai).

Goal of the challenge : Predict the risk of death

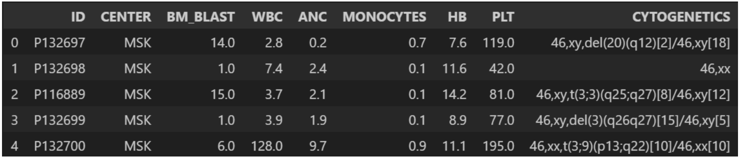
Being able to predict this risk helps doctors adapt treatments and patient follow-up, in order to improve survival.

Introduction

Data overview

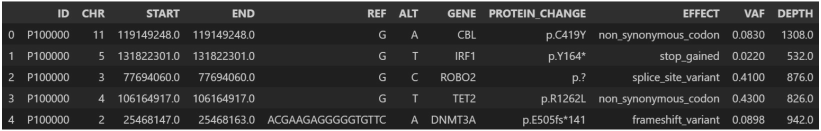
Data is divided in two parts : **Clinical Data** and **Molecular Data**
Very large dataset :

- **3,323** patients in train set
- **1,193** patients in test set



	ID	CENTER	BM_BLAST	WBC	ANC	MONOCYTES	HB	PLT	CYTOGENETICS
0	P132697	MSK	14.0	2.8	0.2	0.7	7.6	119.0	46,xy,del(20)(q12)[2]/46,xy[18]
1	P132698	MSK	1.0	7.4	2.4	0.1	11.6	42.0	46,xx
2	P116889	MSK	15.0	3.7	2.1	0.1	14.2	81.0	46,xy,t(3;3)(q25;q27)[8]/46,xy[12]
3	P132699	MSK	1.0	3.9	1.9	0.1	8.9	77.0	46,xy,del(3)(q26q27)[15]/46,xy[5]
4	P132700	MSK	6.0	128.0	9.7	0.9	11.1	195.0	46,xx,t(3;9)(p13;q22)[10]/46,xx[10]

Figure 1 – Head of clinical train set



	ID	CHR	START	END	REF	ALT	GENE	PROTEIN_CHANGE	EFFECT	VAF	DEPTH
0	P100000	11	119149248.0	119149248.0	G	A	CBL	p.C419Y	non_synonymous_codon	0.0830	1308.0
1	P100000	5	131822301.0	131822301.0	G	T	IRF1	p.Y164*	stop_gained	0.0220	532.0
2	P100000	3	77694060.0	77694060.0	G	C	ROBO2	p.?	splice_site_variant	0.4100	876.0
3	P100000	4	106164917.0	106164917.0	G	T	TET2	p.R1262L	non_synonymous_codon	0.4300	826.0
4	P100000	2	25468147.0	25468163.0	ACGAAGAGGGGGTGTTC	A	DNMT3A	p.E505fs*141	frameshift_variant	0.0898	942.0

Figure 2 – Head of molecular train set

Each patient is associated with a unique identifier and detailed clinical information :

- `ID` : unique identifier per patient
- `CENTER` : clinical center
- `BM_BLAST` : bone marrow blasts in % (blasts are abnormal blood cells)
- `WBC` : white blood cell count in Giga/L
- `ANC` : absolute Neutrophil count in Giga/L
- `MONOCYTES` : monocyte count in Giga/L
- `HB` : hemoglobin in g/dL
- `PLT` : platelet count in Giga/L
- `CYTOGENETICS` : description of the karyotype observed in blood cells, measured by a cytogeneticist

One line per patient per somatic mutation :

- `ID` : Unique identifier per patient
- `CHR_START_END` : Position of the mutation on the human genome
- `REF_ALT` : Reference and alternate (mutant) nucleotide
- `GENE` : Affected gene
- `PROTEIN_CHANGE` : Consequence of the mutation on the protein expressed by the gene
- `EFFECT` : Broad categorization of the mutation consequence on the gene
- `VAF` : Variant Allele Fraction (proportion of cells carrying the deleterious mutation)
- `DEPTH` : Coverage (total number of reads at the locus)

Target : Overall Survival (OS)

- X_i = time to event (`OS_YEARS`)
- Δ_i = event indicator (1 = death, 0 = censored) (`OS_STATUS`)
- Truncation : $\tau = 7$ years

Metric : Concordance Index (C-Index)

- Measures ranking quality (0.5 = random, 1 = perfect)
- Concordant if : earlier death → higher predicted risk

