UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Universidad Distrital Francisco José de Caldas

Computer Engineering Program
School of Engineering

Workshop No. 1 – Kaggle Project Analysis: CIBMTR Equity in Post-HCT Survival Predictions

Sergio Nicolás Mendivelso Martínez – 20231020227
Sergio Leonardo Moreno Granado – 20242020091
Juan Manuel Otálora Hernández – 20242020018
Juan Diego Moreno Ramos – 20242020009

Professor: Eng. Carlos Andrés Sierra, M.Sc.

A report submitted for Workshop No. 1 in Systems Analysis & Design
Semester 2025-III

September 2025, Bogotá D.C.

Abstract

This workshop presents a comprehensive systems analysis of the CIBMTR (Center for International Blood and Marrow Transplant Research) Kaggle competition focused on equity in post-hematopoietic cell transplant (HCT) survival predictions. The competition challenges participants to develop machine learning models that accurately predict survival outcomes following allogeneic hematopoietic cell transplantation while ensuring fairness across different racial and socioeconomic groups.

Our analysis examines the complexity of the medical system surrounding HCT procedures, identifies key variables affecting patient outcomes, and evaluates the dual requirements of accuracy and equity in predictive modeling. The study reveals the intricate relationships between clinical, genetic, and demographic factors that influence post-transplant survival, highlighting the system's sensitivity to small parameter variations and the presence of chaotic behavior in medical outcomes.

Key findings include the identification of critical sensitivity parameters such as patient age, disease risk indices, genetic compatibility scores, and comorbidities. The analysis emphasizes the importance of the stratified C-index evaluation metric, which ensures model performance is consistent across ethnic subgroups. This work contributes to understanding how systems thinking can be applied to improve both the accuracy and fairness of medical prediction models, with direct implications for personalized medicine and transplant care.

The integration of recent advances in machine learning approaches for hematopoietic stem cell research demonstrates the evolving landscape of predictive modeling in healthcare. Our recommendations include implementing ensemble methods, maintaining demographic balance in cross-validation strategies, and incorporating domain expertise in feature engineering to address the inherent complexity and randomness in biological systems.

Keywords: Hematopoietic Cell Transplantation, Machine Learning, Healthcare Equity, Systems Analysis, Survival Prediction, Medical Informatics, Artificial Intelligence, Predictive Modeling

Contents

List of Figures

Chapter 1

Introduction and System Analysis

1.1 Overview

The "CIBMTR – Equity in Post-HCT Survival Predictions" competition has as its primary objective the development of models that accurately predict survival outcomes after allogeneic hematopoietic cell transplantation (HCT) Kaggle and CIBMTR (2025). What distinguishes this competition is its dual approach: predictions must not only be accurate but also equitable, addressing existing disparities by ensuring that models do not disadvantage patients based on their demographic or socioeconomic background.

The provided dataset contains clinical, genetic, and demographic information for each patient. Relevant variables include age, ethnicity, disease risk indices, genetic compatibility metrics, and detailed medical histories CIBMTR (2025). By leveraging this multifaceted data, the goal is to promote innovations in predictive modeling that can directly inform and improve personalized medicine and transplant care.

The main competition rules are as follows: Data use is strictly limited to the datasets provided within the competition framework; external data is not permitted unless explicitly authorized. Additionally, participants must meet equity criteria, as solutions will be evaluated not only for their overall accuracy but also for their equitable performance across different patient groups Kaggle and CIBMTR (2025).

1.2 Objectives

This workshop aims to achieve the following interconnected objectives:

- Understand the structure of the competition's system
- Learn more about the medical context to apply new ideas as accurately as possible
- Identify the system's elements, relationships, and boundaries
- Check the sensitivity and complexity of the problem
- Analyze the evaluation metric
- Propose recommendations to improve the accuracy and fairness of the models

1.3 Medical Background

1.3.1 Allogeneic Hematopoietic Cell Transplantation

Allogeneic Hematopoietic Cell Transplantation (HCT) is a medical procedure where a patient receives healthy blood stem cells from a compatible donor UpToDate (2025a). These new cells replace the patient's damaged or sick bone marrow and help rebuild a working blood and immune system. Recent advances in transplantation have significantly improved patient survival over time, though high morbidity and mortality risks remain critical challenges Auletta, Kou, Chen, and Shaw (2020).

