Executive Summary: Predicting Survival Times After Bone Marrow Transplant

Ray Karpman, Yang Li, Elzbieta Polak, Chi-Hao Wu, Ruibo Zhang
Github: https://github.com/RuiboZhang98/CIBMTR_post_hct_survival

Introduction

Bone marrow transplant, also called hematopoietic stem-cell transplant (HCT), is a key treatment for blood cancers such as leukemia. In an HCT procedure, bone marrow is harvested from a donor and then infused into the recipient's bloodstream. Stem cells from the donor enter the recipient's marrow and replace the recipient's faulty cells. While HCT can be curative, it also has significant risks, with a five-year survival rate of around 49% (Wong, 2020). Our goal is to develop a model that can predict survival times for patients who have undergone HCT, using data shared by CIBMTR on Kaggle.com.

Dataset

The dataset is synthetic, but was generated by CIBMTR using deep-learning methods that capture the statistical properties of real patient data. There are 28799 observations total. The dataset has 58 features, which cover a range of demographic and medical characteristics of both recipients and donors. The target variable is jointly encoded by two columns:

1. **efs:** a binary flag which encodes whether a patient underwent an event (death or relapse), or left the study without an event
2. **efs_time:** time the patient died or left the study

For patients who left the study without experiencing an event, the exact survival time is unknown. This phenomenon, known as *censoring*, introduces unique challenges for predictive modeling.

Preprocessing

We split the data into an 80% training set and a 20% testing set. We considered several imputation strategies, including constant simple imputer, KNN imputer, random forest imputer, and Bayesian Ridge imputer. After comparing the distribution of the imputed data with available data, we decided to keep only the imputation from a constant simple imputer and a KNN imputer. We used one-hot encoding for categorical variables, which creates binary features for each level of a categorical feature.

Model Selection and Results

The Cox proportional hazard (CoxPH) model is a classic parametric model in survival analysis. The model is essentially a regression, so we used it as our baseline. We then use a five-fold cross-validation to compare different models, including:

- **CoxPH**
- **XGBoost**

- **Survival Random Forest**
- **CatBoost**
- **Hybrid models (combine logistic regression or random forest with regressors)**

After fine-tuning model parameters and imputation parameters, all the models improved. However, none of them is significantly better than the others. Considering this, we chose to use the fine-tuned hybrid CoxPH as the final model since it confers a mechanistic model that quantifies the effect of features. Our final model achieved a stratified concordance index of 0.653 on test data, indicating solid predictive performance.

By computing the confidence intervals and p-values for each feature, we identified a small set of 78 post-processed features as significant features. The significant features we identified include

- **DRI score:** Describe the disease-related parameters of a patient
- **Comorbidity score:** Predict the mortality of a patient with concurrent conditions
- **Conditioning density:** Reflect how aggressive the conditioning regimen is before receiving donor stem cells

With these insights, our analysis can assist physicians in treatment planning.

Conclusion

We developed a survival analysis model for patients who undergo bone marrow transplants, identifying key clinical features such as the Disease Risk Index and the comorbidity score that may help in risk stratification and treatment planning. Our hybrid CoxPH model achieved a stratified concordance index of 0.653, reflecting strong performance despite data complexity. Robust preprocessing and interpretable modeling contribute to reliable prediction. Future work will explore deep learning approaches to further improve performance.

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.