Equity in Post–HCT Survival Predictions

Modeling, Fairness, and Ensemble Methods

Valeria Izvoreanu

December 4, 2025

Outline

1. Project Overview
2. Data Overview
3. Models
4. Results
5. Conclusions
6. Conclusions

Motivation

- Predicting post–hematopoietic stem cell transplant (HCT) survival.
- Build accurate survival prediction and risk-ranking models.
- Explore different of approaches: classical statistics, gradient-boosted trees, and deep survival models.

Workflow Summary

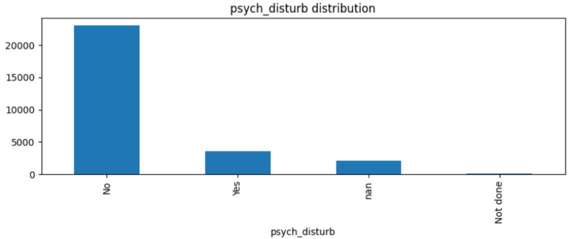
The notebook follows a linear predictive pipeline:

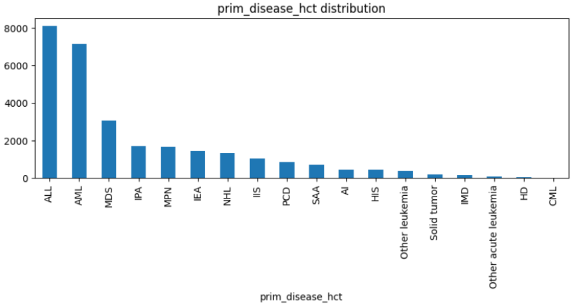
1. Environment setup and data loading.
2. Exploratory Data Analysis (EDA).
3. Feature engineering & preprocessing.
4. Evaluation functions (C-Index, stratified metrics).
5. Logistic Regression.
6. DeepSurv model.
7. Weighted XGBoost.
8. Ensemble of all three models.

Exploratory Data Analysis: Main Findings

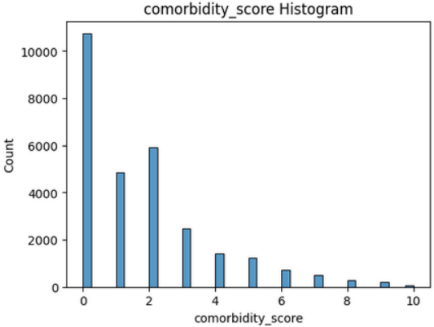
- Several variables have a lot of missing data, requiring careful preprocessing.
- Many categorical fields show **severe imbalance** with rare labels (e.g., psych_disturb, primary disease categories).
- Some variables encoded as numeric (e.g., HLA mismatch values) are **biologically categorical**.
- Some numerical features display **strong skew** and **outliers** (e.g., comorbidity score).

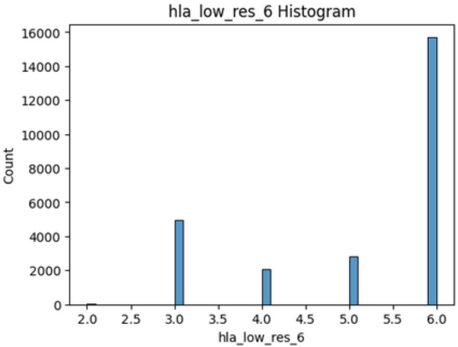
Categorical Variables: Imbalances and Structure





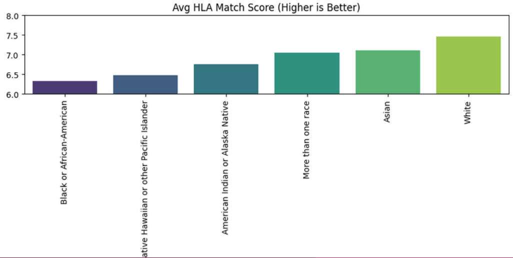
Numerical Variables: Skewness, Outliers, and Interpretation

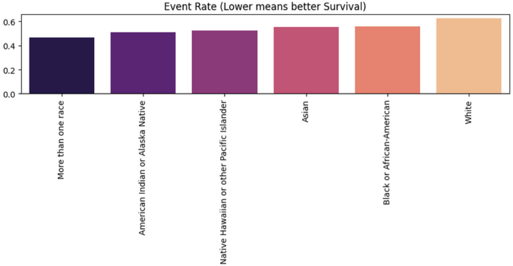




HLA Compatibility and Survival Outcomes by Race

- **HLA match scores** differ across racial groups, with White patients showing the highest compatibility.
- However, **event rates** do not follow the same pattern—some groups with lower HLA compatibility show comparable survival.
- This mismatch suggests **complex, nonlinear relationships** between HLA, race, and outcomes.
- This motivates the use of flexible models such as **XGBoost**, which can capture these patterns without relying on spurious correlations.





