

A Survival-Driven Machine Learning Framework for Donor-Recipient Matching in Liver Transplantation: Predictive Ranking and Optimal Donor Profiling

by

Yingke Wang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2024

© Yingke Wang 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Liver transplantation is a life-saving treatment for patients with end-stage liver disease. However, donor organ scarcity and patient heterogeneity make finding the optimal donor-recipient matching a persistent challenge. Existing models and clinical scores are shown to be ineffective for large national datasets such as the United Network for Organ Sharing (UNOS). In this study, I present a comprehensive machine-learning-based approach to predict post-transplant survival probabilities at discrete clinical important time points and to derive a ranking score for donor-recipient compatibility. Furthermore, I developed a recipient-specific "optimal donor profile", enabling clinicians to quickly compare waiting-list patients to their ideal standard, streamlining allocation decisions. Empirical results demonstrate that my score's discriminative performance outperforms traditional methods while maintaining clinical interpretability. I further validate that the top compatibility list generated by our proposed scoring method is non-trivial, demonstrating statistically significant differences from the list produced by the traditional approach. By integrating these advances into a cohesive framework, our approach supports more nuanced donor-recipient matching and facilitates practical decision-making in real-world clinical settings.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Prof. Xi He and Prof. Sirisha Rambhatla, for their invaluable guidance on the methodology and evaluation metrics of this work. I appreciate their kindness in allowing me to complete my degree requirements early.

I also wish to thank our clinical collaborators at the University Health Network (UHN)—Eunice Tan, Yingji Sun, and Dr. Mamatha Bhat—for their clinical insights and direction, which ensured that this thesis would have a meaningful clinical impact. Without the support of all these individuals, I would not have been able to complete this thesis.

Additionally, I am grateful to my committee member, Prof. Paulo Alencar, for his consistent support and understanding. I am also thankful for the funding provided by Prof. Xi He, Prof. Sirisha Rambhatla, and Dr. Mamatha Bhat, and for their belief in my potential as a researcher.

I acknowledge the support from my lab mates and friends for supporting me whenever I struggle. Without any of the people I mentioned above, I cannot complete this work.

With regard to the use of the UNOS dataset, I would like to acknowledge the support of the Health Resources and Services Administration contract HHS250-2019-00001C. The content is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Table of Contents

Author's Declaration	ii
Abstract	iii
Acknowledgements	iv
List of Figures	viii
List of Tables	ix
1 Introduction	1
1.1 Problem Statement	2
1.1.1 Problem Setup	2
1.2 Our Contributions	3
2 Related Work	4
2.1 Introduction of Survival Analysis	4
2.2 Kaplan-Meier Estimator	6
2.3 The Cox Proportional Hazards Model	6
2.4 DeepSurv Model	7
2.5 DeepHit Model	7
2.6 Tree-based Survival Methods	8

2.6.1	XGBoost Model	9
2.7	Clinical Scoring Systems and Domain-Specific Indices	9
3	Preliminaries	10
3.1	C-index (Harrell’s Concordance Index)	10
3.2	K-Prototypes Clustering	11
3.3	Hypergeometric Test	11
3.4	Time-Dependent ROC Curves and AUC	12
4	Data and Preprocessing	14
4.1	Data Source (UNOS)	14
4.2	Study Population	14
4.3	Features Selection	15
4.4	Challenge and Previous Attempt with UNOS Dataset	15
5	Methodology	17
5.1	Motivation for Using XGBoost as the Prediction Model	17
5.2	Motivation for Using Multi-label Survival Probability Prediction and Ranking Score	18
5.3	Discrete Survival Time Point Construction	18
5.4	Survival XGBoost Loss Function Construction	19
5.4.1	Handling Uncensored Data	19
5.4.2	Handling Censored Data	19
5.4.3	Combining Loss Functions	20
5.5	Ranking Score Construction	20
5.6	Feature Selection and Model Refinement	21
5.6.1	Initial Feature Selection via Our Model	22
5.6.2	Statistical Validation of Features	22
5.6.3	Combining Feature Importance and Statistical Significance	22
5.6.4	Secondary Model and Compatibility Scoring	22

6	Experiment and Evaluation	24
6.1	Evaluating RMSE for Direct Survival Time Estimation	25
6.1.1	Evaluation Through Harrell’s C-statistic	26
6.1.2	Evaluation of Model Performance Using Time-Dependent AUC and ROC	29
6.2	Evaluating Optimal Donors Selection Using Hypergeometric Test	32
7	Impact and Application of the Proposed Framework	34
7.1	Evaluating Model Performance in Different MELD Group	34
7.2	Compute Average Donor Profile	38
7.2.1	Interpreting the Cluster Centers	39
8	Conclusion	41
8.1	Future Work	41
	References	43
	APPENDICES	49
A	Proof of Ranking Score as Survival Time Lower bound	50
A.0.1	Discrete Approximation	50
B	Study Samples and Features Selection of UNOS Dataset	52
B.1	Workflow of Preprocessing Study Samples	53
B.2	Workflow of Preprocessing Transplant Features	54
B.3	Workflow of Deriving Final Study Features	55

List of Figures

2.1	Related Works in Survival Analysis	5
5.1	Model Refinement Workflow	23
6.1	Top 20 Most Frequently Used Features	25
6.2	ROC Curves at Some Time Points	31
6.3	P-Values Distribution of Hypergeometric Test	32
7.1	t-SNE of Top Donor Profile Clusters	38
B.1	Flow Chart of Processing Study Samples	53
B.2	Flow Chart of Transplant Feature Selection	54
B.3	Flowchart of the feature selection process Used to Derive the Final Study Features	55

List of Tables

6.1	RMSE in Months of the Uncensored Data in Survival Models	26
6.2	C-Index between TX year of 2009-2019	27
6.3	Performance C-Index between TX year of 2009-2019 on 72-months Survival different models	28
6.4	AUC at 1-Year Post-Transplant	29
6.5	AUC at 3-Year Post-Transplant	30
6.6	AUC at 5-Year Post-Transplant	30
7.1	An Example Illustration of the Recipient Selection Process Low MELD Group	35
7.2	An Example Illustration of the Recipient Selection Process Medium MELD Group	35
7.3	An Example Illustration of the Recipient Selection Process High MELD Group	36
7.4	C-Index of Our Ranking between TX year of 2009-2019 Censoring on 72- month Survival on Different MELD Group	36
7.5	Cluster 1 Donor Features (Mean and Std)	39
7.6	Cluster 2 Donor Features (Mean and Std)	39
7.7	Cluster 3 Donor Features (Mean and Std)	40
7.8	Sampled Recipient Features	40

Chapter 1

Introduction

Liver transplant (LT), a surgical procedure that replaces a patient’s diseased liver with a healthy liver from a donor, remains the definitive treatment for disease and hepatocellular carcinoma, often providing long-term survival benefits and improved quality of life [50, 44, 26]. However, the allocation of donor livers is complicated by the persistent problem of organ shortage, where the demand for transplants far exceeds the supply [54]. This scarcity pressures clinicians to optimize both recipient clinical emergency and donor-recipient matching criteria to ensure that each precious draft gets allocated to the most needed potential recipient and offers the maximum possible survival benefit.

Current allocation systems primarily rely primarily on clinical urgency rather than the prediction of post-transplant outcomes. Standard metrics such as the Model for End-Stage Liver Disease (MELD) score and state-of-art liver pre-transplant risk metric MELD 3.0 score prioritize the sickest patients first, which is clinically justified but may not always guarantee the best long-term graft survival as it does not consider donor characteristics in their evaluation [31, 35]. To supplement urgency-based metrics, clinicians also use established risk prediction models and donor-recipient scoring systems, including the Balance of Risk (BAR) score [14], the Donor Risk Index (DRI) [16], and other specialized indices [47, 49]. While these tools provide valuable heuristics, they often fail to capture the full complexity of donor-recipient and do not consistently yield optimal outcomes over the years [51]. The complexity of donor-recipient matching arises from the heterogeneous nature of both donors and recipients. Donors vary in age, liver quality, and ischemia times, while recipients differ in disease etiology, MELD scores, and overall health status. Moreover, traditional risk models tend to apply fixed weights derived from historical cohorts, overlooking critical interaction between donor and recipient. Recent advances in machine learning (ML) and deep learning have shown promise in improving predictive accuracy for transplant outcomes

[15, 7, 9]. By harnessing high-dimensional clinical data, ML-driven approaches can offer more individualized survival estimates. Such approaches can identify non-obvious patterns in donor-recipient compatibility and optimize graft utility. Yet, despite these promises, these approaches are not fully integrated into clinical practice. Challenges related to data quality and model interpretability still remain [1].

