Bridging the Bias Gap: A Human-Centered Approach to Fair Post-Transplant Survival Prediction Across Racial Groups

Introduction:

ML models are used in healthcare to predict outcomes like cancer detection, survival rate and drug efficacy. Current predictive models often fall short in addressing disparities related to socioeconomic status, race, and geography. Addressing these gaps is crucial for enhancing patient care, optimizing resource utilization, and rebuilding trust in the healthcare system. These biased predictions are propagated due to imbalanced datasets and incorrect algorithms. This project tries to tackle this critical challenge in human-centered AI.

Problem statement:

Develop ML models to improve the prediction of transplant survival rates for patients undergoing allogeneic Hematopoietic Cell Transplantation (HCT) regardless of their ethnic and racial background.

Motivation:

I felt this is an important topic because I can potentially learn how to work on biases in medical dataset across different demographic groups and to develop models that provide fair predictions for the same. The goal is to address disparities by bridging diverse data sources, refining algorithms, and reducing biases to ensure equitable outcomes for patients across diverse race groups. Creating interpretable predictions that healthcare providers can trust across demographic groups is the main challenge of this project. Data analysis for bias, understanding fairness metrics, model development and fairness optimization will be the phases of the proposed project.

Dataset:

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.

Dataset description:

I felt this dataset is perfect because:

a. It encompasses a range of demographic and medical characteristics of both recipients and donors, such as age, sex, ethnicity, disease status, and treatment details.
b. It features equal representation across recipient racial categories including White, Asian, African American, Native American, Pacific Islander, and More than One Race.
c. There are no privacy concerns since this dataset is synthetically generated.

Background and Related Work:

First, let's discuss what the allogeneic hematopoietic cell transplantation (HCT) procedure is. The human immune system comprises cells that develop from hematopoietic stem cells, a special type of cells that reside in the bone marrow. These stem cells are responsible for generating all blood cells, including red blood cells, platelet-producing cells, and immune system cells such as T cells, B cells, neutrophils, and natural killer (NK) cells. HCT can be used to replace an individual's faulty hematopoietic stem cells with stem cells that can produce normal immune system cells. In other words, a successful HCT can help fix a person's immune system by introducing healthy stem cells into their body. When hematopoietic stem cells are transferred from one person to another, the recipient is referred to as the HCT recipient. The term "allogeneic" indicates that the stem cells being used come from someone else, the hematopoietic stem cell donor. If the HCT is successful, the donor's hematopoietic stem cells will replace the recipient's cells, producing blood and immune system cells that work correctly. The source of hematopoietic stem cells can be bone marrow, peripheral blood, or umbilical cord blood. Depending on the source of the stem cells, HCT procedures may be called bone marrow transplants (BMT), peripheral blood stem cell transplants, or cord blood transplants. [1]

The evaluation metric for survival prediction is discussed next. The Concordance index (C-index) represents the global assessment of the model discrimination power: this is the model's ability to correctly provide a reliable ranking of the survival times based on the individual risk scores. The standard C-index (SCI) is adjusted to account for racial stratification, thus ensuring that each racial group's outcomes are weighed equally in the model evaluation. The stratified c-index is calculated as the mean minus the standard deviation of the c-index scores calculated within the recipient race categories, i.e., the

score will be better if the mean c-index over the different race categories is large and the standard deviation of the c-indices over the race categories is small. This value will range from 0 to 1, 1 is the theoretical perfect score, but this value will practically be lower due to censored outcomes. [1]

Earlier various statistical methods were used to predict survival rates and to perform survival analysis. The Cox proportional hazards (CPH) model is the most widely used multivariate statistical model for survival analysis. In outcomes research, especially clinical trials, a hazard ratio is often estimated from a CPH model and is reported as the main measure of therapeutic efficacy. While powerful, the **main limitation** of the Cox model is that it assumes proportional hazards over time, which isn't always the case in complex clinical scenarios like HCT. It may also struggle with high-dimensional data where the number of predictors exceeds the number of events. [2]