IPCW C-Index (handles censoring)

We weight pairs using the probability of being uncensored :

$$\hat{C}_\tau = \frac{\sum_{i,j} \Delta_i \hat{G}(X_i)^{-2} \mathbf{1}\{X_i < X_j\} \mathbf{1}\{\text{Risk}_i > \text{Risk}_j\}}{\sum_{i,j} \Delta_i \hat{G}(X_i)^{-2} \mathbf{1}\{X_i < X_j\}}$$

where $\hat{G}(t)$ is the Kaplan–Meier estimate of $\mathbb{P}(\{\text{not censored at } t)\}$ (the censoring survival function).

Introduction

Data preprocessing

- Missing clinical values imputed using optimized XGBoost models (trained on the training set and applied to validation data)
- Continuous variables scaled using RobustScaler (median and IQR) to reduce the influence of outliers
- Log(`1+p`) transformation applied to highly skewed variables to stabilize variance and reduce asymmetry

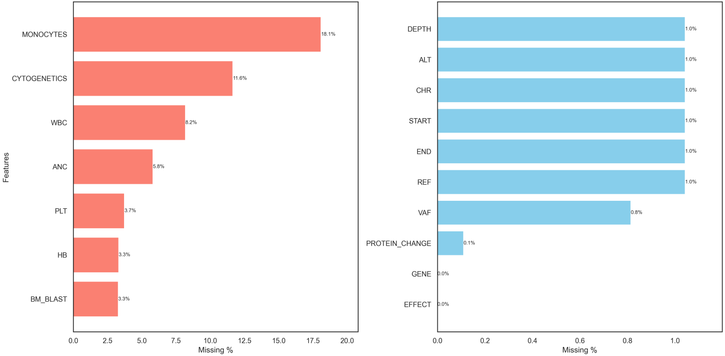


Figure 3 – Missing values in Clinical train and Molecular train

2. Feature Engineering

Feature Engineering

Determinants of AML prognosis

Prognostic Standard (ELN 2017)

Cytogenetic abnormalities and gene mutations are the primary determinants of prognosis in **Acute Myeloid Leukemia** (Döhner et al., 2017).

1. Cytogenetics (ISCN)

- **Normal :** *46,XX* (F) or *46,XY* (M). 23 standard pairs.
- **Abnormal :** Structural or numerical changes.
- *Example :* **-7** (Monosomy 7) indicates a high-risk profile.

2. Gene Mutations

- Comprehensive list of **mutated genes** per patient.
- Detailed descriptions of each mutation variant.
- Integrated with cytogenetics to define final ELN risk groups.

→ **Goal :** Transform these complex raw strings (ISCN/Mutations) into numerical features for downstream tasks.

Feature Engineering

Cytogenetic feature extraction

Converting ISCN descriptions into structured features :

```
46,XX,t(8;21)(q22;q22),del(5q),-7,+8[12]/47,XX,+13,inv(3)(q21q26)[8]
```

1. Abnormality burden

- Total number of events
- Affected chromosomes
- Ploidy status (Hypo/Hyper)

2. Clinical lesions

- Deletions ($-5/7, 5q/7q$)
- Rearrangements (CBF, APL)
- Specific mutations (17p, inv3)

3. Risk summary

- Monosomal/Complex karyotype
- ELN risk class
- Binary score (Adverse / Non-adverse)

4. Clonal structure

- % of abnormal metaphases
- Size of the dominant clone
- Severity of the worst clone

Feature engineering

Effect of cytogenetic features on prediction

Setup

- Models :
  1. Clinical + Molecular only
  2. Clinical + Molecular + Cytogenetics
- Cox elastic-net, fully nested CV (5-fold outer, 3-fold inner)

Performance (outer CV mean ± SD)

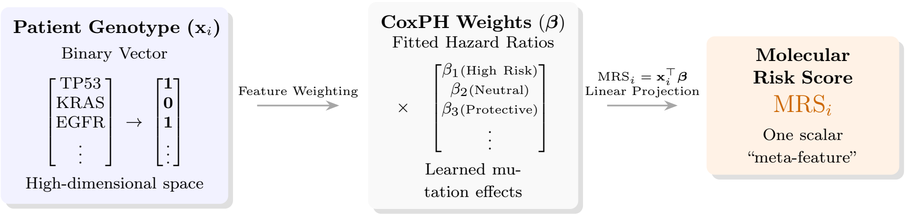
Metric	No cytogenetics	With cytogenetics
C-index	0.741 ± 0.004	0.742 ± 0.004
IBS	0.161 ± 0.009	0.161 ± 0.009
AUC (1 year)	0.795 ± 0.010	0.796 ± 0.010