1.1 Problem Statement

To sum up, liver transplant continues to face a critical shortage of suitable donors compared to the growing number of patients requiring transplants [54]. In addition, suboptimal donor-recipient matches can significantly increase the risk of graft failure and possibly higher mortality. The core challenge of providing a better matching is the inherent complexity of allocating organs: each recipient and donor presents a heterogeneous clinical profile [51]. So far, no universally accepted, data-driven algorithm exists to determine the most appropriate donor for a given recipient. Instead, surgeons and transplant coordinators often rely on clinical experience and limited scoring systems that do not comprehensively capture the multifaceted donor-recipient interactions.

Under these circumstances, there is a pressing need to reduce guesswork and improve resource allocation, ensuring that each available donor organ is utilized where it can bring about the greatest long-term benefit. Consequently, our overarching aim is to propose a framework capable of ensuring precision and efficiency in organ allocation, thereby increasing post-transplant survival rates and reducing adverse outcomes.

1.1.1 Problem Setup

To formalize the allocation challenge, we define a set of potential liver transplant recipients

$$\mathcal{R} = \{r_1, r_2, \dots, r_n\},$$

where each recipient r_i has a clinical feature vector $\mathbf{x}_{r_i}$, and a set of donors

$$\mathcal{D} = \{d_1, d_2, \dots, d_m\},$$

each with attributes $\mathbf{x}_{d_j}$. Our goal is to find a function

$$f: (\mathbf{x}_{r_i}, \mathbf{x}_{d_j}) \mapsto \text{compatibility score},$$

which accurately reflects the expected survival outcome $T(r_i, d_j)$ for the donor-recipient pair (r_i, d_j) . A higher score indicates a more favorable match, implying a greater likelihood of successful transplantation and prolonged graft survival.

Once this function is estimated, for each recipient r_i we can then identify the donor d_j^* that maximizes

$$f(\mathbf{x}_{r_i}, \mathbf{x}_{d_j}).$$

By systematically ranking the donors according to their compatibility scores and finding the optimal donor for each patient, the problem solution aims to facilitate more data-driven and reliable allocation decisions. The end goal is to equip transplant teams with a principled method to highlight which donor-recipient pairs stand to achieve the most favorable long-term outcomes.

1.2 Our Contributions

The work introduces a new donor-recipient matching framework for liver transplantation that moves beyond the traditional clinical scoring system. Instead of focusing on urgency-based criteria or limiting the model to fixed, simplistic scoring formulas, we leverage a data-driven, machine-learning approach to predict robustly and rank long-term survival outcomes. Specifically, we apply a survival-augmented variant of XGBoost [13], enriched with a specialized survival loss function, to model the probability of graft survival at multiple clinically relevant time horizons. We then use this reliable method to derive a ranking score that identifies the most suitable donors for a given patient on the waiting list rather than merely producing a predicted survival time. A key aspect of our method is that it integrates a careful feature selection process, ensuring that only informative and clinically meaningful variables influence the final matching decision. By doing so, we minimize the noise and focus on factors that affect post-transplant survival. The ranking score is constructed by integrating survival probabilities across multiple time points, producing a theoretically meaningful representative measure of each candidate’s potential benefits. To validate the significance of our results, we incorporate a hypergeometric test-based statistical evaluation, confirming that the identified top 5% donors, according to our ranking, differ in a non-random, meaningful way from those prioritized by the conventional method (BAR score). We also provide example recipient selection processes for low, medium, and high MELD score patient groups. We evaluate our framework performance with each MELD score group to acknowledge clinicians when and how to utilize our framework in different MELD groups.

Chapter 2

Related Work

Predicting post-transplant outcomes, such as graft survival or patient mortality following liver transplantation, remains a critical challenge in transplantation medicine. Over the past several decades, various methods have been employed, ranging from basic statistical estimators and classical survival models to more sophisticated machine learning and deep learning techniques. This chapter reviews the existing ML methods in the time flow shown in Figure 2.1 as well as some existing liver transplant matching clinical scores. We discuss each method in detail and explain why or why not it is a good choice for our issue, beginning with an introduction to survival analysis, followed by the evolution of traditional models and the emergence of complex ML and DL methodologies.

2.1 Introduction of Survival Analysis

Successful allocations depend on predicting whether a recipient will experience graft failure or death within a critical time window. Traditional regression approaches often fail to handle these outcome variables properly since not all patients reach the event of interest (graft failure in our case) by the end of the study. Hence, a dedicated survival analysis framework is essential. In subsequent sections, we will take a look at several ML models that deploy survival frameworks. Therefore, it is necessary to introduce it first.

Survival analysis focuses on modeling the time until the occurrence of an event [29]. A defining characteristic of survival data is the presence of *censoring*: for some individuals, the event may not have occurred by the end of the observation period, leaving their true event time unknown. In the context of liver transplantation, the event of interest may be

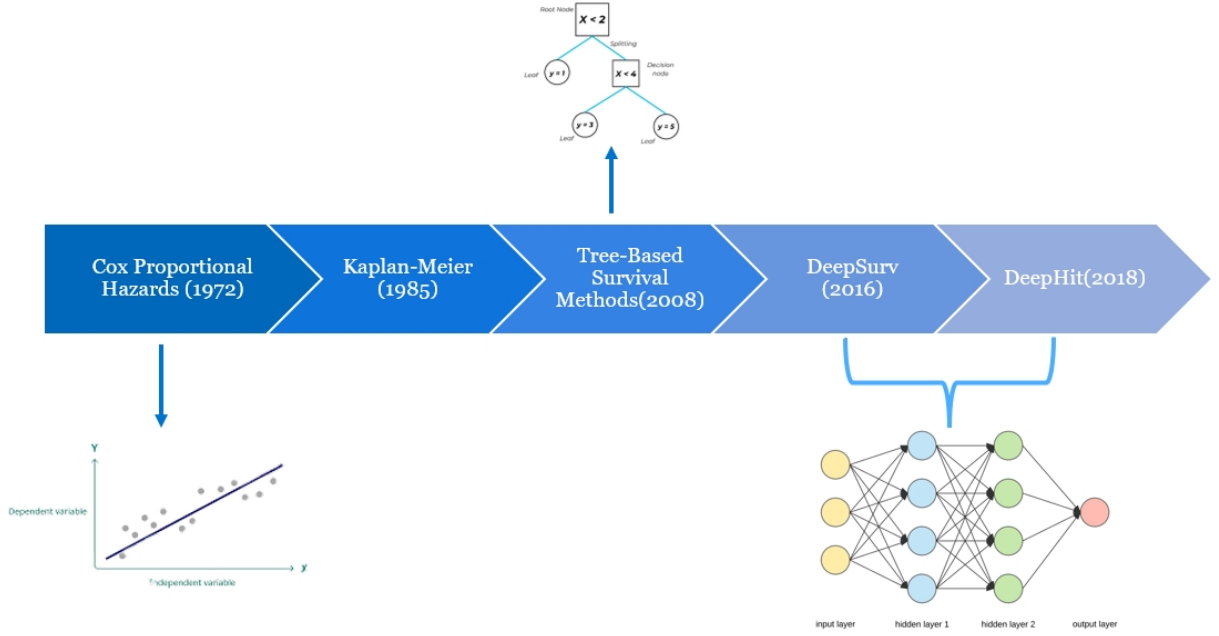


Figure 2.1: Related Works in Survival Analysis

graft failure or patient death, and the censoring time is the moment that clinicians stop recording the recipient's post-transplant status. All the related methods we discuss in the following are based on survival analysis.

Formally, each individual i provides an observed time $\tilde{T}_i = \min(T_i, C_i)$, where C_i is the censoring time, T_i is the event time, and an event indicator δ_i defined as:

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i, \\ 0 & \text{otherwise.} \end{cases}$$

The covariates $X_i \in \mathbb{R}^d$ describe the donor and recipient characteristics. These observations $(\tilde{T}_i, \delta_i, X_i)$ form the basis for learning risk patterns in a survival setting.

Survival Function. The survival function is defined as:

$$S(t) = \mathbb{P}(T > t), \quad (2.1)$$

which gives the probability that the event time T exceeds a given time t . By definition, $S(t)$ is a non-increasing function with $S(0) = 1$ and $\lim_{t \rightarrow \infty} S(t) = 0$.

Hazard Function. The hazard function characterizes the instantaneous risk of the event occurring at time t , given survival up to that time:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \quad (2.2)$$

The hazard function does not represent a probability but rather a rate. It directly relates to the survival function via:

$$S(t) = \exp \left(- \int_0^t h(u) du \right). \quad (2.3)$$

2.2 Kaplan-Meier Estimator

Among the earliest and most widely used tools in survival analysis is the Kaplan-Meier (KM) estimator [48]. The KM estimator provides a non-parametric estimate of the survival function $S(t)$:

$$\hat{S}(t) = \prod_{\tilde{T}_i \leq t} \left(1 - \frac{d_i}{r_i} \right), \quad (2.4)$$

where d_i is the number of individuals at time $\tilde{T}_i$ and r_i is the number of individuals at risk just prior to $\tilde{T}_i$. KM estimator effectively handles censored observations, offering a simple and intuitive way to visualize survival probabilities over time. However, KM curves do not accommodate covariates, making it challenging to produce individualized predictions that account for donor and recipient factors influencing outcomes.

