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

Winning solution

Arthur De Rouck & Ruben Barata

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

Context

This year's challenge was about a subtype of blood cancer called myeloid leukemia.

- Accumulation of abnormal immature myeloid cells
- The bone marrow produces dysfunctional blood cells

Goal of the Challenge : Predict risk disease

The risk is measured through the overall survival of patients, i.e., the duration of survival from the diagnosis of the blood cancer to the time of death or last follow-up.

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

Clinical Data

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

Molecular Data

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 & Metric

The goal of the challenge was to predict Overall Survival (OS).
Two outcomes : `OS_YEARS` (time) and `OS_STATUS` (event).

Metric : IPCW-Concordance Index

To take censoring into account, we use an *Inverse Probability of Censoring Weighted* (IPCW) version of the C-index, truncated at $\tau = 7$ years :

$$\hat{C}_\tau = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i \, \hat{G}(X_i)^{-2} \, \mathbf{1}\{X_i < X_j,\ X_i < \tau\} \, \mathbf{1}\{\hat{\beta}' Z_i > \hat{\beta}' Z_j\}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i \, \hat{G}(X_i)^{-2} \, \mathbf{1}\{X_i < X_j,\ X_i < \tau\}}$$

- X_i : follow-up time (`OS_YEARS`)
- Δ_i : event indicator (`OS_STATUS`)
- $\hat{G}(X_i)$: survival function of the censoring distribution
- $\hat{\beta}' Z_i$: predicted risk score
- $\tau = 7$ years : the loss is truncated at 7 years