Interpretation

- Cytogenetics provide a *small but consistent* improvement
- Gains are stable across folds (nested CV → low overfitting risk)
- Directionally aligned across all metrics (C-index, IBS, AUC)

Feature Engineering : Molecular Risk Score (MRS)

Dimensionality reduction of mutation data via CoxPH



Patient Genotype ($\mathbf{x}_i$)

Binary Vector

$$\begin{bmatrix} \text{TP53} \\ \text{KRAS} \\ \text{EGFR} \\ \vdots \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \\ \mathbf{1} \\ \vdots \end{bmatrix}$$

High-dimensional space

Feature Weighting →

CoxPH Weights ($\boldsymbol{\beta}$)

Fitted Hazard Ratios

$$\times \begin{bmatrix} \beta_1 \text{(High Risk)} \\ \beta_2 \text{(Neutral)} \\ \beta_3 \text{(Protective)} \\ \vdots \end{bmatrix}$$

Learned mutation effects

$\text{MRS}_i = \mathbf{x}_i^\top \boldsymbol{\beta}$

Linear Projection →

Molecular Risk Score

MRS_i

One scalar "meta-feature"

1. Input preprocessing

- **Prevalence Filter :** Genes kept if $1\% \leq \text{freq} \leq 99\%$.
- **Rationale :** Eliminates noise from ultra-rare variants and non-informative ubiquitous mutations.

2. Semantic compression

- Transforms p sparse features into a single continuous prognostic index.
- Efficiently handles right-censored survival data.

Feature engineering : Molecular Risk Score (MRS)

Cox Proportional Hazards Model

Cox Proportional Hazards Model

The instantaneous risk of death (hazard) at time t for patient i is defined as :

$$h(t \mid \mathbf{x}_i) = h_0(t) \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$$

where $h_0(t)$ is the baseline hazard and $\mathbf{x}_i^\top \boldsymbol{\beta}$ is the **Molecular Risk Score (MRS)**.

Model Assumptions

- **Proportionality :** The ratio of hazards between two patients is constant over time : $\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_j)} = \exp(\boldsymbol{\beta}(\mathbf{x}_i - \mathbf{x}_j))$.
- **Baseline Agnosticism :** $h_0(t)$ remains unspecified, focusing on the *relative risk* of genomic features.

Coefficient interpretation

- $\beta_j > 0$: Mutation j increases hazard (**pro-tumoral**).
- $\beta_j < 0$: Mutation j decreases hazard (**protective**).
- $\beta_j \approx 0$: No significant impact on survival.

Feature Engineering : The Molecular Risk Score (MRS)

Prognostic performance of the Molecular Risk Score

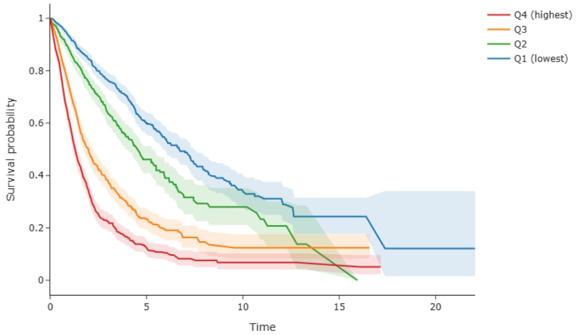


Figure 4 – Kaplan–Meier Survival Curves stratified by MRS quartiles.

Predictive accuracy

- **Harrell's C-index :** 0.70 (5-fold CV)
- **Bootstrap (n=1000) :**
  - Mean : 0.700
  - 95% CI : [0.686; 0.713]

Clinical stratification

- **Monotone Trend :** Clear survival separation between all quartiles ($p < 0.001$).
- **Interpretation :** Higher MRS directly correlates with increased mortality risk.

Feature Engineering : The Molecular Risk Score (MRS)

Biological interpretation

CoxPH coefficients highlight genomic lesions that most strongly shape survival risk, separating **pro-tumoral** from **protective** events.