2.3 The Cox Proportional Hazards Model

The Cox Proportional Hazards (CoxPH) model [24] revolutionized survival analysis by introducing a semi-parametric framework that avoids specifying a baseline hazard function. It assumes the hazard function for an individual with covariates X can be expressed as:

$$h(t|X) = h_0(t) \exp(\beta^\top X),$$

where $h_0(t)$ is an unspecified baseline hazard and β is a vector of coefficients to be estimated from data.

Due to its semi-parametric nature, the CoxPH model provides a more flexible approach than strict parametric models. In medical research, CoxPH models have been widely used to assess the effect of individual characteristics, surgical factors, and clinical treatment on the potential event of interest outcomes [36, 43, 36]. Its interpretable coefficients representing log hazard ratios associated with covariates make it attractive to clinicians and policymakers. However, CoxPH relies on the assumption of proportional hazards and may not capture complex non-linearities or interactions in high-dimensional donor-recipient data.

2.4 DeepSurv Model

As the volume and complexity of clinical data increased, researchers began to explore whether more flexible function approximators could outperform traditional models. DeepSurv [33] is one prominent example. It adapts the CoxPH framework by replacing the linear predictor $\beta^\top X$ with a neural network that learns a non-linear representation of covariates:

$$h(t|X) = h_0(t) \exp(f_\theta(X)),$$

where f_θ is a neural network with parameters θ . DeepSurv thus retains the relative risk interpretation of CoxPH while allowing for complex, non-linear feature interactions.

In several clinical contexts, including oncology and cardiovascular disease, DeepSurv has demonstrated improvements in discrimination and calibration over CoxPH [12]. DeepSurv can handle large feature sets and potentially complex variables by applying modern optimization and regularization techniques. However, performance improvements are often incremental and can be sensitive to hyperparameter tuning and data preprocessing steps. Moreover, deep neural networks may be prone to overfitting, particularly if data quality and sample sizes are not sufficient. In addition, the dataset’s missingness and heterogeneity pose significant challenges in the stage of training [33].

2.5 DeepHit Model

While DeepSurv builds upon the Cox model’s risk-based approach, DeepHit [37] takes a different route by modeling the joint distribution of survival times. DeepHit employs a deep neural network to learn a discretized representation of the survival function at a list of time points at competing events (events that cannot happen simultaneously).

Formally, let T be a discrete random variable denoting the time-to-event (with $T \in \{1, 2, \dots, L\}$ for some maximum follow-up time L), and let J be a random variable representing the event type, with $J \in \{1, 2, \dots, K\}$ for K possible event types (including a censored outcome when the event is unobserved). DeepHit aims to estimate:

$$p_{ij}(t) = \Pr(T = t, J = j \mid X_i),$$

for each individual i with covariates X_i . By modeling $p_{ij}(t)$, DeepHit produces a full joint distribution over time and events, allowing for flexible and granular prediction of survival probabilities across a spectrum of competing outcomes.

DeepHit uses a loss function that encourages both discrimination and proper calibration of survival probabilities $p_{ij}(t)$. Empirical studies have shown that DeepHit can outperform traditional survival methods and even DeepSurv in certain settings [37]. Modeling the joint distribution provides a richer representation of outcomes and is particularly useful when competing events are prominent. However, this added complexity translates into increased computational overhead and potentially intricate tuning procedures. Additionally, simply aggregating the prediction result over discrete time points may degrade its predictive power to that of Cox PH [3].

2.6 Tree-based Survival Methods

Unlike previous methods, Tree-based survival methods fit particularly well in our problem setting. These models adapt decision-tree algorithms for time-to-event data. This adjustment allows the algorithm to handle censored observations and to model the time until an event occurs. When many survival trees are aggregated into an ensemble, as in Random Survival Forest (RSF) [27], overall predictive performance can improve substantially, and heterogeneous clinical features can be incorporated more effectively.

In the context of donor-recipient matching for liver transplantation, tree-based approaches confer several advantages. They capture non-linear and interaction effects among donor and recipient attributes without requiring explicit transformations, making them well-suited to the high-dimensional, heterogeneous data commonly encountered in medical registries. These methods also manage missing data by assigning missing records to branches dedicated to incomplete information or by employing surrogate splits [46]. Furthermore, they handle categorical variables without imposing artificial numeric ordinality, a valuable characteristic given that most transplant-related features are distinctly non-ordinal. While large-scale tree ensembles can evolve into black-box models, measures such as variable

importance, enable investigators to identify the main drivers of survival and to provide explanatory insights to clinicians [46]. Overall, tree-based survival methods are greatly suited to our problem needs.

2.6.1 XGBoost Model

We introduce XGBoost (*eXtreme Gradient Boosting*) here as we build our method using it as the base, adapting it to time-to-event data. It was originally designed for regression and classification, yet its efficiency and robust performance on tabular datasets have made it a favorite in many practical scenarios [13]. Compared to Random Forest (RF), the foundation of RSF, XGBoost incrementally refines an ensemble of weak learners by iteratively adding new trees that focus on the residual errors left by previous iterations. In addition, its optimized parallel implementation and efficient data handling often result in faster training times and lower memory usage than RF, which is particularly beneficial for high-dimensional clinical variables in survival analysis.

2.7 Clinical Scoring Systems and Domain-Specific Indices

In parallel with statistical and ML models, numerous clinical scoring systems have been developed to assist in liver allocation and post-transplant prognostication. The Donor Risk Index (DRI) [16] is one early example that summarizes donor quality into a single metric. D-MELD score [23], improves draft compatibility of traditional MELD score by considering donor age. The Survival Benefit model [49] attempts to quantify the net survival gain from transplantation compared to remaining on the waitlist, while the Survival Outcomes Following Transplantation (SOFT) score and the Balance of Risk (BAR) score [45, 14] integrates donor and recipient factors to reflect clinical reality better.

Although these scoring systems are easy to implement and interpret, their static nature and limited feature sets often fail to capture complex, time-varying interactions. Moreover, all precision scores have declining discrimination over time [51]. As a result, they may yield suboptimal stratification of patients, especially in scenarios where patient populations and donor criteria evolve over time. Some newer models attempt to refine these indices, incorporate additional variables, or consider non-linear relationships among donor-recipient pairs [4]. Still, consensus on a definitive, universally applicable clinical scoring system remains elusive.

Chapter 3

Preliminaries

This chapter introduces the necessary evaluation metrics and the donor classification method for our subsequent validations and analyses.

3.1 C-index (Harrell's Concordance Index)

In survival analysis, Harrell's Concordance Index (C-index) [25] is a widely used metric for evaluating a model's discriminative ability. The C-index (C) measures how well the model's predicted risk scores respect the ordering of true event times as the following:

$$C = \frac{\sum_{i \neq j} I(X_i < X_j) I(T_i > T_j) d_j}{\sum_{i \neq j} I(T_i > T_j) d_j}.$$

where I is an indicator function, T_i and T_j are the times to event for patients i and j , X_i and X_j are risk scores for patients i and j , and d_j is the censoring indicator for patient j .

Values of C near 0.5 indicate that the risk score predictions are no better than a coin flip in determining which patient will live longer. Values near 1 indicate that the risk scores are good at determining which of two patients will have the disease first.

3.2 K-Prototypes Clustering

Clustering techniques are used to identify subgroups within a dataset. While k -means clustering is well-suited to numerical data [2], clinical datasets often include both numerical (e.g., recipient age, MELD score) and categorical features (e.g., blood type). The *K-Prototypes* algorithm [28] extends k -means by combining distance measures for both numerical and categorical variables:

$$d((x, y), (x_c, y_c)) = \sum_{x_j \in x} (x_j - x_{c_j})^2 + \sum_{y_j \in y} \mathbb{I}(y_j \neq y_{c_j}),$$

where (x, y) represents an instance with numerical features x and categorical features y , and (x_c, y_c) is a cluster centroid.

3.3 Hypergeometric Test

The hypergeometric test is used to determine whether the observed number of "successes" in a sample drawn from a finite population significantly deviates from what would be expected by random chance [22]. Consider a population of size N , containing K items classified as "successes." We draw a sample of size n without replacement and observe that exactly k of these sampled items are successes. The hypergeometric distribution models the probability of this outcome.

The probability mass function (PMF) of the hypergeometric distribution is given by:

$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}},$$

where:

- N is the total population size,
- K is the total number of successes in the population,
- n is the sample size,
- k is the observed number of successes in the sample.