The term allogeneic means the cells come from another person (a relative or someone unrelated). Although stem cells can be collected in different ways, bone marrow remains a primary source for transplantation Thomas (2016).

This procedure is mainly used to treat serious blood and immune system diseases, such as:

- Acute and chronic leukemias Salit and Deeg (2024)
- Lymphomas
- Multiple myeloma
- Severe aplastic anemia
- Immune disorders Zeng et al. (2025)

The goal is to replace faulty or harmful cells with stem cells that can create healthy new blood cells like red blood cells, white blood cells (T and B lymphocytes), and platelets UpToDate (2025a).

1.3.2 The Transplantation Process

The HCT process involves several critical stages UpToDate (2025a):

1. **Donor selection:** A genetic compatibility test (HLA typing) is performed. The better the match, the lower the risk of rejection. This is where prediction models from the competition become relevant.
2. **Patient preparation:** Before the transplant, the patient receives chemotherapy and/or radiation to:
   - Destroy diseased cells
   - Suppress the immune system to prevent rejection
3. **Stem cell infusion:** The transplant is not a surgical procedure. Stem cells are administered through an IV, similar to a blood transfusion.
4. **Engraftment:** The donor's stem cells travel to the patient's bone marrow and begin producing new blood cells. Recent studies have shown that robust and polyclonal engraftment occurs across different age groups, though with distinct differences in cellular diversity Kucab et al. (2024).
5. **Follow-up:** Patients require close monitoring to detect and manage complications.

1.3.3 Risks and Complexity

Because the transplant uses cells from another person, there is a risk that the patient's body may recognize them as foreign. Laboratory tests are conducted to reduce the likelihood of rejection or immune reactions UpToDate (2025b).

Possible complications include:

- **Graft-versus-host disease (GVHD):** The donor's immune cells may attack the patient's tissues UpToDate (2025b); Williams, Chien, Gladwin, and Pavletic (2021)
- **Infections**
- **Graft failure:** The new cells fail to engraft or function properly
- **Social and geographic factors:** These can affect access to care, treatment follow-up, and overall outcomes Zubarovskaya et al. (2023)

1.3.4 Importance of Survival Prediction

Allogeneic transplants are high-risk and expensive procedures Auletta et al. (2020). Accurate survival prediction:

- Helps physicians make better decisions about transplant candidacy
- Enables more personalized treatment and care protocols
- Improves utilization of healthcare resources
- Reduces health inequalities, as current models often favor certain population groups Doherty, Char, Goodman, Shah, and Oberst (2024)

1.4 System Analysis

1.4.1 Input Data Sources and Characteristics

The predictive system incorporates comprehensive data sources spanning clinical, transplant-specific, demographic, and temporal dimensions:

Disease and Clinical Characteristics

- Disease characteristics: Type of hematological malignancy, disease stage at transplant, remission status
- Pre-transplant comorbidity indices: Existing health conditions that affect transplant outcomes
- Previous treatment history: Chemotherapy cycles, radiation exposure, prior transplants
- Laboratory values: Blood counts, organ function markers, inflammatory markers
- Karnofsky/Lansky performance scores: Functional status indicators

Transplant-Specific Inputs

- Donor type: Matched related, matched unrelated, haploidentical, cord blood
- HLA matching degree: Level of compatibility between donor and recipient
- Stem cell source: Bone marrow, peripheral blood, umbilical cord
- Conditioning regimen intensity: Myeloablative, reduced intensity, non-myeloablative
- Graft-versus-host disease prophylaxis protocol

Demographic and Socioeconomic Inputs

- Age at transplant (continuous variable)
- Sex/gender (categorical)
- Race/ethnicity (categorical, critical for equity analysis)
- Geographic location (affects access to care)
- Insurance status (impacts follow-up care quality)

Temporal Inputs

- Year of transplant (captures medical advances over time)
- Time from diagnosis to transplant (affects disease progression)
- Follow-up duration markers

1.4.2 System Architecture and Modules

The predictive system employs a sophisticated modular architecture designed to ensure both accuracy and equity in survival predictions:

Data Preprocessing Module

This module handles initial data preparation through multiple stages:

