Predicting Survival Time After Bone Marrow Transplant

The Erdős Institute
Spring 2025

Team: Ray Karpman, Yang Li, Elzbieta Polak, Chi-Hao Wu, Ruibo Zhang

Mentor: Shravan Patankar

Hematopoietic Stem Cell Transplant (HCT)

- Also called bone marrow transplant.
- Key treatment for blood cancers.
- Five-year survival less than 49%.
    - (Wong, 2020)
- Goal: predict survival times for patients after HCT.
    - Data: CIBMTR via kaggle.com

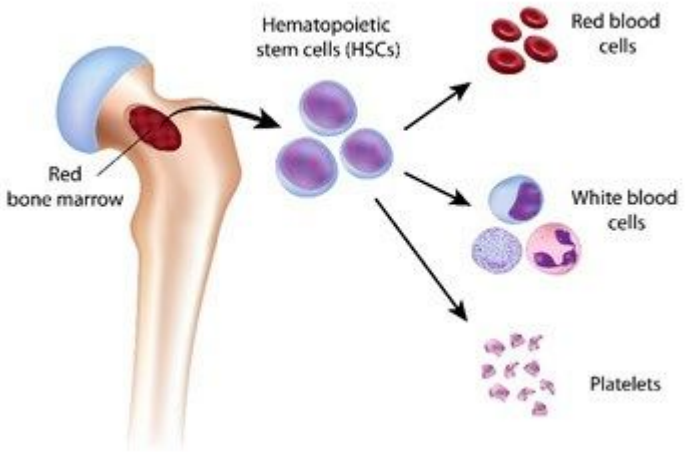
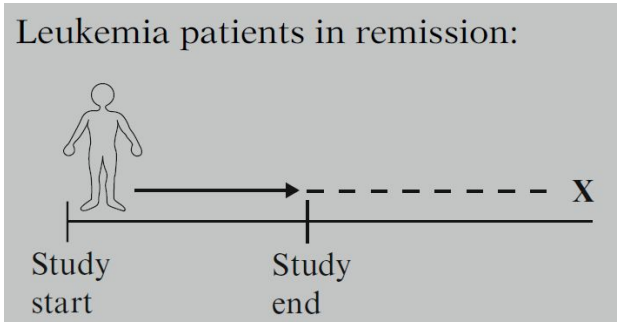


Figure: Hematopoietic stem cells in bone marrow produce all blood cells.

Challenge: Censored Data

- Problem: Many patients exit study before an event (death or relapse)
  - *Censored observations.* True survival time not known.

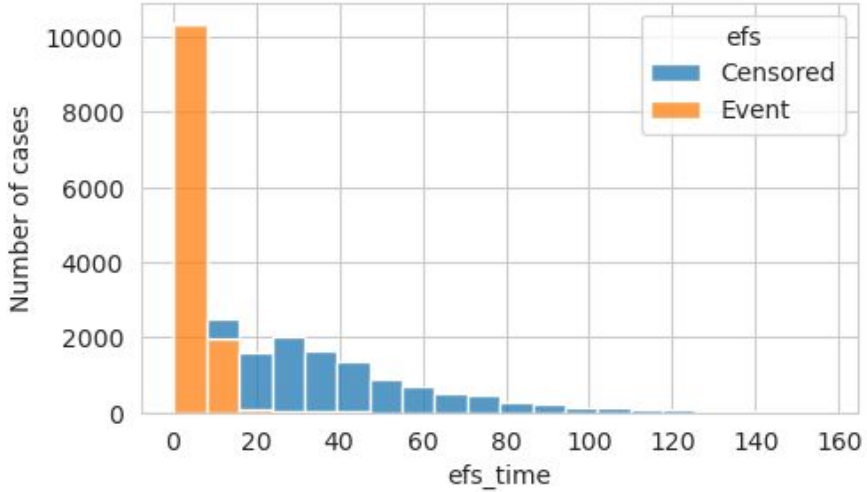


- Requires specialized methods from *survival analysis*.