Data vizualisation

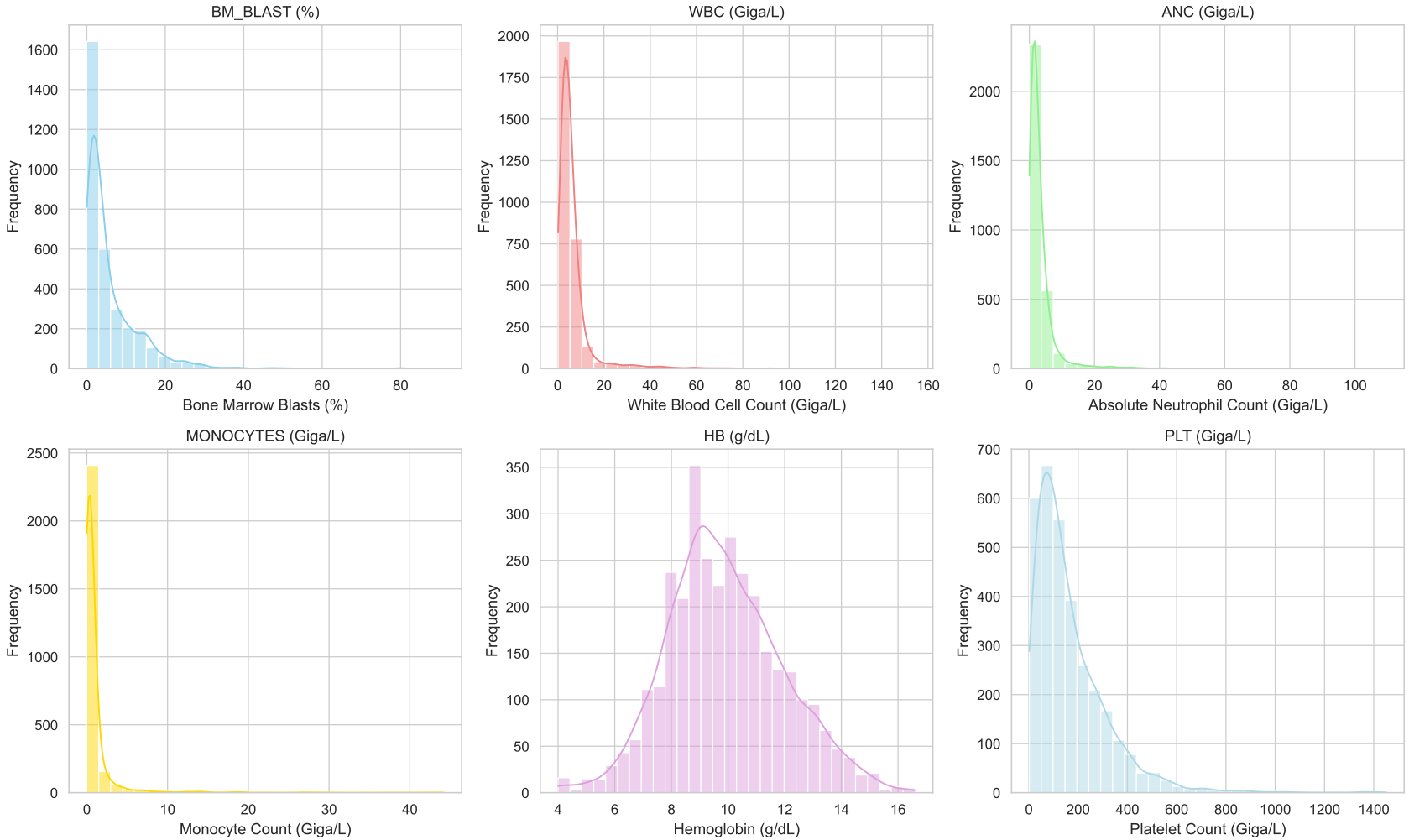


Figure 3 – Clinical variables distributions

Data vizualisation

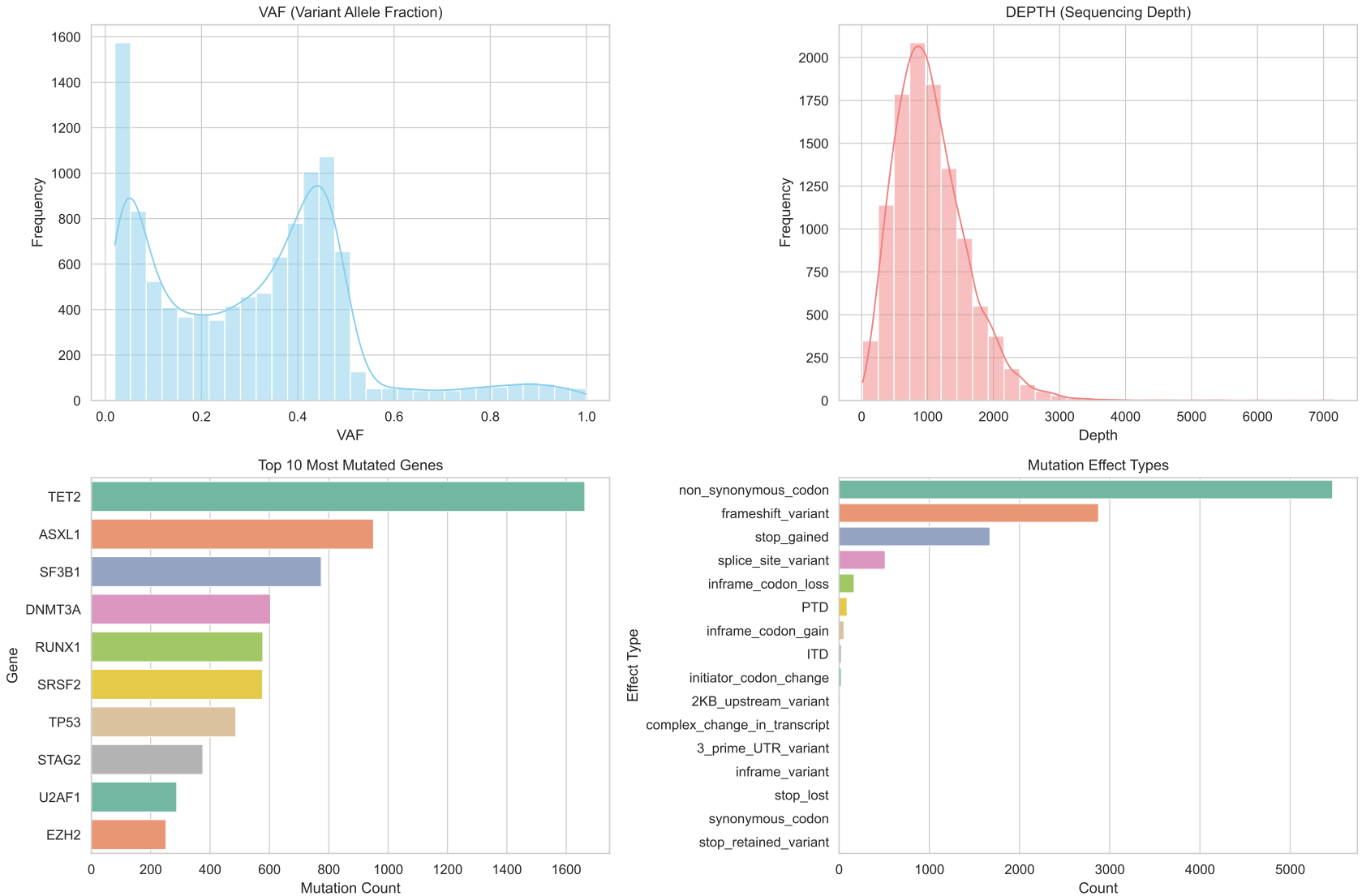


Figure 4 – Molecular variables distributions

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

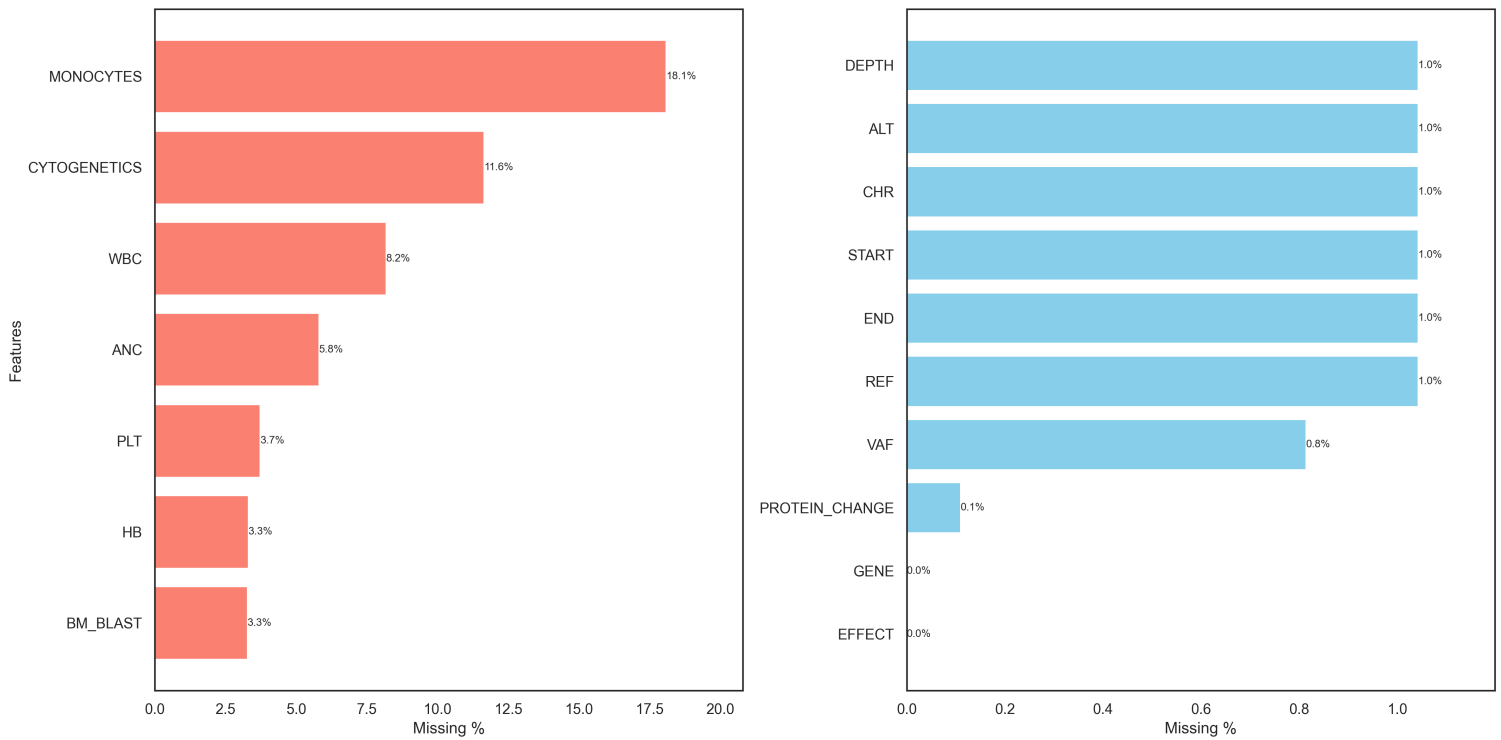


Figure 5 – Missing values in Clinical train and Molecular train

Feature engineering

Overview

- **2017 European LeukemiaNet (ELN) :** cytogenetic abnormalities and gene mutations are the most important determinants of prognosis in acute myeloid leukemia (Döhner et al., Blood, 2017).

- **Cytogenetics :** description of the karyotype observed in the patients' leukemic blood cells, reported using the ISCN (International System for Human Cytogenomic Nomenclature).
  A karyotype can be :
  - **normal**, e.g. *46,XX* (female) or *46,XY* (male), meaning 23 normal pairs of chromosomes,
  - **abnormal**, e.g. loss of chromosome 7 (*-7*, monosomy 7), which is typically associated with higher-risk disease.

  Example : *46,XY,-7* denotes a male karyotype with monosomy 7.

- **Gene mutations :** for each patient we also have the list of mutated genes together with a description of each mutation.