- Feature engineering: Creating interaction terms between clinical and demographic variables, calculating risk scores from raw measurements, generating time-dependent features
- Data standardization and normalization while preserving demographic-related variations that might mask inequities
- Handling missing data and outliers with equity-aware imputation methods

Equity Analysis Module

This component specifically addresses fairness in the model:

- Stratified analysis across demographic groups to identify potential biases
- Bias detection algorithms measuring disparities in data quality, feature availability, and baseline outcome rates
- Fairness-aware preprocessing techniques, such as reweighting samples to balance representation across demographic groups while preserving clinical validity

Feature Selection and Importance Module

This system component determines which variables contribute most to survival predictions:

- Multiple selection strategies: Clinical domain knowledge integration, statistical significance testing, machine learning-based feature importance rankings
- Ensuring predictive features are available equitably across all patient populations
- Avoiding features that might be systematically missing for certain demographic groups

Predictive Modeling Core

The central prediction engine employs an ensemble approach:

- Survival analysis models: Cox proportional hazards and accelerated failure time models for time-to-event predictions
- Machine learning algorithms: Gradient boosting machines and random forests for complex pattern recognition
- Deep learning architectures for capturing non-linear relationships
- Cross-validation specifically designed for survival data with demographic stratification

Fairness Calibration Module

This post-processing component adjusts model predictions for equitable performance:

- Calibration techniques maintaining similar prediction accuracy across different patient populations
- Threshold optimization considering fairness metrics alongside traditional performance measures
- Disparity impact assessment to quantify and minimize prediction inequities

Uncertainty Quantification Module

This component provides confidence intervals and prediction uncertainties:

- Prediction intervals using techniques appropriate for survival analysis
- Risk stratification with associated uncertainty bounds
- Identification of cases where predictions might be less reliable due to data limitations

1.4.3 System Outputs

The system generates comprehensive outputs for clinical decision support and equity monitoring:

Primary Output

- Survival Probability Predictions: Time-dependent survival probabilities at key clinical milestones (100 days, 1 year, 2 years, 5 years post-transplant)
- Risk Stratification Categories: Patient classification into risk groups (low, intermediate, high) with associated survival curves

Secondary Outputs

- Equity Metrics Dashboard: Comprehensive fairness assessments including demographic parity measures showing prediction consistency across groups
- Clinical Decision Support Outputs: Individualized risk factors highlighting important predictors, comparative analysis showing patient risk profiles, treatment modification suggestions based on modifiable risk factors
- Model Interpretability Outputs: Feature importance rankings, partial dependence plots, SHAP values providing detailed explanations for individual predictions
- Quality Assurance Outputs: Model performance metrics, calibration plots, data quality reports identifying input data issues
- Research and Monitoring Outputs: Aggregate statistics for clinical research, temporal trend analyses, center-specific performance metrics for transplant program evaluation

1.5 Complexity and Sensitivity Analysis

1.5.1 Complexity in Post-HCT Survival Modeling

The challenge of predicting survival outcomes after hematopoietic cell transplantation (HCT) lies fundamentally in the system's complexity Auletta et al. (2020). Multiple interconnected clinical, genetic, and demographic factors influence patient survival, from age, disease stage, and comorbidities to donor compatibility and care protocols. The interaction of these variables forms a nonlinear, high-dimensional system where unexpected effects can arise from subtle parameter changes Harrington et al. (2025).

Furthermore, feedback effects, such as immune responses, transplant complications, or secondary interventions, can introduce new layers of complexity, making simple models insufficient to capture the true patient trajectory Kucab et al. (2024).

1.5.2 Sensitivity Analysis

Small variations in input variables can cause significant differences in survival predictions. For example, changing a patient's age or disease risk index value can substantially modify the survival estimate, relocating the patient to another risk group or altering their prognosis Zubarovskaya et al. (2023). Similarly, the presence or absence of certain comorbidities (such as renal failure or previous infections) can generate major changes in the final model outcome.

A technique used to measure this complexity is subgroup analysis, whereby dividing the sample into subgroups (by age, gender) and examining how predictions change can analyze the system's sensitivity to these changes.