Various machine learning and deep learning-based models are created on top of existing statistical solutions to predict these survival rates. One of the most cited papers in this field is the systematic review of ML techniques in HCT carried out by Gupta, et.al. It provides a thorough overview of terms, metrics, methods and results needed for this project. [3]

A stacked ensemble of Cox Proportional Hazard (CPH) regression and 7 machine-learning algorithms was applied to develop a prediction model. The stacked ensemble model achieved better predictive accuracy evaluated by C-index than other state-of-the-art competing risk models (ensemble model: 0.670; Cox-PH: 0.668; Random Survival Forest: 0.660; Dynamic DeepHit: 0.646). [4]

This study from 2024 is one of the latest studies in this field. It focused on all age groups combined data and used Naïve-Bayes machine learning models that incorporated longitudinal data. They were significantly better than models constructed from baseline variables alone at predicting whether patients would be alive or deceased at the given time points. This proof-of-concept study demonstrated that unlike traditional prognostic tools that use fixed variables for risk assessment, incorporating dynamic variability using clinical and laboratory data improves the prediction of mortality in patients undergoing HCT. [5]

While focusing on optimizing survival prediction in children, one of the studies combined Chaotic mapping Harris Hawk Optimization (CHHO) and enhanced the conventional k-means clustering procedure to form CHHO with Deep clustering Model (CHHO-DCM). It performs the effective clustering of instances with the advantage of both local and global optimization. [6] A study also used gradient boosting machine (GBM) for

predicting long-term survival after allogeneic HCT in patients with hematologic malignancies.[7]

A study was undertaken to monitor potential disparities in survival after allogeneic hematopoietic stem cell transplantation (HSCT) with the aim of optimizing access and outcomes for minority and low-income patients. The overall survival probability was 61.8% at 36 months. Non-Hispanic white (63.6%) and especially Hispanic patients (49.2%) had lower survival probabilities at 36 months than non-Hispanic Black patients (75.6%, $p = 0.04$) [8]. Compared to matched related donor and matched unrelated donor HCT, more ethnically diverse patients received mismatched unrelated donor, haploidentical donor, and cord blood HCT. [9]

Methodology:

XGBoost (Extreme Gradient Boosting) will be used as baseline. It is optimized for speed and performance, handling large datasets efficiently. It uses parallel processing and hardware optimization, making it faster than many other algorithms. Medical datasets frequently have missing values, which it inherently manages by learning the best direction to take when it encounters a missing value in a split. This reduces the need for extensive preprocessing. Healthcare data often contains complex interactions and nonlinear relationships. XGBoost's ability to capture these patterns makes it particularly suited for predicting survival outcomes post-transplant. Through regularization parameters like eta, gamma, and lambda, XGBoost provides control over model complexity, helping prevent overfitting—a common concern where the goal is to generalize well to unseen data. It offers insights into feature importance, which is invaluable for understanding which factors are most influential in predicting survival rates.

After setting a benchmark following tasks will be considered on top of the baseline:

- **Hyperparameter Tuning:** Tools like Grid Search will be used to fine-tune hyperparameters such as max_depth, n_estimators, and learning_rate for better performance.
- **Cross-Validation Strategy:** Stratified k-fold cross-validation will be explored to ensure that the model generalizes well across different subsets of the data, especially if the dataset is imbalanced.

- **Ensembling Methods:** Ensembling XGBoost with other models like LightGBM or CatBoost will be explored. Ensembling can capture different aspects of the data and potentially boost performance.

In the proposed solution, to understand impact and correlation of race/ethnicity on survival, techniques in EDA and data preprocessing will be performed. Using techniques in feature engineering, new features that capture the interaction between race/ethnicity and other variables, we can manage the weight given to race/ethnicity. Modeling will be based on following concepts:

- **Pairwise Ranking Loss Neural Network (PRL-NN):** The PRL-NN is designed to model survival data by focusing on the relative risk between pairs of subjects rather than predicting absolute survival times. The core idea is to train a neural network that can rank individuals based on their risk of experiencing the event. By customizing the loss function, we can weigh certain features more heavily. For instance, we might design the pairwise ranking loss to prioritize minimizing errors related to race and ethnicity.