Gene	**Effect**	**HR**	β	**Z-score**	**Significance**
TP53	Pro-tumoral	1.29	0.25	10.84	$p < 10^{-25}$
RUNX1	Pro-tumoral	1.14	0.13	5.39	$p < 10^{-6}$
ASXL1	Pro-tumoral	1.10	0.09	3.59	$p < 10^{-3}$
NRAS	Pro-tumoral	1.06	0.06	2.40	$p = 0.017$
STAG2	Pro-tumoral	1.05	0.05	2.12	$p = 0.034$
SF3B1	Protective	0.96	-0.04	-1.69	$p = 0.09$

Key points

- **TP53** = strongest adverse driver.
- Chromatin/splicing genes (*ASXL1*, *STAG2*) ↑ risk.
- RAS-pathway activation contributes (*NRAS*).

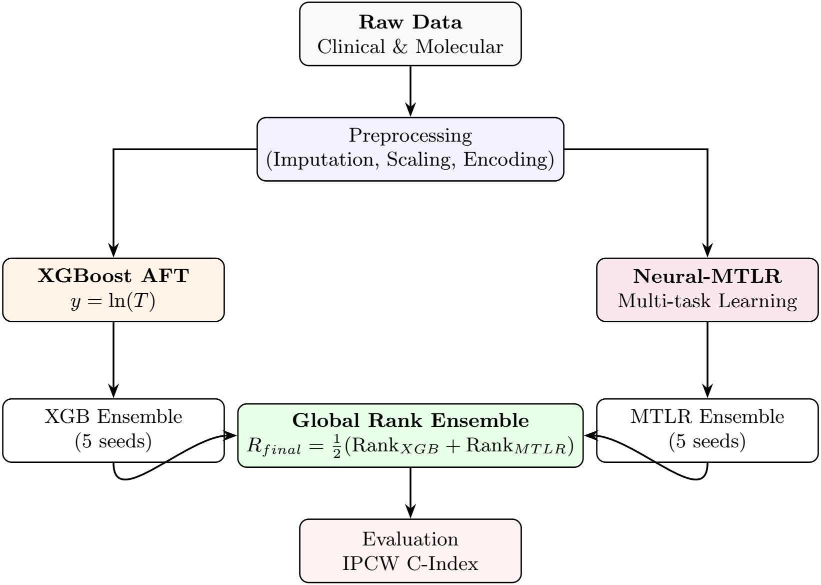
Biological insight

- MRS reflects **genomic aggressiveness**.
- **SF3B1** suggests subtype-specific biology.
- Adds prognostic value beyond clinical data.

3. Modeling Strategy

Modeling Strategy

Model architecture overview



Raw Data
Clinical & Molecular

Preprocessing
(Imputation, Scaling, Encoding)

XGBoost AFT
$y = \ln(T)$

Neural-MTLR
Multi-task Learning

XGB Ensemble
(5 seeds)

Global Rank Ensemble
$R_{final} = \frac{1}{2}(\text{Rank}_{XGB} + \text{Rank}_{MTLR})$

MTLR Ensemble
(5 seeds)

Evaluation
IPCW C-Index

Modeling Strategy

XGBoost AFT

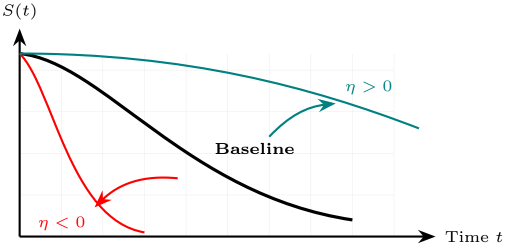
XGBoost AFT model

The **Accelerated Failure Time (AFT)** model assumes

$$\ln T = \eta(\mathbf{x}) + \sigma\varepsilon,$$

where ε follows a chosen error distribution. XGBoost learns the non-linear risk score $\eta(\mathbf{x})$ by maximizing the likelihood of the observed (possibly censored) times.