DG Kleinbaum, M Klein (1996)

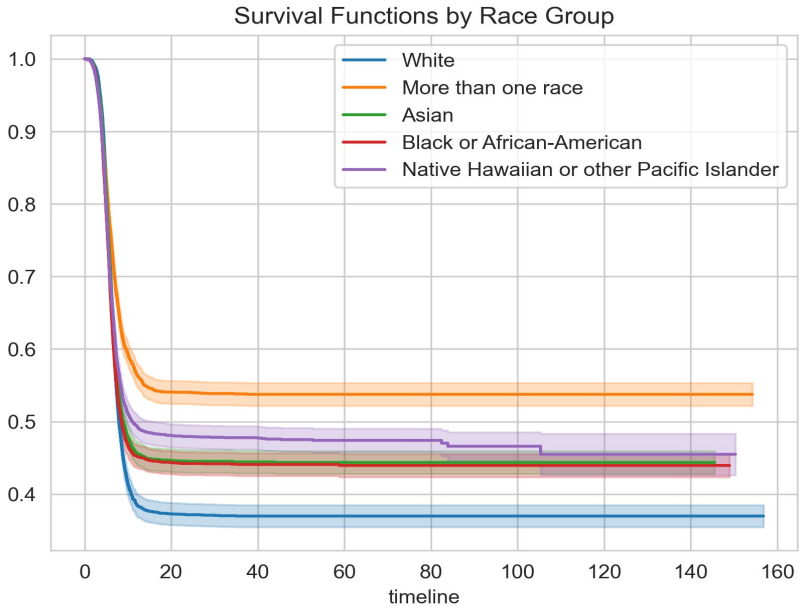
EDA

- The dataset dimensionality: 58 feature columns and 2 target columns.
- Target columns:
  - **efs (event-free survival):** whether an adverse event has occurred (1 or 0).
  - **efs_time (event-free survival time):** event-free survival time in months.



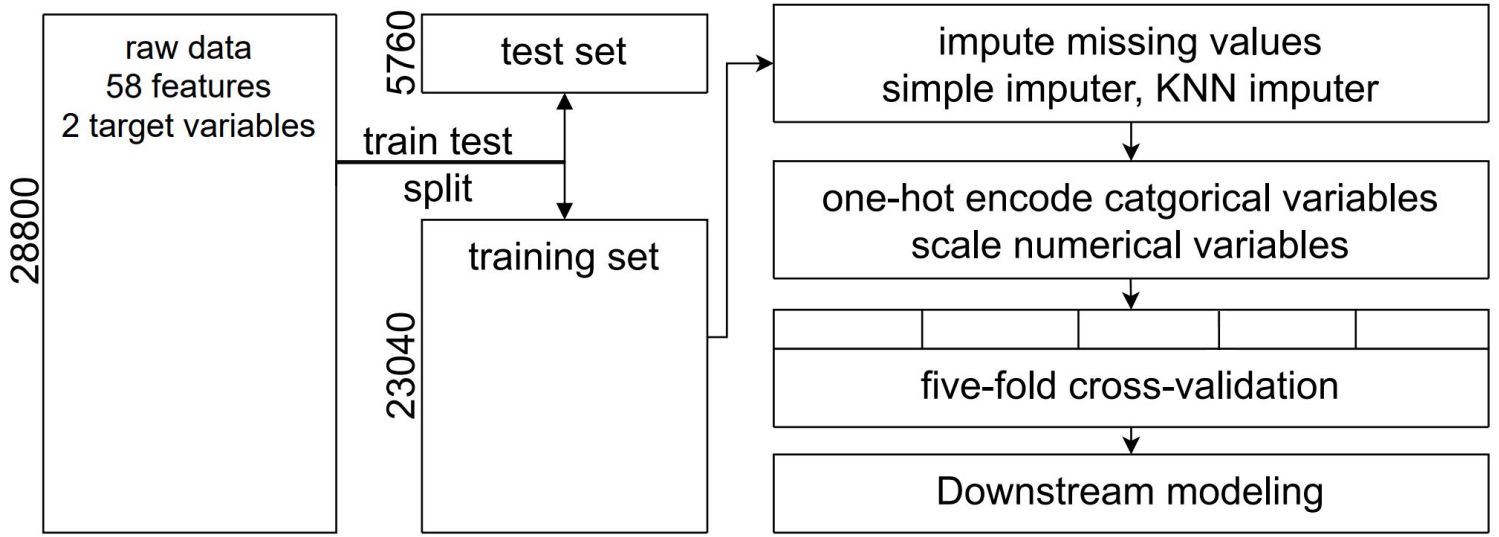
Stratified Concordance Score (SC-Index)

- There is a strong correlation between survival rates and the race group:



- These survival discrepancies are the reason why it is preferred to compute a **metric** called **stratified concordance index**, along the race group.

Data Preprocessing / Imputation



Challenge: Missing features in the test set

KNN Imputer: Data imputation based on k-nearest neighbors.