Feature Engineering

ISCN mapping

1. Global Abnormality Load

- Any abnormal clone
- Total number of events
- Number of chromosomes affected
- Hypodiploidy / hyperdiploidy
- Baseline chromosome count

2. Clinically-Relevant Lesions

- -5, del(5q), -7, del(7q)
- $+8$
- CBF / APL rearrangements
- inv(3), t(6 ;9), t(9 ;22), 17p abnormalities

3. Risk Synthesis (ELN-like)

- Monosomal karyotype
- Complex karyotype
- ELN-like cytogenetic risk class
- Binary adverse / non-adverse flag

4. Clonality Structure

- Total metaphases analysed
- Largest clone size
- Proportion abnormal metaphases
- Severity of the worst clone

ISCN cytogenetic descriptions are mapped into a structured set of prognostic features for survival modelling.

Feature engineering

Added prognostic value of cytogenetics

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

Gene Survival Analysis – Summary

Global statistics

- 124 genes analyzed
- 27 significant genes (FDR < 0.05)
- 21 significant genes (Bonferroni < 0.05)
- 26 higher–risk genes (HR > 1)
- 1 protective gene (HR < 1)

Gene	HR [95% CI]	FDR	Effect
TP53	2.74 [2.41–3.13]	4.91×10^{-49}	Higher risk