To assess statistical significance, we compute the probability of observing k or more successes by chance. This corresponds to the upper tail of the hypergeometric distribution:

$$P(X \geq k) = \sum_{i=k}^{\min(K,n)} \frac{\binom{K}{i} \binom{N-K}{n-i}}{\binom{N}{n}}.$$

A small p -value (e.g., $p < 0.05$) obtained from this cumulative probability indicates that the observed number of successes is unlikely to have occurred by random chance, suggesting a statistically significant enrichment of the observed trait or category in the sample.

3.4 Time-Dependent ROC Curves and AUC

Traditional receiver operating characteristic (ROC) curves and their associated Area Under the Curve (AUC) metrics are designed for binary classification tasks, where the outcome is either an event occurrence or a non-occurrence at a fixed point in time. In survival analysis, however, the event of interest may occur at various times, and censored observations complicate the direct application of standard ROC methods. To address these challenges, time-dependent ROC curves and their corresponding time-dependent AUCs have been developed [30].

Time-Dependent ROC Curve For a chosen time $t > 0$, consider the following classification rule: an individual is predicted to fail by time t if their predicted risk is above a certain threshold c . To construct the time-dependent ROC curve at time t , we must define cases and controls with respect to the event occurrence by t , where δ represent where the individual is censored ($\delta = 1$ when uncensored and $\delta = 0$ when censored):

- **Cases:** Individuals who experience the event by time t (i.e., $T \leq t$ and $\delta = 1$).
- **Controls:** Individuals who are known to remain event-free past time t (i.e., $T > t$). This includes both those observed event-free beyond t and those censored after t .

Given a continuous risk score X , we vary the threshold c to obtain pairs of sensitivity and specificity at time t . Specifically:

$$\text{Sensitivity}(t, c) = P(X > c \mid T \leq t, \delta = 1),$$

$$\text{Specificity}(t, c) = P(X \leq c \mid T > t).$$

Plotting $(1 - \text{Specificity}(t, c), \text{Sensitivity}(t, c))$ as c varies produces the time-dependent ROC curve at time t . Just as in the classical ROC framework, a perfect classifier would achieve 100% sensitivity and specificity, and hence its time-dependent ROC curve would pass through the upper-left corner of the unit square.

Time-Dependent AUC The time-dependent AUC is then defined as the probability that, given two individuals, one who fails by time t and one who survives beyond t , the model assigns a higher risk score to the one who fails earlier. Formally:

$$\text{AUC}^{i,j}(t) = P(X_i > X_j \mid T_i \leq t < T_j),$$

where (i, j) is a pair of individuals drawn from the population, where X_i and X_j represent their risk score, such that $T_i \leq t$ and $T_j > t$. Aggregating over all individuals i and j :

$$\text{AUC}(t) = \frac{\sum_{i=1}^n \sum_{j=1}^n I(T_j > t) I(T_i \leq t) \omega_i I(X_j \leq X_i)}{(\sum_{i=1}^n I(T_i > t)) (\sum_{i=1}^n I(T_i \leq t) \omega_i)}$$

where ω_i are inverse probability of censoring weights (IPCW).

Under this definition, $\text{AUC}(t)$ quantifies the model's capacity to correctly rank individuals with respect to their risk of failing before time t . Just as in the classical AUC framework, A time-dependent AUC of 0.5 indicates no better discrimination than random guessing at time t , while a value closer to 1.0 indicates strong discriminative ability.

Chapter 4

Data and Preprocessing

This chapter focuses on the data we use in this study. It introduces the liver transplant data source and explains how we process the study population and its features. Most importantly, it analyzes the challenge imposed by this dataset’s nature, which will be helpful when designing our methodology in the next chapter.

4.1 Data Source (UNOS)

The study utilized data from STAR (Standard Transplant Analysis and Research) files. STAR files are datasets that contain Organ Procurement and Transplantation Network (OPTN) patient-level information about LT procedures from all U.S. transplant hospitals, including both donor and recipient features, collated from (UNOS) United Network for Organ Sharing Data [11]. The author of the thesis has submitted a request and signed the required Data Use Agreement to use the STAR data files for research purposes.

4.2 Study Population

All liver transplants performed from January 2009 to December 2019 were included in the study. This time frame was chosen to contain a transplant recent enough to reflect today’s reality as well as provide a long enough post-transplant period to observe survival outcomes. We included all patients aged at least 18 years who underwent deceased donor LT. However, several transplants were excluded to maintain a focused, homogeneous study population.

We excluded pediatric transplants, living donor LT, and multi-organ transplants. Patients with Status 1 are also excluded as those patients require immediate action and should always be on top of the rank. The full inclusion and exclusion criteria are shown in Supplementary Figure B.1.

4.3 Features Selection

To identify clinically relevant and non-leaking features suitable for the liver transplant context, we initially followed the data cleaning methodology proposed by Berrevoets *et al.* [6]. The step-by-step feature exclusion process is illustrated in Supplementary Figure B.2.

Subsequently, we refined the feature set through a series of statistical assessments. For example, variables reflecting post-transplantation outcomes were excluded to prevent data leakage. Features with more than 20% missingness or those exhibiting zero or near-zero variability were also removed. In addition, expert clinicians reviewed the remaining variables to identify clinically important features and discard irrelevant ones. The additional refinement pipeline for this feature selection process is presented in Supplementary Figure B.3.

4.4 Challenge and Previous Attempt with UNOS Dataset

Several studies have attempted to use the UNOS liver transplant dataset (the one we use) to develop predictive models for post-transplant survival outcomes. Briceño et al. [21] investigated the functionality of Artificial Neural Networks (ANNs) using the United Network for Organ Sharing (UNOS) database. This database included data from 39,189 liver transplants, encompassing donor, recipient, and logistical factors. The study focused on predicting outcomes for the majority class (graft survival) and the minority class (non-survival) at several intervals (3 months, 1, 2, and 5 years). The researchers compared classical statistical approaches (naïve Bayes and logistic regression) with various machine learning techniques: ANNs, random forests, gradient boosting, and support vector machines. At the 5-year mark, traditional logistic regression ($AUC = 0.654$) outperformed machine learning methods such as ANNs ($AUC = 0.599$) and random forests ($AUC = 0.644$). The authors noted that UNOS dataset contains a significant amount of incomplete data, with only 28 variables having less than 10% missingness data.

The poor performance problem that existed in the UNOS dataset further exploded in the evaluation of Cox models and machine learning techniques' prediction power in

survival time after liver transplant by Kantidakis et al [32] . It evaluates the 62294 patients who underwent liver transplant surgery from 2005 to 2015. The average C-index, a rank statistic between all Cox models and Random Survival Forest observations, is around 0.62, suggesting suboptimal model performance. It also mentions that deep learning models suffer from overfitting and instability across training due to the large number of weights. The two studies above indicate that difficulties in accurately predicting post-transplant survival stem from the dataset’s inherent characteristics and the structural limitations of existing models. While addressing this challenge is difficult, it, in the meantime, offers substantial potential for improvement.

Chapter 5

Methodology

In this chapter, we detail the technical framework of our approach to donor-recipient matching, beginning with the motivations and culminating constructions of loss function and scoring system tailored to the feature set refinement process.

5.1 Motivation for Using XGBoost as the Prediction Model

We selected XGBoost for our prediction model for three primary reasons. First, Kantidakis et al. [32] demonstrated that Random Survival Forest outperforms traditional methods like the Cox Proportional Hazards model on similar tasks. This performance advantage is particularly relevant for the UNOS dataset, which contains a substantial amount of missing data. Tree-based models like XGBoost can inherently handle missing values without the need for imputation by treating them as a separate category during the splitting process [13], thus simplifying the data preparation process.

Moreover, since the features in the UNOS dataset are either categorical (for example, Patient Ending Status, ABO, or whether the patient had an exception point) or numerical features on diverse scaling (for example, patient weight, age, creatine level at transplant, XGBoost’s tree-based structure naturally handles those variables without requiring normalization or extensive encoding. This preserves the data’s inherent properties and reduces the preprocessing complexity.

Additionally, the tree-based structure of XGBoost offers a level of interpretability that

deep learning methods often lack, enabling clinicians and researchers to gain insights into the features most influential for survival predictions, increasing the model’s credibility [13].

Overall, these features make XGBoost particularly suitable for predicting post-transplant survival probabilities using the UNOS dataset.

5.2 Motivation for Using Multi-label Survival Probability Prediction and Ranking Score

Traditional methods often produce a direct point estimate of survival time, which is straightforward to implement and convenient for numerical comparison. However, in the UNOS dataset, empirical results in Table 6.1 suggest that these direct estimations yield high error margins. Models designed to output a single survival-time estimate exhibit substantial Root Mean Squared Error (RMSE), indicating difficulty in capturing the precise event time for each individual. These high RMSE values may be attributed to the difficulty of directly predicting exact survival time from complex, high-dimensional data where event times vary widely.