Sensitive parameters include:

- **Age:** An increase of only one year can move a patient from a low to high-risk category Kucab et al. (2024)
- **Disease risk index:** Small changes in the score can alter prognosis or eligibility for specific treatments
- **Genetic compatibility:** Minimal differences can affect transplant rejection rates and survival
- **Severe comorbidities:** Their presence or absence can double or drastically reduce projected survival

1.5.3 Chaos and Randomness in Post-HCT Survival Prediction

Medical outcomes following allogeneic hematopoietic cell transplantation (HCT) are intrinsically unpredictable and complex, with many processes exhibiting characteristics of chaos and randomness Harrington et al. (2025). Chaos theory, which describes how small variations in initial conditions can lead to vastly different outcomes, is highly relevant for survival modeling in this context. Even small inaccuracies or variations in clinical measurements, such as patient age, genetic compatibility, or comorbidities, can dramatically change survival predictions.

This sensitivity is sometimes known as the "butterfly effect" in nonlinear systems. Randomness also plays a key role. Patient data contain elements of stochasticity arising from biological diversity, incomplete records, and unknown confounding factors Zeng et al. (2025). For example, two patients with apparently identical risk factors may experience completely different outcomes after transplantation due to hidden genetic or environmental variables.

Therefore, the proposed model must be robust against noise, outliers, and missing data, as these introduce additional layers of unpredictability in survival outcomes. Furthermore, feedback loops and nonlinearities, such as immune responses or treatment complications, can give rise to emergent behaviors that are difficult to predict using traditional linear models Krummey and Gareau (2022). These dynamic interactions illustrate how system evolution cannot be fully explained or predicted by examining its individual parts in isolation. Even the most advanced predictive models may struggle to capture all aspects of post-HCT survival, highlighting the ever-present influence of chaos and randomness in personalized medicine.

1.6 Evaluation and Metrics

1.6.1 The C-Index Metric

The C-Index is a metric used to evaluate a survival model's ability to correctly order risk among pairs of patients Kaggle and CIBMTR (2025). In the Kaggle competition on post-HCT survival, a "stratified" version is used, meaning the C-index is calculated within each racial group of patients and then these results are averaged with dispersion adjustment. This stratified C-Index dictates that our model must be not only accurate overall but consistent and fair across all ethnic subgroups present in the dataset.

1.6.2 Importance of Equity in Medical Predictions

Equity in medical predictions is fundamental because it ensures that all patients, regardless of their race, gender, socioeconomic status, or other factors, receive fair and accurate care Doherty et al. (2024). When predictive models do not consider equity, they can perpetuate existing inequalities in the health system.

1.7 Visual Representation

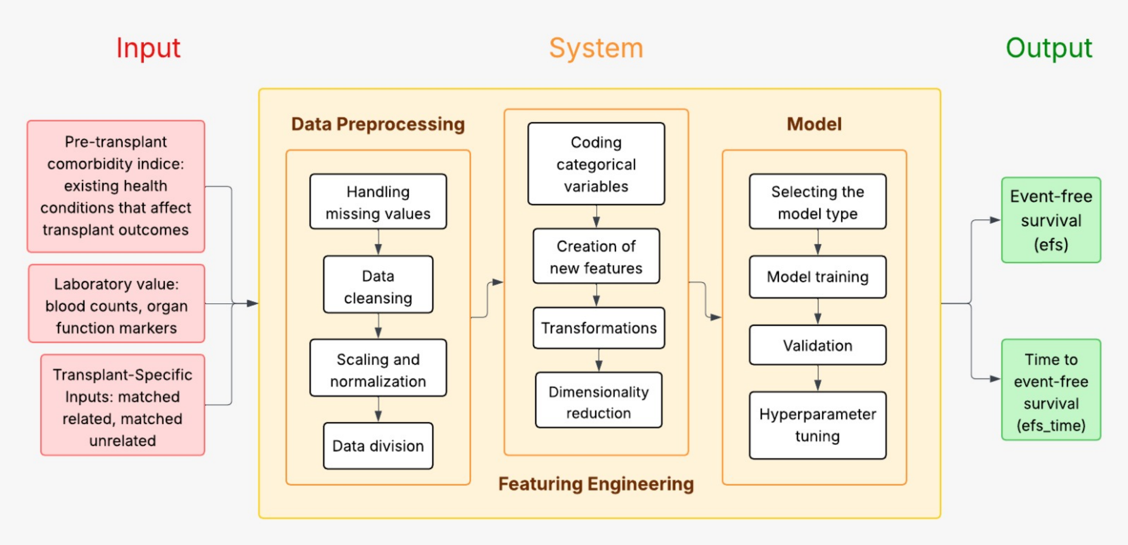
Input
System
Output
Pre-transplant comorbidity indice: existing health conditions that affect transplant outcomes
Laboratory value: blood counts, organ function markers
Transplant-Specific Inputs: matched related, matched unrelated
Data Preprocessing
Handling missing values
Data cleansing
Scaling and normalization
Data division
Coding categorical variables
Creation of new features
Transformations
Dimensionality reduction
Featuring Engineering
Model
Selecting the model type
Model training
Validation
Hyperparameter tuning
Event-free survival (efs)
Time to event-free survival (efs_time)