RUNX1	2.14 [1.89–2.43]	5.19×10^{-31}	Higher risk
ASXL1	1.67 [1.50–1.85]	2.18×10^{-20}	Higher risk
STAG2	1.96 [1.69–2.28]	2.33×10^{-17}	Higher risk
SF3B1	0.64 [0.56–0.72]	1.12×10^{-11}	Protective

Methods

- Log-Rank test (comparison of survival curves)
- Univariate Cox model (Hazard Ratios)
- Benjamini-Hochberg FDR correction
- Bootstrap (200 iterations) for stability assessment

Feature Engineering & Model Strategy

Extracting Signal from High-Dimensional Genomic Data

1. High-Dimensional Input & Filtering

Representing patient i as a binary sparse vector : $\mathbf{x}_i \in \{0,1\}^p$

- **Noise Reduction :** Retain genes with prevalence $\in [1\%, 99\%]$.
- **Rationale :** Eliminates "uninformative constants" to maximize the **signal-to-noise ratio**.

2. Survival Modeling : The Cox Framework

Modeling the **Hazard Rate** to quantify death risk over time.

- **Flexibility :** Semi-parametric (no baseline hazard assumption).
- **Censoring :** Robust handling of non-uniform follow-up periods.

Objective : Molecular Risk Score

Consolidate the sparse mutational landscape into a single, continuous metric to forecast survival and stratify risk.