Therefore, instead of relying on a single predicted survival time, a more flexible approach is to predict survival probabilities at a series of discrete time points. Using discrete time points to model post-transplant outcomes on the liver transplant dataset has been shown in prior research [8] to yield higher accuracy and precision. By modeling each time point as a separate label, the model can learn a probability that captures how likely the event is to occur within each interval. This multi-label scheme is generally more straightforward to estimate because it focuses on the probability of surviving past critical milestones rather than pinpointing the exact event time. In addition, it can better align with clinical protocols, which often monitor patients at specific checkpoints such as three months, six months, one year, and beyond. Most importantly, this approach aligns with the primary research question of this thesis: to design a compatibility score of LT transplant based on patients and donors to inform treatment strategies and resource allocation rather than estimate the patients’ post-transplant survival time.

5.3 Discrete Survival Time Point Construction

Our model is specifically trained to predict the survival probability at various clinical-important time points (e.g., 3 months, 6 months, 12 months, 36 months, 60 months, and

72 months) for each patient-donor pair. Such time points are chosen by consulting with clinicians and are able to capture patients' post-transplant survival in a diverse time range. Let these time points be denoted by:

$$0 < t_1 < t_2 < \dots < t_M, \text{ where } M = 6, t_M = 72$$

5.4 Survival XGBoost Loss Function Construction

Below, we illustrate how the losses for both uncensored and censored data are calibrated in our Survival XGBoost Model.

5.4.1 Handling Uncensored Data

For uncensored individuals, we have complete information about their event time. Let $\tilde{T}_i$ be the observed event time for individual i . For each discrete time point t_j , define the true survival status $y_{i,j}$ as:

$$y_{i,j} = \begin{cases} 1 & \text{if } t_j < \tilde{T}_i \text{ (individual } i \text{ is alive after the event),} \\ 0 & \text{if } t_j \geq \tilde{T}_i \text{ (individual } i \text{ has experienced the event at or before } t_j \text{).} \end{cases}$$

The model predicts $p_{i,j}$, the probability of surviving through time t_j for individual i . To encourage the model to predict high survival probabilities before the event and low survival probabilities at or after the event, we employ a cross-entropy based loss over all uncensored individuals and all time points:

$$L_{uncensored} = - \sum_{i \in \mathcal{U}} \sum_{j=1}^M [y_{i,j} \log(p_{i,j}) + (1 - y_{i,j}) \log(1 - p_{i,j})],$$

where $\mathcal{U}$ denotes the set of uncensored individuals.

5.4.2 Handling Censored Data

For censored individuals, we only know that they survived up to their censoring time C_i , but not beyond. For each censored individual i , we set:

$$y_{i,j} = 1 \quad \text{for all } t_j \leq C_i,$$

indicating that the individual is event-free through time C_i . Beyond C_i , we have no direct information, so we do not penalize the model’s predictions for $t_j > C_i$.

Let $t_{m(i)}$ denote the largest discrete time such that $t_{m(i)} \leq C_i$. Then, the partial-entropy loss for censored data is given by:

$$L_{censored} = - \sum_{i \in \mathcal{C}} \sum_{j=1}^{m(i)} \log(p_{i,j}),$$

where $m(i)$ is the index corresponding to the last discrete time point $t_{m(i)} \leq C_i$ for individual i , and $p_{i,j}$ is the predicted survival probability at time t_j .

5.4.3 Combining Loss Functions

Finally, we combine the fully observed (uncensored) and partially observed (censored) losses into a single function. Introducing a hyperparameter $\alpha \in [0, 1]$ allows us to control the relative importance of each loss component. Specifically, we define:

$$L = L_{uncensored} + \alpha \cdot L_{censored}$$

The hyperparameter α can be tuned to mitigate the imbalance inherent in the dataset, particularly when censored instances are more prevalent. By adjusting α , we balance the model’s focus, ensuring that it learns robust survival patterns without overly relying on either fully observed or partially observed data. As the result, the combined loss L fosters more accurate and generalizable survival predictions across the entire study population.

5.5 Ranking Score Construction

To compute a ranking score for each donor-recipient pair, we begin by obtaining estimates of the survival probability using a trained XGBoost survival model:

$$S(t_j) = \hat{p}_j, \quad \text{for } j = 1, 2, \dots, M.$$

Next, we approximate the expected survival time by discretely integrating the survival curve. For each interval $[t_j, t_{j+1}]$, we assume the survival probability is approximately

constant at $S(t_j)$. The incremental contribution of each interval to the expected survival is then:

$$\Delta_j = S(t_j) \cdot (t_{j+1} - t_j).$$

Summing these increments over all intervals yields the ranking score:

$$\text{RankingScore}(D, R) = \sum_{j=1}^{M-1} S(t_j)(t_{j+1} - t_j).$$

This score represents a lower-bound estimation of the expected survival time for the donor-recipient pair, as outlined in the proof provided in Appendix A, which shows that the proposed ranking score is statistically significant. The higher the score is, the more compatible of the donor-recipient pair.

In this study, clinically meaningful time points were identified by expert clinicians as critical intervals for evaluating post-transplant survival: 3, 6, 12, 36, 60, and 72 months, denoting by:

$$\begin{aligned} t_1 &= 3 \text{ months,} \\ t_2 &= 6 \text{ months,} \\ t_3 &= 12 \text{ months,} \\ t_4 &= 36 \text{ months,} \\ t_5 &= 60 \text{ months,} \\ t_6 &= t_M = 72 \text{ months.} \end{aligned}$$

The result of the ranking score calculation remains the same.

5.6 Feature Selection and Model Refinement

A central challenge in leveraging large-scale registry data like UNOS data, such as that provided by the national transplant system, is clinical variables' heterogeneous quality and completeness. Differences in hospital-level data recording practices often lead to high proportions of missing values in national datasets like UNOS, reducing the utility of certain features [21]. Therefore, relying solely on clinical experts' guidance may not yield helpful features. To address this, we employ a two-step process using statistical validation to improve feature selection and rankings.

5.6.1 Initial Feature Selection via Our Model

Our workflow commences with the proposed model trained to differentiate two groups of data instances. Group 1 contains donor-recipient pairs that exhibit both favorable survival outcomes and high predicted ranks. In contrast, Group 2 represents pairs that, despite being ranked highly by the model, have comparatively poor survival outcomes. By examining these groups, we can identify which features the model relies upon when distinguishing strong donor-recipient matches from weaker ones, which our initial model cannot capture.

5.6.2 Statistical Validation of Features

Once we have the two groups from the initial model, we refine the feature set using statistical tests. Continuous variables are examined using two-sample T-tests [34] to assess differences in means between the two groups. For categorical variables, we apply chi-square tests [52] to determine whether feature distributions differ significantly. Together, these two tests produce certain features that carry meaningful survival-related information that differs between long and low-surviving pairs we did poorly in the first model.

5.6.3 Combining Feature Importance and Statistical Significance

The features that pass statistical significance thresholds are then merged with the top 70 most frequently utilized features identified by the initial XGBoost model. This integration ensures that the final feature set is grounded not only in model-driven salience but also in statistical evidence of relevance.

5.6.4 Secondary Model and Compatibility Scoring

With this enhanced feature set, we train a secondary XGBoost survival model optimized to predict survival outcomes and generate a donor-recipient compatibility score. The final model is better equipped to distinguish between favorable and less compatible donor-recipient pairs, thereby supporting clinicians in making more informed decisions during the transplantation process. A flowchart of the entire model refinement process is depicted in Figure 5.1.