Figure 1.1: Diagram illustrating the modular pipeline of the predictive system. The diagram shows the flow from input data sources through preprocessing, equity analysis, feature selection, predictive modeling, fairness calibration, and uncertainty quantification modules, culminating in the output of survival predictions and equity metrics.

1.8 Conclusions

1.8.1 Key Findings Summary

This analysis reveals several critical aspects of the CIBMTR competition system:

- The complexity of post-HCT survival prediction stems from multiple interacting clinical, demographic, and transplant-specific factors
- The modular system architecture enables both accurate predictions and equitable performance across demographic groups
- Small parameter variations can significantly impact predictions, requiring robust sensitivity analysis
- The stratified C-index metric ensures equitable model performance while maintaining clinical accuracy
- System sensitivity requires sophisticated modeling approaches with built-in fairness calibration

1.8.2 System Strengths and Weaknesses

Strengths:

- Comprehensive multi-modular architecture addressing both accuracy and equity Kaggle and CIBMTR (2025)
- Integrated fairness calibration throughout the prediction pipeline
- Real-world medical application with high impact potential Auletta et al. (2020)
- Sophisticated handling of clinical complexity through ensemble modeling approaches Harrington et al. (2025)
- Comprehensive output system supporting clinical decision-making and equity monitoring

Weaknesses:

- High system complexity requires substantial computational resources
- Inherent randomness and chaos in biological systems Zeng et al. (2025)
- Potential for hidden confounding variables not captured in current data sources
- Dependence on data quality and completeness across all demographic groups
- Challenges in generalizing across different healthcare systems and populations

1.8.3 Recommendations

To improve accuracy and equity in predictions:

- Implement advanced ensemble methods to handle system complexity while maintaining interpretability Harrington et al. (2025)
- Develop cross-validation strategies that maintain demographic balance across all folds
- Create robust preprocessing techniques specifically designed for clinical missing data patterns
- Incorporate domain expertise in feature engineering to ensure clinical relevance
- Apply fairness-aware regularization techniques to prevent overfitting to majority groups
- Consider temporal dynamics and evolving medical practices in patient outcomes Kucab et al. (2024)
- Enhance uncertainty quantification to support clinical decision-making under uncertainty

1.8.4 Clinical and Healthcare System Implications

This work has significant implications for:

- Enhanced clinical decision-making in transplant medicine through personalized risk assessment Auletta et al. (2020)
- More equitable allocation of medical resources based on accurate and fair predictions
- Advancement of personalized medicine approaches through comprehensive risk profiling Zeng et al. (2025)
- Reduction of healthcare disparities through systematic bias detection and mitigation Doherty et al. (2024)
- Development of fair AI systems in healthcare that can be trusted across diverse populations
- Improved patient outcomes through better risk stratification and targeted interventions
- Quality improvement in transplant programs through comprehensive monitoring and evaluation outputs

References

Auletta, J. J., Kou, J., Chen, M., & Shaw, B. E. (2020). Indications for hematopoietic cell transplantation and immune effector cell therapy: Guidelines from the american society for transplantation and cellular therapy. *Biology of Blood and Marrow Transplantation*, *26*(7), 1247–1261. Retrieved from https://www.astctjournal.org/article/S1083-8791(20)30114-2/fulltext doi: 10.1016/j.bbmt.2020.03.002