The Predictive Engine : Cox Framework

Quantifying Mutational Impact on Survival

Model Specification

For patient i, the instantaneous risk of death (Hazard) is modeled as :

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

Core Assumption

- **Proportional Hazards :** The Risk Ratio between patients is **time-invariant**.
- **Flexibility :** $h_0(t)$ is unspecified, allowing focus on the **relative alpha** of mutations.

Interpretation

- $\beta > 0$: High-risk mutation.
- $\beta < 0$: Protective effect.
- **Censoring :** Naturally handles "survivors" or lost-to-follow-up data via partial likelihood.

Robust Estimation & Risk Scoring

Elastic Net Cox Model with Nested CV

Preprocessing

- Retain genes with prevalence $0.5\% \leq p_j \leq 99\%$
- Standardization within each training fold

Elastic Net-Penalized Cox Model

Estimate $\hat{\beta}$ by maximizing the penalized partial log-likelihood :

$$\ell_{\text{partial}}(\beta) - \lambda\big((1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\big)$$

- `lifelines.CoxPHFitter` (`penalizer` $= \lambda$, `l1_ratio` $= \alpha$)
- **Nested CV** : 5 outer folds, 3 inner folds, 30 Optuna trials
- Hyperparameters selected by maximizing Harrell C-index
- Final (λ, α) = median across outer folds, then refit on full data

Molecular Risk Score (MRS)

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

Higher MRS $\Rightarrow$ higher mortality risk (used for patient stratification).