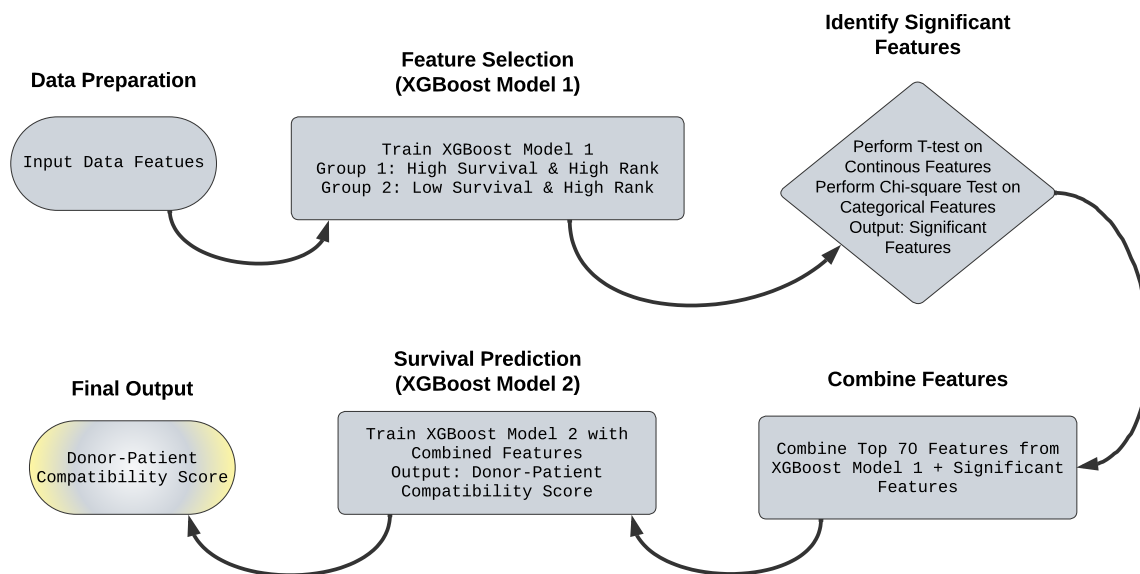


Figure 5.1: Model Refinement Workflow

Chapter 6

Experiment and Evaluation

The dataset was divided into training (80%) and testing (20%) subsets. To ensure both the training and testing sets have a similar distribution of survival times, the train-test split was based on binned graft survival times. The training subset was further split into training (95%) and validation (5%) for tuning hyper-parameters. Notice here that we incorporate a relatively small validation set as the XGBoost model does not require a detailed hyper-parameter tuning discussed in previous sections. The model was trained on the training set, and the testing set was used for testing. To eliminate and minimize the possibility of obtaining results due to randomness, we performed the train-test split using 10 different random seeds and aggregated the performance results.

From now on, we will call our survival variant XGBoost model *DiscTimeSurvXGBoost* to emphasize we are using survival analysis as well as do prediction on discrete time points. Due to the inherent transparency and interpretability of XGBoost, we present the top 20 most frequently selected features in the DiscTimeSurvXGBoost after feature set refinement in Figure 6.1. These features align closely with the clinical variables that practitioners regularly consider in transplant decision-making. As a result, the model’s outputs are not only empirically robust but also clinically credible, strengthening its utility in real-world medical contexts. Additionally, our predicting model and ranking score performance was assessed using Root Mean Squared Error (RMSE) on uncensored data, time-dependent AUC at various time points, and Harrell’s C-statistic [55] for the study period and at the 72-month post-transplant time point (the last discrete time point used in the customized XGBoost model). Our model demonstrated superior performance across all evaluation metrics compared to other benchmarks, including both clinical scores and machine learning models.

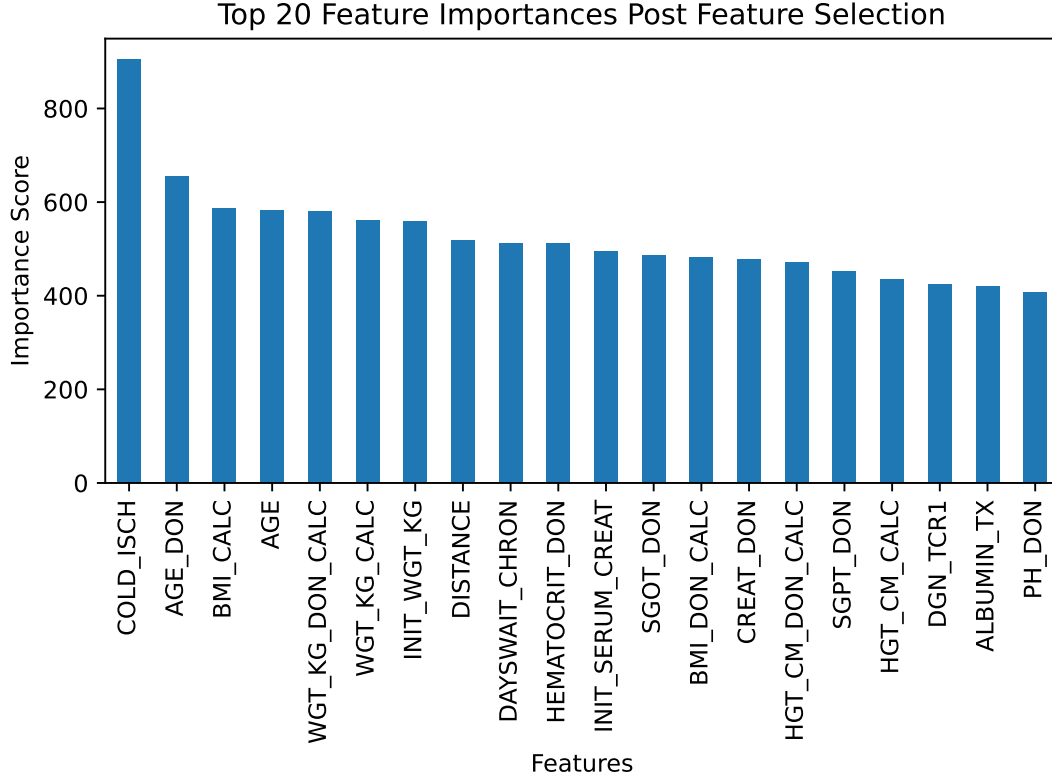


Figure 6.1: Top 20 Most Frequently Used Features

6.1 Evaluating RMSE for Direct Survival Time Estimation

Table 6.1 presents the RMSE in months for various survival models, considering only uncensored data up to 72 months of follow-up. We computed the RMSE by comparing each model’s estimated survival times to the observed event times. For our models, we use our ranking score as the predicted survival time, for DeepSurv and Coxph methods, we choose the time when the survival probability hits 0.5, and for DeepHit, we aggregate the area under probability mass function (pmf) curve as the survival time.

While our DiscTimeSurvXGBoost model achieves a lower RMSE than the competing approaches, the absolute values of these errors remain quite high, indicating that directly predicting an exact survival time remains imprecise. Similarly, DeepHit, despite its con-

Table 6.1: RMSE in Months of the Uncensored Data in Survival Models

Model	Mean	Std	95% CI
DeepHit	76.92	2.88	[74.74, 79.09]
CoxPH	35.52	0.36	[35.25, 35.8]
DeepSurv	34.24	0.30	[34.01, 34.01]
DiscTimeSurvXGBoost (before feature set refinement)	28.90	0.28	[28.68, 29.11]
DiscTimeSurvXGBoost (after feature set refinement)	29.12	0.30	[28.89, 29.35]

ceptual advantages, did not translate into lower RMSE, in part because it aggregates probabilities across multiple time points, introducing additional uncertainty and error accumulation.

The persistently large RMSE values observed across all methods suggest that direct survival time prediction, in absolute terms, is challenging and may not be the most reliable or clinically useful goal. Given these findings, our focus shifts from aiming to predict exact survival times to improving the discriminative capacity of the model. In other words, rather than striving for precise time-to-event estimations, we emphasize the model’s ability to rank donor-recipient pairs according to their relative survival prospects.

6.1.1 Evaluation Through Harrell’s C-statistic

To assess the discriminatory power of our ranking score, we calculated Harrell’s C-statistic [25] for the entire study period (2009 to 2023) and censoring over the 72-month post-transplant time point. The overall C-statistic evaluates how well our model distinguishes between different survival times throughout the study duration. The results of the entire study period are presented in Table 6.2 together with ML and clinical score baselines.

Before applying feature selection, our DiscTimeSurvXGBoost model demonstrated a strong ability to differentiate between shorter and longer donor-recipient pair survival times, achieving a C-index of 0.8601 with a 95% confidence interval (CI) of [0.8590, 0.8610]. This performance surpassed that of the best-performing baseline model, DeepSurv, which had a C-index of 0.8500 (95% CI: [0.8480, 0.8520]). After feature selection, the performance of our DiscTimeSurvXGBoost model improved further, attaining a C-index of 0.8623 (95% CI: [0.8610, 0.8640]). Noticing the clinical scores, DMLD and BAR are at a much lower precision level due to their simple expression.

Table 6.2: C-Index between TX year of 2009-2019

Model	Mean	Std	95% CI
DMELD	0.5174	0.0053	[0.513, 0.521]
Balance of Risk (BAR)	0.5261	0.0035	[0.523, 0.529]
Coxph	0.8435	0.0025	[0.842, 0.845]
DeepSurv	0.8500	0.0038	[0.848, 0.852]
DeepHit	0.8381	0.0040	[0.836, 0.840]
DiscTimeSurvXGBoost (before feature set refinement)	0.8601	0.0023	[0.859, 0.861]
DiscTimeSurvXGBoost (after feature set refinement)	0.8623	0.0026	[0.861, 0.864]

These findings indicate that our model not only outperforms existing benchmarks but also benefits from feature selection, enhancing discrimination in predicting post-transplant survival outcomes.