CIBMTR. (2025). *Publicly available datasets.* Retrieved from https://cibmtr.org/CIBMTR/Resources/Publicly-Available-Datasets (Accessed: 2025-09-26)

Doherty, T. S., Char, D. S., Goodman, S. N., Shah, N. H., & Oberst, M. (2024). Addressing ai algorithmic bias in health care. *JAMA*. Retrieved from https://jamanetwork.com/journals/jama/fullarticle/2823006 doi: 10.1001/jama.2024.16564

Harrington, P., de Lima Zuchner, C., McGowan, R., Gormley, N., Hill, Q. A., & Snowden, J. A. (2025). Editorial: Improving stem cell transplantation delivery using computational modelling. *Frontiers in Immunology*, *16*. Retrieved from https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2025.1579353/full doi: 10.3389/fimmu.2025.1579353

Kaggle and CIBMTR. (2025). *Cibmtr - equity in post-hct survival predictions.* Retrieved from https://www.kaggle.com/competitions/equity-post-HCT-survival-predictions (Accessed: September 2025)

Krummey, S. M., & Gareau, A. J. (2022). Donor specific hla antibody in hematopoietic stem cell transplantation: Implications for donor selection. *Frontiers in Immunology*, *13*. Retrieved from https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2022.916200/full doi: 10.3389/fimmu.2022.916200

Kucab, M., Mitchell, T. J., Ariffin, H., Campbell, P. J., Rahman, N. A., Scelo, G., … Prokunina-Olsson, L. (2024). Characterization of clonal dynamics using duplex sequencing in diagnoses and relapses of acute myeloid leukemias. *Science Translational Medicine*, *16*(770), eado5108. Retrieved from https://www.science.org/doi/10.1126/scitranslmed.ado5108 doi: 10.1126/scitranslmed.ado5108

Salit, R. B., & Deeg, H. J. (2024). Transplant in all: who, when, and how? *Hematology*, *2024*(1), 93–100. Retrieved from https://ashpublications.org/hematology/article/2024/1/93/526246/Transplant-in-ALL-who-when-and-how doi: 10.1182/hematology.2024000503

Thomas, E. D. (2016). Allogeneic stem cell transplantation: A historical and scientific overview. *Cancer Research*, *76*(22), 6445–6451. Retrieved from https://aacrjournals.org/cancerres/article/76/22/6445/613932/Allogeneic-Stem-Cell-Transplantation-A-Historical doi: 10.1158/0008-5472.CAN-16-0892

UpToDate. (2025a). *Allogeneic hematopoietic cell transplantation: Indications, eligibility, and prognosis.* UpToDate. Retrieved from https://www.uptodate.com/contents/allogeneic-hematopoietic-cell-transplantation-indications

-eligibility-and-prognosis (Accessed: September 2025)

UpToDate. (2025b). *Clinical manifestations, diagnosis, and grading of acute graft-versus-host disease.* UpToDate. Retrieved from https://www.uptodate.com/contents/clinical-manifestations-diagnosis-and-grading-of-acute-graft-versus-host-disease (Accessed: September 2025)

Williams, K. M., Chien, J. W., Gladwin, M. T., & Pavletic, S. Z. (2021). Updates in chronic graft-versus-host disease. *Hematology*, *2021*(1), 648–654. Retrieved from https://ashpublications.org/hematology/article/2021/1/648/482938/Updates-in-chronic-graft-versus-host-disease doi: 10.1182/hematology.2021000301

Zeng, W., Yu, X., Chen, Y., Gu, W., Chen, S., Xu, X., ... Sun, L. (2025). *Curing autoimmune diabetes with islet and hematopoietic cell transplantation under nonmyeloablative and immunosuppression-free conditioning.* bioRxiv. Retrieved from https://www.biorxiv.org/content/10.1101/2025.09.05.673576v1 doi: 10.1101/2025.09.05.673576

Zubarovskaya, L. S., Moiseev, I. S., Vladovskaya, M. D., Mikhailova, N. B., Morozova, E. V., Bykova, T. A., ... Afanasyev, B. V. (2023). Trends in outcome of hematopoietic stem cell transplantation: 5000 transplantations and 30 years of single-center experience. *Cancers*, *15*(19), 4758. Retrieved from https://www.mdpi.com/2072-6694/15/19/4758 doi: 10.3390/cancers15194758