Risk Score Function: For each individual i with feature vector x, the neural network computes a risk score s :

$$s_i = f(\mathbf{x}_i; \theta)$$

Pairwise Ranking Loss: The goal is to ensure that individuals who experience the event earlier have higher risk scores.

$$L = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \cdot \ell(s_i, s_j)$$

- **Event Masking**: This technique can be used to adjust the risk based on race and ethnicity predictions, increasing their impact on the outcome.

Ensemble approach of combining multiple models, including XGBoost and PRL-NN, leverages the strengths of each model, resulting in improved performance. Additionally, by using weighted predictions and focusing on ranking, the ensemble model achieves better overall performance.

Experimental Design:

Dataset #1:

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.

Dataset description:

The dataset consists of 59 variables related to hematopoietic stem cell transplantation (HSCT), encompassing a range of demographic and medical characteristics of both recipients and donors, such as age, sex, ethnicity, disease status, and treatment details. The data features equal representation across recipient racial categories including White, Asian, African-American, Native American, Pacific Islander, and More than One Race.

The primary outcome of interest is event-free survival, represented by the variable efs, while the time to event-free survival is captured by the variable efs_time. These two variables together encode the target for a censored time-to-event analysis.

- **train.csv** - the training set, with target efs (Event-free survival)
- **test.csv** - the test set; task is to predict the value of efs for this data
- **data_dictionary.csv** - a list of all features and targets used in dataset and their descriptions.

The **data_dictionary.csv** gives a good description (usage and meaning), type of data (numerical or categorical), and the unique values of each variable. Some important variables for this dataset are:

- Disease-related variables (dri_score, prim_disease_hct, cyto_score)
- Patient demographics (age_at_hct, race_group, ethnicity)
- Donor characteristics (donor_age, donor_related)
- HLA matching information (multiple variables like hla_high_res_8, hla_match_a_high)
- Comorbidities (diabetes, obesity, cardiac, pulmonary issues)
- Treatment details (conditioning_intensity, graft_type, rituximab, tbi_status)
- Outcome measures (efs, efs_time)

Data preprocessing:

1. **Handle Missing Values**: Many variables have 'nan' or missing values that need to be addressed. For categorical variables, a "Missing" category can be created, or imputation can be done. For numerical variables, mean/median imputation can be done.

2. **Encoding Categorical Variables**: Most categorical variables need encoding before analysis. One-hot encoding for variables with no natural ordering (like graft_type, sex_match) is used. Ordinal encoding for ordered categories (like dri_score: Low, Intermediate, High, Very high) is used.

3. **Standardize/Normalize Numerical Variables**: Variables like age_at_hct, donor_age, comorbidity_score need standardization to prevent one metric or unit from skewing results.

4. **Feature Engineering**: Creating interaction terms between relevant variables (e.g., donor-recipient matching variables) will help in better understanding the relationship. Since there are multiple comorbidity variables, we can try aggregating comorbidity information into composite scores. Deriving new features from existing ones (e.g., age difference between donor and recipient) is a good way to extract most information from the dataset.

5. **Data Validation**: For categorical variables to only contain expected values, missing values need to be filled/ or rows to be dropped. For numerical variables, they should be standardized/normalized and should be within reasonable ranges.

6. **Special Considerations**: Performing correlation analysis to identify which to keep. Variables like 'year_hct' introduce temporal trends, hence they need specialized handling. The complex categories in variables like 'gvhd_proph' need simplification.

7. **Target Variable Preparation**: There are 2 target variables: efs and efs_time. The outcome variable 'efs' (event-free survival) is categorical with 'Event' or 'Censoring'. Paired with 'efs_time' in months. We need to convert these to appropriate format for survival analysis (e.g., status indicator and time variable).

8. **Data Splitting**: Stratified sampling to maintain class distribution when splitting into training/validation/test sets will be used. Key prognostic variables like 'dri_score' or 'prim_disease_hct' will be focused on first.

9. **Class Imbalance**: Techniques like SMOTE and SMOTE-TOMEK will be used to ensure class imbalance is avoided.