Feature Engineering

- Computed **HLA mismatch scores** as biologically meaningful risk indicators, because they inform models about donor–recipient immunological compatibility and the likelihood of adverse outcomes.
- Created **interaction features** capturing donor–recipient relationships (age differences, sex mismatch), which often influence transplant outcomes.
- Simplified highly imbalanced categorical variables into **binary or grouped categories** to reduce sparsity and stabilize model training.
- Encoded categorical variables using **one-hot encoding** and standardized continuous features.
- Imputed missing values through **median/mode filling** and domain-appropriate strategies to reduce bias introduced by uneven missingness.

C-Index (Concordance Index)

The C-Index evaluates how well the model ranks survival times. It is defined as the proportion of comparable pairs where the model correctly predicts a higher risk for the patient who experiences the event earlier.

$$C = \frac{\sum_{i<j} \mathbf{1}\left(T_i < T_j\right) \cdot \mathbf{1}\left(\hat{h}_i > \hat{h}_j\right)}{\sum_{i<j} \mathbf{1}\left(T_i < T_j\right)}$$

where:

- T_i is the observed survival time of patient i,
- $\hat{h}_i$ is the predicted risk or hazard score,
- The numerator counts correctly ordered risk pairs,
- The denominator counts all valid comparable pairs.

Goal: maximize C while maintaining fairness across subgroups.

Logistic Regression

- Logistic Regression serves as the baseline model for comparison.
- **Performance:**
  - **F1-score: 0.6829**
  - **Accuracy: 0.6704**
  - **ROC-AUC: 0.7432**
  - **C-Index: 0.6478**
- Has a good performance, but its linear nature limits its ability to capture nonlinear interactions present in HCT clinical variables.
- It cannot fully represent complex patterns such as donor–recipient matching, HLA effects, or disease-stage interactions, which reduces its ranking power.

DeepSurv

- DeepSurv is a neural network extension of the **Cox proportional hazards model**, allowing nonlinear representations while preserving the Cox ranking objective.
- It can learn complex interactions among clinical variables, donor characteristics, and HLA features.
- **Performance:**
  - **C-Index: 0.6575**
  - **ROC-AUC: 0.7316**
  - **F1-score: 0.6777**
  - **Accuracy: 0.6650**
- Provides improved discrimination over linear models by capturing nonlinear structure in the data.
- Still inherits Cox-model constraints: it sorts patients from highest to lowest risk, but it cannot estimate actual survival probabilities. It also assumes that risk factors affect patients the same way over time, which may not always match real-life medical behavior.

Weighted XGBoost (Cox Model)

- XGBoost is a **gradient-boosted tree ensemble** capable of learning nonlinear interactions through additive decision-tree structures.
- In this project, XGBoost is trained with **custom sample weights** that give higher penalty to misclassified **failure events**, especially in underrepresented subgroups.
- **Performance:**
  - **C-Index: 0.6354**
  - **ROC-AUC: 0.5920**
  - **F1-score: 0.5764**
  - **Accuracy: 0.5597**
- The weighting scheme strengthens event ranking for minority and low-frequency patterns but leads to a **drop in global accuracy and ROC-AUC**, showing a trade-off between robust subgroup performance and overall metrics.

Final Ensemble

- Combined:
  1. DeepSurv (best ranking power)
  2. Logistic Regression (statistical stability)
  3. Equity-XGBoost (fairness tuning)

Results Summary

Model	**C-Index**	**F1-score**	**Accuracy**
Logistic Regression	0.6487	0.6866	0.6743
DeepSurv	0.6584	0.6744	0.6616
Weighted XGBoost	0.6427	0.5924	0.5764
Final Ensemble	**0.6616**	**0.6646**	**0.6514**

(All models were evaluated after fine-tuning their hyperparameters.)

Discussion

- The **Final Ensemble achieves the highest C-Index (0.6616)**, making it the best-performing model for survival ranking in this task.
- Although threshold-based metrics (Accuracy, F1) vary across models, the **C-Index is the most clinically meaningful** because it evaluates whether patients are correctly ordered by relative risk.
- In survival analysis, the goal is not to classify patients as “fail” or “survive,” but to identify those at **higher relative risk** so clinicians can prioritize monitoring, choose conditioning regimens, or evaluate donor options.

Discussion

- Logistic Regression performs well on classification metrics but is limited in ranking performance due to its linear structure.
- The weighted XGBoost model introduces sample-dependent penalties that improve the ranking of rare failure events, adding complementary signal despite lower overall metrics.
- DeepSurv benefits from nonlinear representation learning and provides the strongest individual model performance, which is reflected in its contribution to the ensemble.

How This Model Compares to Others

- The **Final Ensemble achieved a C-Score of 0.6616**.
- For context, the **top solution achieved a C-Score of 0.6980**, indicating that the ensemble model performs reasonably well but leaves room for improvement with more sophisticated architectures, feature engineering, or ensembling strategies.

The End

Thank you for your attention.