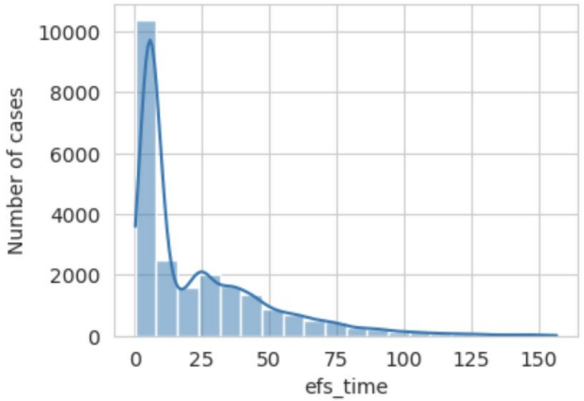
Target quantities

Target quantities:

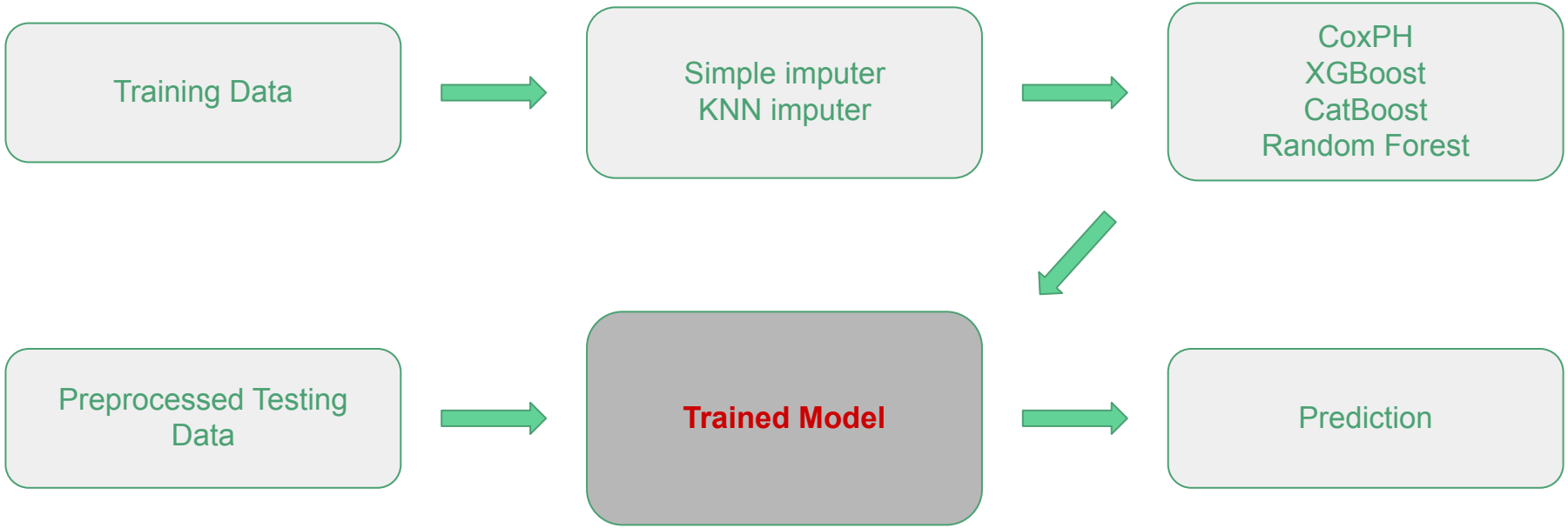
- **Hazard rate**: measure the risk that an event (death) happens to a patient
- **Expected survival time**: measure the expected time a patient lives

Sample models:

- **Cox proportional hazard model (CoxPH)**
- **accelerated failure time model (AFT)**



Model Training



Results

Stakeholder Metric: Stratified Concordance Index (SC-index)