Dataset #2:

Tush Blue, B. J., Brazauskas, R., Chen, K., Patel, J., Zeidan, A. M., Steinberg, A., Ballen, K., Kwok, J., Rotz, S. J., Perez, M. A. D., Kelkar, A. H., Ganguly, S., Wingard, J. R., Lad, D., Sharma, A., Badawy, S. M., Lazarus, H. M., Hashem, H., Szwajcer, D., Knight, J. M., ... Majhail, N. S. (2023). Racial and Socioeconomic Disparities in Long-Term Outcomes in ≥1 Year Allogeneic Hematopoietic Cell Transplantation Survivors: A CIBMTR Analysis. *Transplantation and cellular therapy*, *29*(11), 709.e1–709.e11. https://doi.org/10.1016/j.jtct.2023.07.013

Dataset description:

The dataset consists of 28 variables related to hematopoietic cell transplantation (HCT), encompassing a range of demographic and medical characteristics of recipients, such as age, sex, race, disease status, and treatment details. The data includes four recipient racial categories: Non-Hispanic white, Non-Hispanic black, Hispanic, and Asian.

The primary outcomes of interest are overall survival, represented by the variable "dead" and disease-free survival, represented by "dfs". The time measurements are captured by variables "intxsurv" (time from HCT to death/last follow-up) and "intxrel" (time from HCT to relapse). These variables together encode the targets for survival analysis.

- **Dataset name:** hs1802_datafile_05172021
- **Primary targets:** dead (Overall survival) and dfs (Disease free survival)

A detailed description of all features and targets used in the dataset is in **data_dictionary.docx**. The data dictionary provides descriptions, types, and possible values for each variable. Some important variables for this dataset are:

- Demographic variables (age, sex, racegp, marstat, ruralzip)
- Socioeconomic indicators (insurgrp, povarea)
- Disease-related variables (disease, drigp)
- Treatment details (condintb, tbigp, donorgp, graftype)
- Complications (agvhd, cgvhd1y)
- Outcome measures (dead, dfs, intxsurv, intxrel)

Experimental design:

Performance Metrics:

Overall Survival (OS) will be our primary endpoint using the "dead" variable to evaluate the effectiveness of different treatments with Kaplan-Meier analysis. Disease-Free Survival (DFS) will serve as our secondary metric by analyzing the "dfs" variable and

calculating time until relapse or death occurs. Treatment-Related Mortality (TRM) will be analyzed as a competing risk to relapse using the "trm" variable, providing crucial safety information for risk-benefit assessment. For analyzing GVHD outcomes, we'll include cumulative incidence of acute GVHD (agvhd variable) and chronic GVHD (cgvhd1y variable) as key metrics to assess treatment-specific toxicity profiles. We'll also calculate the area under the ROC curve (AUC) to evaluate our model's discriminative ability in predicting treatment-related mortality across different patient subgroups. Given the potential imbalance in mortality outcomes, F1-score will be incorporated to balance precision and recall, particularly when evaluating prediction models for high-risk subpopulations.

Experimental Procedure:

We will implement 5-fold stratified cross-validation to ensure balanced representation of key subgroups (disease type, recipient race) across all folds while maintaining statistical power. Data will be partitioned using an 80-20 train-test split with stratification to preserve the distribution of critical variables like disease risk index and donor type throughout testing. To enhance result stability and reduce selection bias, we will perform ten random repeats with different random seeds (42-51) and report both mean outcomes and confidence intervals.

Statistical Tests and Reproducibility:

Cox proportional hazards model to analyze survival outcomes while adjusting for covariates. Log-rank test to compare survival curves between different treatment groups. Primary comparison is different GVHD prophylaxis regimens (variable "gvhdgpn") on overall survival. Subgroup analysis by recipient race (variable "racegp") to identify potential disparities in outcomes. Non-parametric Wilcoxon rank-sum tests will be applied for comparing continuous variables that don't meet normality assumptions. For normally distributed continuous outcomes, we'll use t-tests or ANOVA with post-hoc Tukey tests to identify significant differences between multiple treatment groups. Fixed random seed (42) for all randomization procedures to ensure reproducibility. Standardized feature scaling applied consistently across all splits.