We also evaluated our model’s discrimination on dataset censoring over 72 months of post-transplant graft survival. Unlike the evaluation of the entire data span, here we focused solely on the discrimination among grafts that survived less than 72 months, censoring those grafts that survived longer than 72 months. The results are shown in Table 6.3. Our DiscTimeSurvXGBoost after the feature set refinement model also led the performance among all models, achieving a C-index of 0.8407 (95% CI: [0.8390, 0.8420]).

These findings highlight the effectiveness of our model in predicting shorter-term survival outcomes, reinforcing its potential utility in clinical decision-making for patients within this critical time frame.

Table 6.3: Performance C-Index between TX year of 2009-2019 on 72-months Survival different models

Model	Mean	Std	95% CI
DMELD	0.5180	0.0056	[0.514, 0.522]
Balance of Risk (BAR)	0.5281	0.0041	[0.525, 0.531]
Coxph	0.8388	0.0037	[0.836, 0.842]
DeepSurv	0.8442	0.0047	[0.841, 0.848]
DeepHit	0.7796	0.0330	[0.755, 0.805]
DiscTimeSurvXGBoost (before feature set refinement)	0.8546	0.0037	[0.8352, 0.857]
DiscTimeSurvXGBoost (after feature set refinement)	0.8559	0.0036	[0.853, 0.859]

Table 6.4: AUC at 1-Year Post-Transplant

Model	Mean AUC	Std	95% CI
Coxph	0.7287	0.0121	[0.723, 0.734]
DeepSurv	0.7418	0.0129	[0.736, 0.748]
DiscTimeSurvXGBoost (before feature set refinement)	0.7641	0.0090	[0.760, 0.768]
DiscTimeSurvXGBoost (after feature set refinement)	0.7687	0.0077	[0.765, 0.772]

6.1.2 Evaluation of Model Performance Using Time-Dependent AUC and ROC

To evaluate and compare the predictive performance of the survival models at specific time points, we calculated the time-dependent Area Under the Receiver Operating Characteristic Curve (AUC) at 1-year, 3-year, and 5-year post-transplant periods. The time-dependent AUC measures the ranking score/survival prediction’s ability to discriminate between patients who experience the event (e.g., graft failure or death) by a certain time t and those who do not. Tables 6.4, 6.5, and 6.6 present the mean, standard deviation, and 95% confidence intervals of time-dependent AUC for each model at 1-year, 3-year, and 5-year post-transplant time points, respectively.

At the 1-year post-transplant time point (Table 6.4), our DiscTimeSurvXGBoost (after feature set refinement) model achieved the highest mean AUC of 0.7687 (95% CI: [0.765, 0.772]), indicating superior discriminatory ability compared to other models. The DiscTimeSurvXGBoost (before feature set refinement) model also performed well with a mean AUC of 0.7641, outperforming both the Coxph and DeepSurv, which had mean AUCs of 0.7287 and 0.7418, respectively.

At the 3-year post-transplant time point (Table 6.5), the DiscTimeSurvXGBoost (after feature set refinement) model demonstrated the best performance with a mean AUC of 0.7942 (95% CI: [0.791, 0.797]). This result surpasses the DiscTimeSurvXGBoost (before feature set refinement) model’s mean AUC of 0.7907 and significantly exceeds the performance of Coxph (mean AUC: 0.7570) and DeepSurv (mean AUC: 0.7697).

At the 5-year post-transplant time point (Table 6.6), the DiscTimeSurvXGBoost (after feature set refinement) model continued to lead with a mean AUC of 0.8065 (95% CI: [0.804, 0.809]), followed closely by the DiscTimeSurvXGBoost (before feature set refinement) model

Table 6.5: AUC at 3-Year Post-Transplant

Model	Mean AUC	Std	95% CI
Coxph	0.7570	0.0092	[0.753, 0.761]
DeepSurv	0.7697	0.0096	[0.765, 0.774]
DiscTimeSurvXGBoost (before feature set refinement)	0.7907	0.0076	[0.787, 0.794]
DiscTimeSurvXGBoost (after feature set refinement)	0.7942	0.0067	[0.791, 0.797]

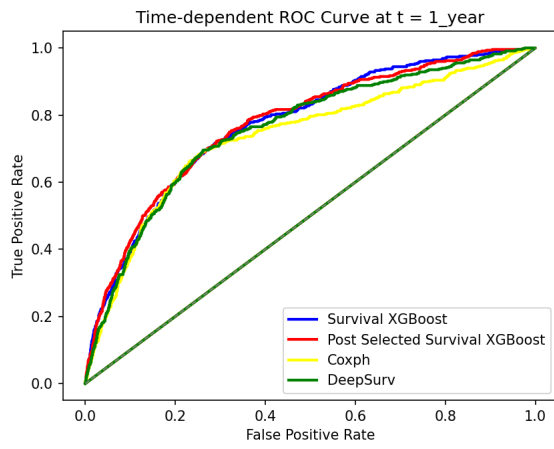
with a mean AUC of 0.8033. Both models show substantial improvement over Coxph and DeepSurv, which achieved mean AUCs of 0.7697 and 0.7828, respectively.

Table 6.6: AUC at 5-Year Post-Transplant

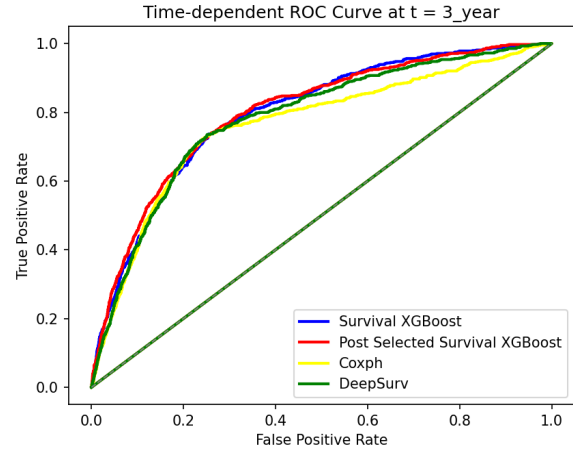
Model	Mean AUC	Std	95% CI
Coxph	0.7697	0.0068	[0.766, 0.773]
DeepSurv	0.7828	0.0073	[0.779, 0.786]
DiscTimeSurvXGBoost (before feature set refinement)	0.8033	0.0060	[0.800, 0.806]
DiscTimeSurvXGBoost (after feature set refinement)	0.8065	0.0053	[0.804, 0.809]

Additionally, we present the receiver operating characteristic (ROC) curves at the specified time points in Figure 6.2. Across all figures and time points, our two proposed models—before and after feature refinement—consistently outperform the baseline models.

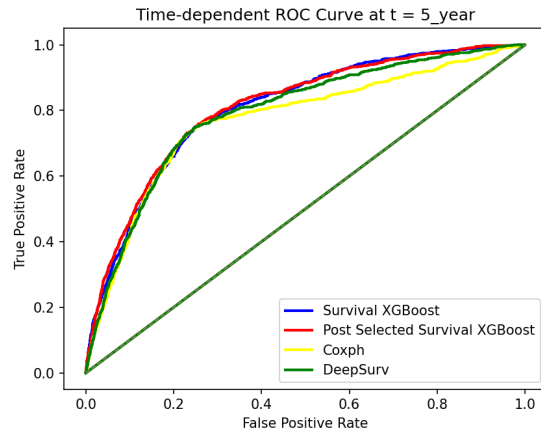
The time-dependent AUC and ROC curve results indicate that our DiscTimeSurvXGBoost (after feature set refinement) model outperforms both the traditional Coxph model and the DeepSurv neural network model. Mathematically, a higher AUC(t) value implies a greater probability that the model assigns a higher risk score to a patient who experiences the event by time t compared to a patient who does not. This indicates better discriminatory power. The consistent improvement in AUC across different time points underscores the robustness of our model in predicting survival outcomes over time.