Feature engineering

Evaluation of the Composite Risk Score

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

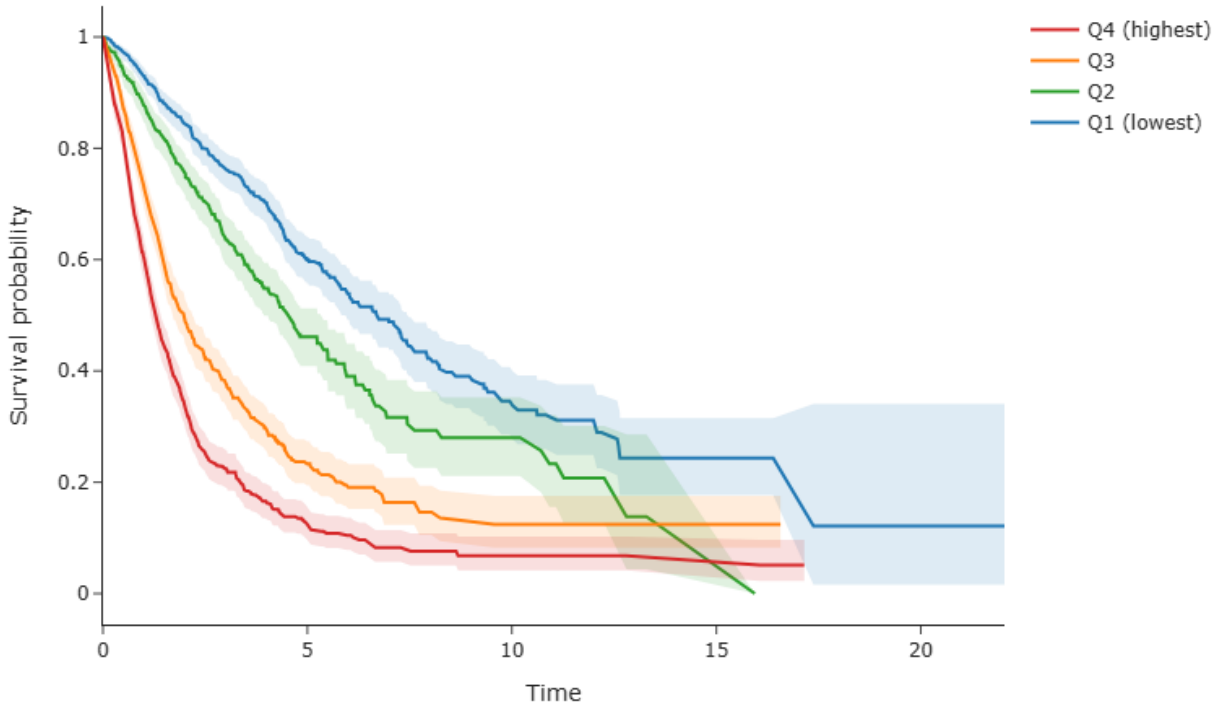
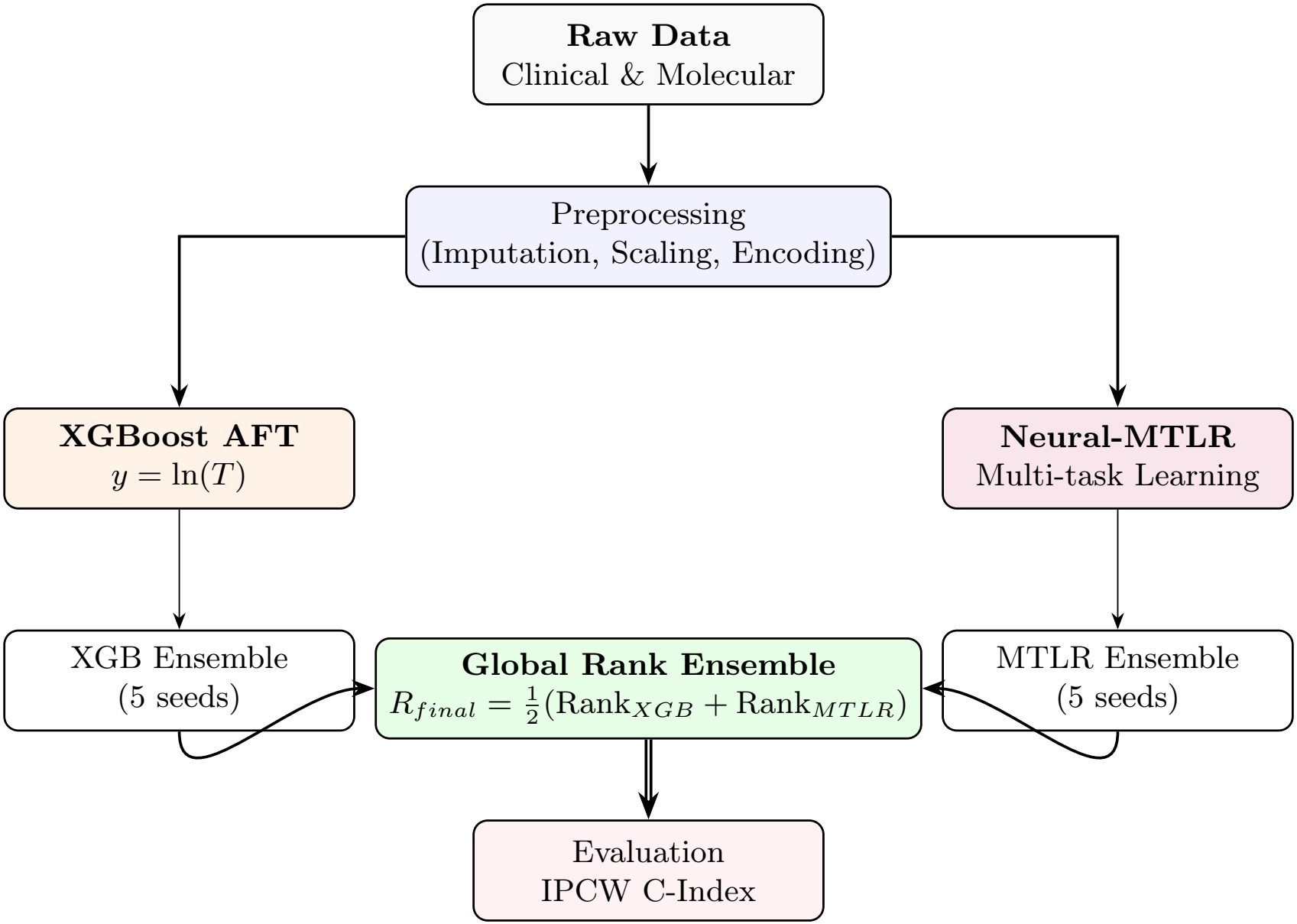


Figure 6 – **Kaplan–Meier** survival curves by molecular risk quartile

- Clear monotone pattern : higher risk quartiles are associated with poorer survival.