Dataset #3:

Turcotte, L. M., Wang, T., Beyer, K. M., Cole, S. W., Spellman, S. R., Allbee-Johnson, M., Williams, E., Zhou, Y., Verneris, M. R., Rizzo, J. D., & Knight, J. M. (2024). The health risk of social disadvantage is transplantable into a new host. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(30), e2404108121. https://doi.org/10.1073/pnas.2404108121

Dataset description:

The dataset consists of 56 variables related to hematopoietic cell transplantation (HCT), encompassing demographic and medical characteristics of both recipients and donors, including age, gender, race, disease status, and treatment details. The data categorizes recipient race into White and Non-White groups.

The primary outcomes of interest are overall survival, represented by the variable "dead" and disease-free survival, represented by "dfs". The time measurements are captured by variables "intxsurv" (interval from HCT to last contact date) and "intxrel" (interval from HCT to relapse/progression). These variables together encode the targets for survival analysis.

- **Dataset name:** IB20-03
- **Primary targets:** dead (Overall survival) and dfs (Disease-free survival)

A detailed description of all features and targets used in the dataset is in **Data Dictionary IB20-03.docx.** The data dictionary provides descriptions, types, and possible values for each variable. Some important variables for this dataset are:

- Demographic variables (Age, Sex, Racegp)
- Socioeconomic indicators (MedHHInc_zcta_adj_R, pctfpov_zcta_r, r_score, r_score_cat)
- Donor characteristics (dnrage, Dnrrace, d_score, d_score_cat)
- Disease-related variables (Disease, Status, Indxtx)
- Treatment details (Condint, TBI, Gvhdgp, Atgcampathgp)
- Outcome measures (dead, dfs, intxsurv, intxrel)

Experimental Design:

Performance Metrics:

Overall survival will be assessed as our primary outcome measure using the "dead" variable. Disease-free survival (DFS) will be evaluated as a composite endpoint combining the "rel" (relapse/progression) and survival status to provide insight into both disease control and patient longevity. Treatment-related mortality (TRM) will be analyzed

independently from disease relapse using the variable "trm" to distinguish between treatment complications and disease progression as competing causes of failure.

To comprehensively assess socioeconomic impacts, we'll employ the C-index (Harrell's concordance index) to measure the discriminative power of our predictive models for survival outcomes across SES quartiles. Net reclassification improvement (NRI) will be calculated to quantify the added value of including socioeconomic variables beyond traditional clinical factors in predicting one-year mortality and relapse. Additionally, calibration plots will be generated to assess how well the predicted probabilities align with observed outcomes across the SES spectrum, ensuring our models perform consistently for all patient subgroups.

Experimental Procedure:

We will employ 10-fold cross-validation stratified by disease type and recipient SES score quartiles to ensure balanced representation of socioeconomic factors across all analytical subsets. The dataset will be divided using a 75-25 train-validation split with temporal validation, where earlier transplants (by "Yeartx") are used for training and more recent ones for validation to simulate real-world implementation. Multiple iterations (n=15) will be performed with bootstrap resampling to generate robust confidence intervals and assess model stability across different patient populations.

Statistical Tests and Reproducibility:

Multivariate Cox regression to evaluate the impact of socioeconomic status (r_score) on outcomes while controlling for clinical variables. Competing risk analysis using Fine-Gray model to distinguish between relapse and non-relapse mortality. For comparing survival outcomes between specific treatment groups, log-rank tests will determine statistical significance in Kaplan-Meier curves. When proportional hazards assumptions are violated, we'll employ stratified analyses or time-dependent coefficients to ensure valid comparisons. Interaction tests will be performed to identify differential treatment effects across SES quartiles and racial groups. Association between recipient SES score quartiles (r_score_cat) and survival outcomes will be key to understanding the relation with social factors and survival outcomes. Fixed random seed (42) for all randomization procedures to ensure reproducibility. Standardized feature scaling applied consistently across all splits.