- Similar to AUC, SC-index is between (0, 1). **1 means perfect prediction.**

Baseline (CoxPH): 0.6507

Kaggle Winner: 0.7012

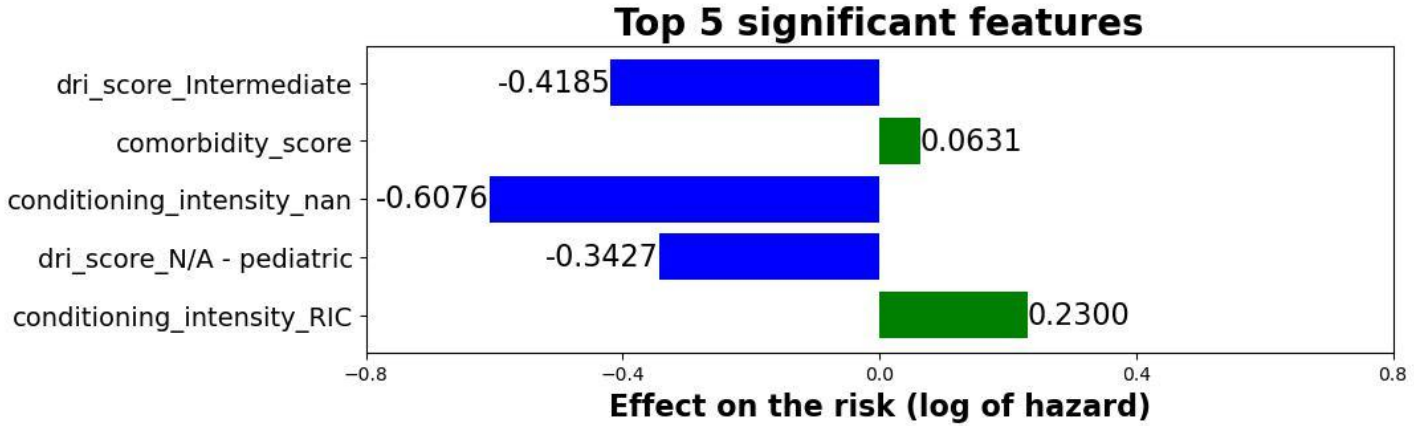
Fine-tuned Models	CoxPH	XGboost	Survival Random Forest	CatBoost
Performance CV 80%	0.6523	0.6533	0.6340	0.6578
Performance Test Set 20%	0.6532	0.6545	0.6334	0.6581

We select the tuned CoxPH model to be our final solution.

- The coefficients of CoxPH directly explains the effect of each feature on the risk.

Conclusion

Clinical impact: Our analysis reveals key features linked to post-HCT outcomes, which may assist physicians in risk stratification and treatment planning.



Model performance: A SC-Index of 0.653 suggests the model effectively captures survival patterns despite data complexity. Robust preprocessing and survival modeling enhance interpretability and utility.

Future directions: Explore deep learning-based survival models to boost accuracy.

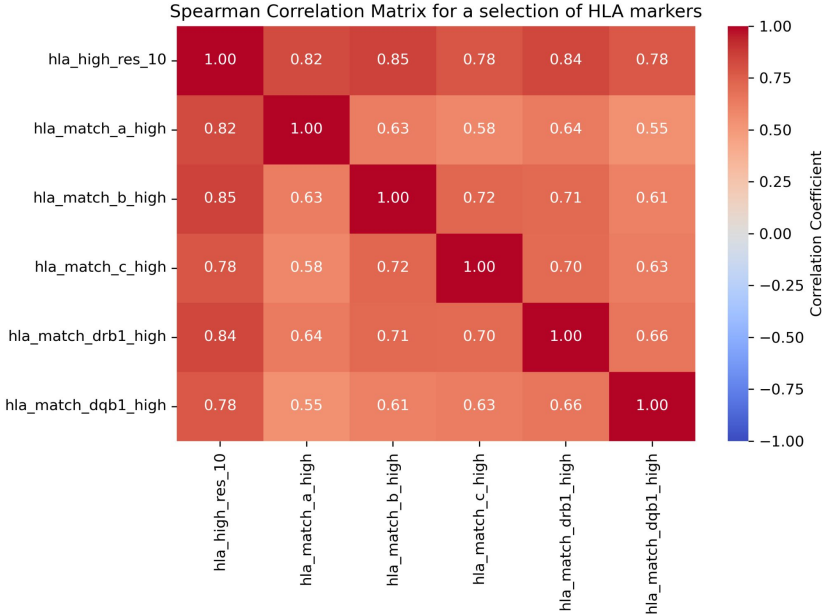
References

- Tushar Deshpande, Deniz Akdemir, Walter Reade, Ashley Chow, Maggie Demkin, and Yung-Tsi Bolon. CIBMTR - Equity in post-HCT Survival Predictions. https://kaggle.com/competitions/equity-post-HCT-survival-predictions, 2024. Kaggle.
- F Lennie Wong, Jennifer Berano Teh, Liezl Atencio, Tracey Stiller, Heeyoung Kim, Dayana Chanson, Stephen J Forman, Ryotaro Nakamura, Saro H Armenian, Conditional Survival, Cause-Specific Mortality, and Risk Factors of Late Mortality After Allogeneic Hematopoietic Cell Transplantation, JNCI: Journal of the National Cancer Institute, Volume 112, Issue 11, November 2020, Pages 1153–1161.
- Kleinbaum, David G., and Mitchel Klein. Survival analysis a self-learning text. Springer, 1996.