Model Architecture Overview



XGBoost AFT

Log-Linearity and Time Acceleration

The **Accelerated Failure Time (AFT)** model is a direct alternative to the Cox proportional hazards model.

Acceleration Assumption

Covariates x act as a multiplicative factor $\exp(\eta(x))$ on survival time :

$$T = e^{\eta(x)} \cdot T_0 \quad \Longrightarrow \quad \ln(T) = \eta(x) + \ln(T_0)$$

- If $\eta(x) > 0$: the event occurs earlier (time is accelerated).
- If $\eta(x) < 0$: the event is delayed (longer survival).

Role of XGBoost : Unlike a classical linear AFT model, XGBoost learns a highly flexible function $\eta(x)$ through tree boosting, capturing complex non-linear effects and interactions among genomic and clinical variables.

XGBoost AFT

Gradient Boosting and Likelihood Maximization

Model training is performed by minimizing the **negative log-likelihood**.
Right-Censoring Case : If a patient is censored at time t_i, we only know that $T_i > t_i$.

$$\mathcal{L}_i = \delta_i \underbrace{\ln f(z_i)}_{\text{Observed event}} + (1-\delta_i) \underbrace{\ln S(z_i)}_{\text{Censoring}} \quad \text{with} \quad z_i = \frac{\ln t_i - \eta_i}{\sigma}$$

- **Scale parameter σ :** controls the dispersion of log-survival times. A smaller σ corresponds to greater model confidence.
- **Numerical stability :** XGBoost uses a second-order Taylor expansion ; gradients and Hessians depend on the assumed distribution family (e.g., Weibull, Log-Normal).

Neural-MTLR

Time Discretization and Multi-Task Learning

Multi-Task Logistic Regression (MTLR) reformulates survival prediction as a sequence of dependent binary classification tasks.

Time Grid

Define k time points $\tau_1, \ldots, \tau_k$ (e.g., event-time quantiles), partitioning the time axis into $k+1$ intervals.

- **Multi-task view :** For each interval j, the model predicts the probability of surviving past τ_j.
- **Target encoding :** For an event at time t, the target vector consists of 1's (survival) followed by 0's (post-event).
- **Key advantage :** No proportional-hazards assumption — the effect of a covariate may vary over time.

Neural-MTLR

Architecture and Monotonicity

A neural network implemented via `nn.Sequential` extracts non-linear features prior to the MTLR layer.

From Features to Survival Curve

1. **Feature encoding :** Dense layers map clinical data into a latent representation.
2. **Piecewise linear scoring :** The MTLR layer outputs k scores $\{\phi_1, \ldots, \phi_k\}$.
3. **Global Softmax :** Ensures

$$\sum_j P(\text{death in interval } j) = 1.$$

Smoothing Regularization :

$$\text{Penalty} = \gamma \sum_{j=1}^{k-1} \|\theta_{j+1} - \theta_j\|_2^2$$

Model Comparison and Ensemble Strategy

Property	XGBoost AFT	Neural-MTLR
Model type	Parametric AFT (flexible η)	Non-parametric (discrete time)
Core assumption	Log-linearity in time	None on hazard structure
Strength	Robust with limited sample size	Captures non-proportional hazards
Output	Scalar log-time	Full survival distribution

Rank-Based Ensemble

To combine complementary strengths while reducing variance, we aggregate predictions by rank :

$$\text{Score}_{\text{final}} = \text{mean}(\text{rank}(\hat{y}_{\text{AFT}}) + \text{rank}(P_{\text{MTLR}}))$$

This approach is naturally robust to heterogeneous output scales.

Effect of Ensembling — XGBoost AFT

IPCW C-index (5-fold nested cross-validation)

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

Paired comparison (fold-wise) :

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

Interpretation

- Small but consistent improvement.
- Borderline statistical significance at the 5% level.
- Ensemble reduces variance and stabilises rankings.

Effect of Ensembling — Neural-MTLR

IPCW C-index (5-fold nested cross-validation)

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

Paired comparison (fold-wise) :

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

Interpretation

- Clear and statistically significant improvement.
- Rank-based ensembling is particularly beneficial here.
- Suggests complementary inductive bias across random seeds.