Sources:

1. https://www.kaggle.com/competitions/equity-post-HCT-survival-predictions

2. Deo, S. V., Deo, V., & Sundaram, V. (2021). Survival analysis-part 2: Cox proportional hazards model. Indian journal of thoracic and cardiovascular surgery, 37(2), 229–233. https://doi.org/10.1007/s12055-020-01108-7

3. Gupta, V., Braun, T. M., Chowdhury, M., Tewari, M., & Choi, S. W. (2020). A Systematic Review of Machine Learning Techniques in Hematopoietic Stem Cell Transplantation (HSCT). Sensors, 20(21), 6100. https://doi.org/10.3390/s20216100

4. Makoto Iwasaki, Junya Kanda, Yasuyuki Arai, Tadakazu Kondo, Takayuki Ishikawa, Yasunori Ueda, Kazunori Imada, Takashi Akasaka, Akihito Yonezawa, Kazuhiro Yago, Masaharu Nohgawa, Naoyuki Anzai, Toshinori Moriguchi, Toshiyuki Kitano, Mitsuru Itoh, Nobuyoshi Arima, Tomoharu Takeoka, Mitsumasa Watanabe, Hirokazu Hirata, Kosuke Asagoe, Isao Miyatsuka, Le My An, Masanori Miyanishi, Akifumi Takaori-Kondo,; on behalf of the Kyoto Stem Cell Transplantation Group (KSCTG), Establishment of a predictive model for GVHD-free, relapse-free survival after allogeneic HSCT using ensemble learning. Blood Adv 2022; 6 (8): 2618–2627. doi: https://doi.org/10.1182/bloodadvances.2021005800

5. Yiwang Zhou, Jesse Smith, Dinesh Keerthi, Cai Li, Yilun Sun, Suraj Sarvode Mothi, David C. Shyr, Barbara Spitzer, Andrew Harris, Avijit Chatterjee, Subrata Chatterjee, Roni Shouval, Swati Naik, Alice Bertaina, Jaap Jan Boelens, Brandon M. Triplett, Li Tang, Akshay Sharma; Longitudinal clinical data improve survival prediction after hematopoietic cell transplantation using machine learning. Blood Adv 2024; 8 (3): 686–698. doi: https://doi.org/10.1182/bloodadvances.2023011752

6. Arthi, R., Priscilla, G., Maidin, S., & Yang, Q. (2024). Optimizing Survival Prediction in Children Undergoing Hematopoietic Stem Cell Transplantation through Enhanced Chaotic Harris Hawk Deep Clustering. *Journal of Applied Data Sciences, 6*(1), 405-415. doi:https://doi.org/10.47738/jads.v6i1.468

7. Choi E, Jun T, Park H, Lee J, Lee K, Kim Y, Lee Y, Kang Y, Jeon M, Kang H, Woo J, Lee J Predicting Long-term Survival After Allogeneic Hematopoietic Cell Transplantation in Patients With Hematologic Malignancies: Machine Learning–Based Model Development and Validation. JMIR Med Inform 2022;10(3):e32313. URL: https://medinform.jmir.org/2022/3/e32313. DOI: 10.2196/32313

8. Garcia, L., Feinglass, J., Marfatia, H., Adekola, K., & Moreira, J. (2024). Evaluating Socioeconomic, Racial, and Ethnic Disparities in Survival Among Patients Undergoing Allogeneic Hematopoietic Stem Cell Transplants. Journal of racial and ethnic health disparities, 11(3), 1330–1338. https://doi.org/10.1007/s40615-023-01611-8

9. Auletta, J. J., Kou, J., Chen, M., Bolon, Y. T., Broglie, L., Bupp, C., Christianson, D., Cusatis, R. N., Devine, S. M., Eapen, M., Hamadani, M., Hengen, M., Lee, S. J., Moskop, A., Page, K. M., Pasquini, M. C., Perez, W. S., Phelan, R., Riches, M. L., Rizzo, J. D., ... Shaw, B. E. (2023). Real-World Data Showing Trends and Outcomes by Race and Ethnicity in Allogeneic Hematopoietic Cell Transplantation: A Report from the Center for International Blood and Marrow Transplant Research. Transplantation and cellular therapy, 29(6), 346.e1–346.e10. https://doi.org/10.1016/j.jtct.2023.03.007