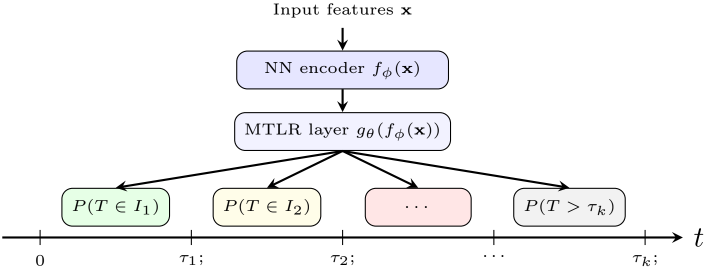
- **High Risk ($\eta < 0$) :** Survival time is **compressed**. The event occurs *earlier.*
- **Low Risk ($\eta > 0$) :** Survival time is **stretched**. The event is *delayed.*



$$\boxed{S(t|\mathbf{x}) = S_0\big(t\, e^{-\eta(\mathbf{x})}\big)}$$

Modeling Strategy

Neural Multi-Task Logistic Regression (N-MTLR)


Input features $\mathbf{x}$
NN encoder $f_\phi(\mathbf{x})$
MTLR layer $g_\theta(f_\phi(\mathbf{x}))$
$P(T \in I_1)$
$P(T \in I_2)$
$\cdots$
$P(T > \tau_k)$
t
0
$\tau_1;$
$\tau_2;$
$\cdots$
$\tau_k;$

Multi-Task Logic

Each interval j defines a classification "task" :

- $P(y_j = 1|\mathbf{x})$ is the probability of survival beyond τ_j.
- The **Softmax** on the mass function ensures that $\sum P(\text{death in } I_j) = 1$.

✓ **Non-Proportional** : The hazard can vary freely across intervals.

✓ **Regularization** : Smoothing penalty $\gamma \sum \|\theta_{j+1} - \theta_j\|^2$ for a "smooth" curve.

4. Results & Ensembling Analysis

Results & Ensembling Analysis

Rank-based ensembling

XGBoost AFT

Parametric Model

- **Modeling bias :** Log-normal survival times
- **Key strength :** Robust to noisy covariates
- **Output scale :** Log survival time ($\ln T$)

Neural-MTLR

Non-Parametric Model

- **Modeling bias :** Piecewise hazard representation
- **Key strength :** Captures complex risk dynamics
- **Output scale :** Survival probability ($S(t)$)

Ensembling strategy

Since the two models output values on different scales (time vs probability), their predictions are mapped into a common **rank space**. This keeps only the relative ordering, which is what drives the C-index.

$$R_{\text{final}} = \frac{1}{2}\Big(\text{Rank}(\hat{y}_{\text{XGB}}) + \text{Rank}(-\hat{S}_{\text{MTLR}})\Big)$$

Effect of ensembling : XGB AFT

Model	**Mean ± SD**	**95% CI**
Single XGB AFT	0.7223 ± 0.0139	$[0.7050\,;\,0.7395]$
Ensemble (5 seeds, rank averaging)	0.7243 ± 0.0131	$[0.7080\,;\,0.7406]$

Paired fold-wise comparison :

$$\Delta C = +0.0021, \quad t = 2.70, \quad p = 0.054$$

Interpretation

- Small but consistent performance gain.
- Borderline statistical significance at the 5% level.
- The ensemble slightly reduces variance and stabilises rankings.

Results & Ensembling Analysis

Effect of ensembling : Neural-MTLR

Model	**Mean ± SD**	**95% CI**
Single Neural-MTLR	0.6997 ± 0.0163	$[0.6795\,;\,0.7199]$
Ensemble (5 seeds, rank averaging)	0.7102 ± 0.0151	$[0.6914\,;\,0.7290]$

Paired fold-wise comparison :

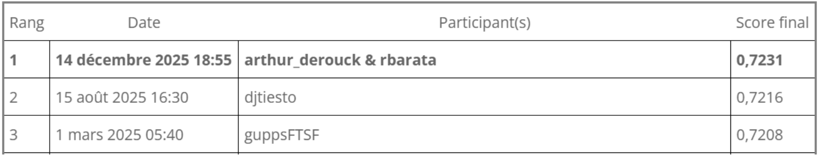
$$\Delta C = +0.0105, \quad t = 6.56, \quad p = 0.0028$$

Interpretation

- Clear and statistically significant improvement.
- Rank-based ensembling is particularly effective for this model.
- Results indicate complementary inductive bias across random seeds.

1$^{\text{st}}$ place out of 634 participants

C-Index on private leaderboard : **0.7231**



Rang	Date	Participant(s)	Score final
1	**14 décembre 2025 18:55**	**arthur_derouck & rbarata**	**0,7231**
2	15 août 2025 16:30	djtiesto	0,7216
3	1 mars 2025 05:40	guppsFTSF	0,7208

Thank you :)