Thank you for listening!

(Bonus) EDA: HLA Features

- One of the most important groups of numerical features in the set are HLA features (17). HLA (Human Leukocyte Antigen) is a set of proteins found on the surface of cells that help the immune system identify foreign invaders.
- In bone marrow or stem cell transplants it's critical for the donor and recipient to have a good match in their HLA markers.
- The HLA features are very correlated:



(Bonus) Score function and Concordance index (C-index)

Score function: assign a score to each sample. A high score means a low survival time. (For example 100 / survival time (weeks)).

The C-index assesses the proportion of pairs of subjects where the model's predicted risk scores align with the observed event times.

T: survival time, η: estimated score, δ: indicator of censored data

$$\text{C-index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j}$$

(Bonus) C-index Example

$$\text{C-index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j}$$

	T	Censored(δ)
Patient 1:	20	1
Patient 2:	30	1
Patient 3:	10	0
Patient 4:	40	0

Concordant pairs (ordered)

(2,1) (3,1) (4,1)

(1,2) (3,2) (4,2)

Ti > Tj concordant pairs

(2,1) (4,1)

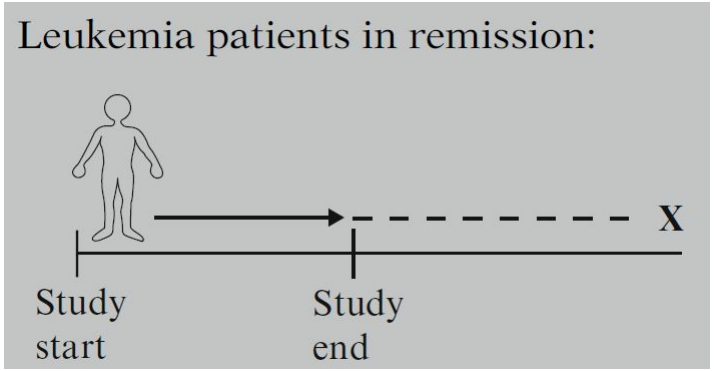
(4,2)

Perfect scores: $\eta 4 > \eta 2 > \eta 1$

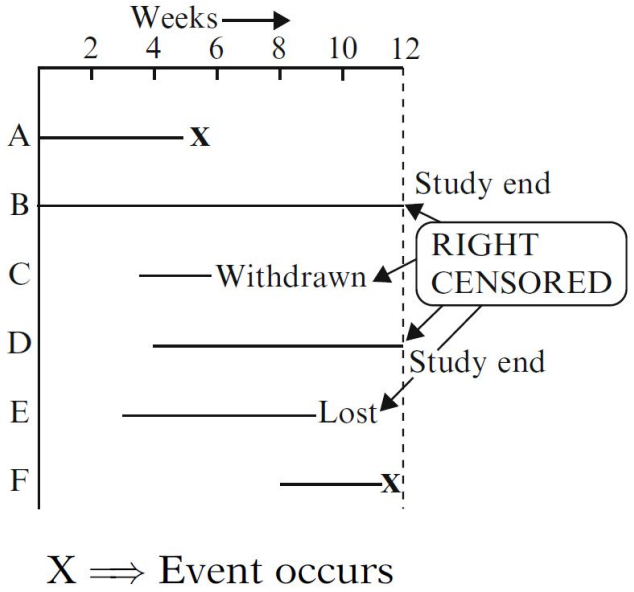
(Bonus) More on Censored Data

Outcome variable: Time until an event occurs

Censoring: we don't know survival time exactly.



Right-censored: true survival time is equal to or greater than observed survival time



Some other events: death, ending subscription, etc.

(Bonus) More on Target quantities

Survival function models the probability that the patient lives longer the time t:

$$S(t) = \mathbb{P}(T > t)$$

Hazard function describes the intensity that the patient dies at the next instant, assuming the patient has lived for time t:

$$h(t) = -\frac{S'(t)}{S(t)}$$

With these quantities, suitable candidates for risk scores include expected survival time or the value of the hazard function.

(Bonus) More on CPH and AFT

Cox proportional hazard model (CPH): Assume the ratio of the hazards is

$$\frac{h(t;X)}{h(t;X')} = \frac{\exp\left(X \cdot \beta\right)}{\exp\left(X' \cdot \beta\right)}$$

Accelerated failure time model (AFT):

$$\log \frac{T}{T'} = \sum_{i=1}^{p} \beta_i (X_i - X_i')$$

CPH models the ratio of the hazard functions, while AFT models the log ratio of the survival times. Using machine learning techniques, we estimate the optimal β.