(a) ROC Curve at 1 Year



(b) ROC Curve at 3 Years



(c) ROC Curve at 5 Years

Figure 6.2: ROC Curves at Some Time Points

6.2 Evaluating Optimal Donors Selection Using Hypergeometric Test

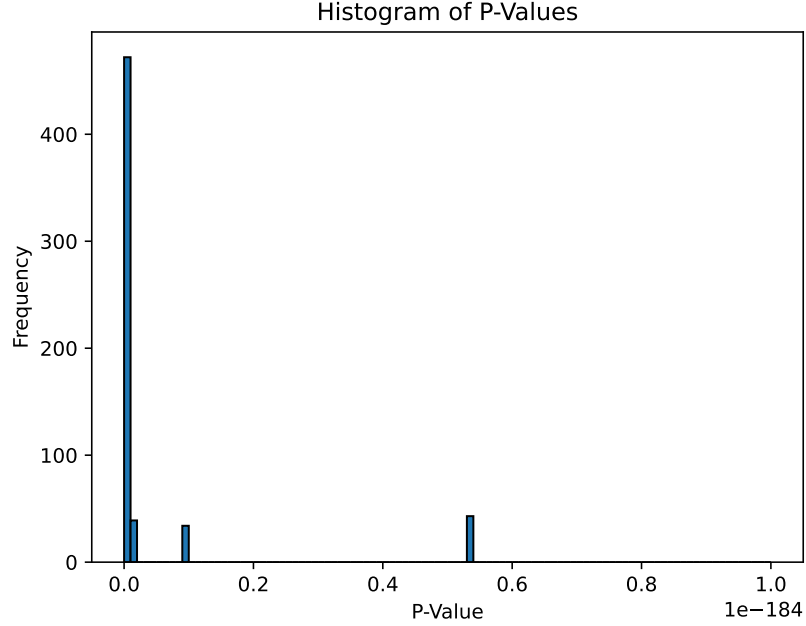


Figure 6.3: P-Values Distribution of Hypergeometric Test

To quantitatively assess whether the top 5% of donors selected by our ranking score differ meaningfully from the top 5% selected by the BAR score, we employ the hypergeometric test [22]. Our underlying hypothesis is that these two top-5% lists are not random samples drawn from the same population. If the observed overlap between the two subsets is considerably smaller than expected under a random sampling assumption, it indicates a meaningful divergence in donor prioritization.

Formally, let N be the total number of donors, and let our method select a subset A (the top 5% by our proposed ranking), and the BAR score selects a subset B (the top 5% by BAR) for a specific patient. We compute the intersection $|A \cap B|$ and use the hypergeometric test to calculate the probability of observing at least this amount of overlap by chance alone.

We apply this procedure for every patient in the test set, comparing our top 5% donor list against the BAR-based 5% donor list. For the entire test set patients, 99.8% of them

have significant p -values on the difference between the two optimal donor lists. Consistently low p -values across all patients confirm that our top 5% selection differs significantly from BAR's top 5%, implying that our ranking criteria identify a fundamentally distinct cohort of suitable donors. Since our approach also achieves a higher C-index, these results suggest that our method better aligns donor characteristics with recipient needs, thereby improving patient selection. Figure 6.3 shows a histogram of the resulting p -values, highlighting the non-random nature of our selected donor sets.

Chapter 7

Impact and Application of the Proposed Framework

Along with validating the utility and accuracy of our proposed framework, we aim to provide clinicians and other potential users with practical insight into how and when the scoring system can be applied. In this chapter, we assess how the model performs across different MELD strata, using concrete examples and empirical evaluation to illustrate scenarios in which the method is best suited for patient evaluation. We also present a case study on constructing optimal donor profiles for a given recipient, enabling clinicians to plan ahead by identifying favorable donor characteristics even before an actual donor organ becomes available.

7.1 Evaluating Model Performance in Different MELD Group

Tables [7.1](#), [7.2](#), and [7.3](#) illustrate examples of the selection process of the best compatible recipient for each donor across three different MELD score strata:

1. **Low MELD Group** (< 20): Recipients in this group have relatively low MELD scores.
2. **Medium MELD Group** ($20 \leq \text{MELD} < 30$): Recipients have moderate MELD scores.

Table 7.1: An Example Illustration of the Recipient Selection Process **Low MELD Group**

	D1	D2	D3	D4
R1	29.36 [4] (51.93)	26.64 [3]	28.28 [4]	26.60 [4]
R2	35.12 [2]	37.42 [1] (33.93)	34.70 [2]	38.16 [2]
R3	21.57 [4]	21.99 [3]	21.87 [4] (14.20)	21.38 [4]
R4	33.03 [4]	34.81 [3]	30.52 [4]	34.00 [4] (24.53)
System Outputs: Selection of Recipient				
	R2	R2	R2	R2

Table 7.2: An Example Illustration of the Recipient Selection Process **Medium MELD Group**

	D1	D2	D3	D4
R1	49.26 [12] (16.80)	49.98 [12]	48.96 [13]	50.61 [12]
R2	30.97 [7]	33.56 [7] (15.90)	31.24 [8]	33.03 [7]
R3	32.54 [7]	33.75 [7]	33.99 [8] (64.17)	37.46 [7]
R4	35.07 [7]	35.85 [7]	35.37 [8]	37.19 [7] (42.07)
System Outputs: Selection of Recipient				
	R1	R1	R1	R1

3. **High MELD Group** (≥ 30): Recipients have high MELD scores, indicating more severe illness and a greater need for transplant.

In each table, the rows (R1, R2, R3, R4) represent recipients and the columns (D1,

Table 7.3: An Example Illustration of the Recipient Selection Process **High MELD Group**

	D1	D2	D3	D4
R1	27.63 [18] 4.00	28.19 [20]	24.83 [19]	28.97 [18]
R2	35.27 [18]	34.16 [20] 45.57	33.38 [19]	36.59 [18]
R3	33.82 [14]	29.99 [16]	28.76 [15] (8.10)	34.16 [14]
R4	26.81 [14]	25.55 [16]	25.99 [15]	26.43 [14] (6.57)
System Outputs: Selection of Recipient				
	R2	R2	R2	R2

Table 7.4: C-Index of Our Ranking between TX year of 2009-2019 Censoring on 72-month Survival on Different MELD Group

Model	Mean	Std	95% CI
Low MELD Group	0.8521	0.0091	[0.845, 0.859]
Medium MELD Group	0.8603	0.0100	[0.853, 0.868]
High MELD Group	0.8724	0.0093	[0.865, 0.879]

D2, D3, D4) represent donors. Recipient and donor with the same number represent that they form a real transplant pair in the UNOS dataset. All such pairs are uncensored (death observed) to give a more informative real survival time. Each cell shows our model’s computed score as well as the corresponding BAR (Balance of Risk) score in red and, when available, the real survival time in blue. After computing the scores for all recipients with respect to a given donor, the system selects the recipient with the highest ranking score as the top-ranked recipient, highlighted in yellow in the final row titled “*System Outputs: Selection of Recipient*”.

From Table 7.1 (Low MELD Group) and Table 7.2 (Medium MELD Group), we observe that our model’s selection does not closely align with what one might consider optimal based on clinical intuition or standard scoring criteria. In other words, the model is not

performing as strongly in these groups. One primary reason for this observation is the limited number of patients in the dataset with low or medium MELD scores. The scarcity of such examples makes it challenging for the model to learn patterns and predict outcomes effectively in these strata.

In contrast, Table 7.3 (High MELD Group) shows that our score ranks better for recipients with high MELD scores, suggesting that it can identify more suitable transplant candidates in the range where patients are sicker, and data is more abundant.

To further quantify and understand these differences in performance, we present a C-index analysis in Table 7.4. The result confirms our hypothesis made from the selection tables: the model achieves better discrimination (higher C-index) in the high MELD group while struggling more with low and medium MELD candidates. Therefore, clinicians should consider our model as a prior method when dealing with sicker patients.

7.2 Compute Average Donor Profile

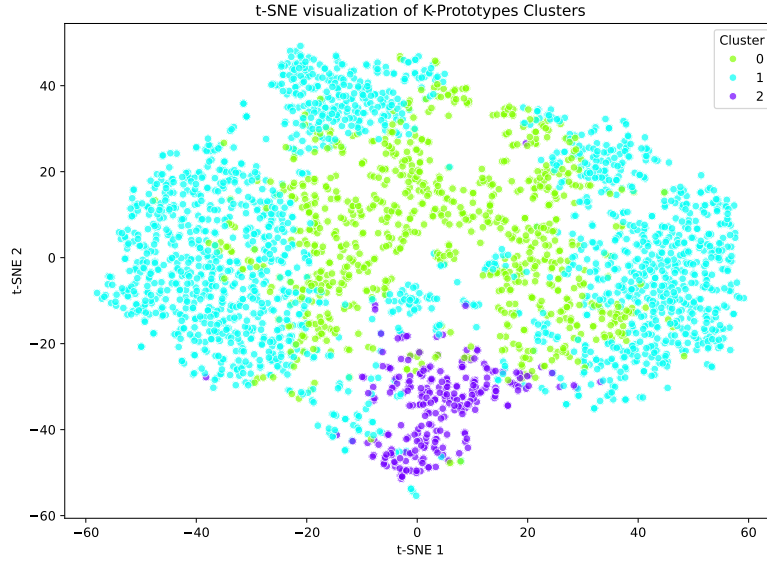


Figure 7.1: t-SNE of Top Donor Profile Clusters

To enhance the clinical utility of our system, we aim not only to provide compatibility scores for existing donor-recipient pairs but also to define an optimal *donor profile* for each individual on the waiting list. By establishing such a reference profile, clinicians can efficiently evaluate incoming donor organs against ideal donor characteristics for a specific patient. As a result, when a new donor organ becomes available, clinicians can rapidly identify which recipients’ optimal donor profile best matches the incoming donor, supporting more informed and timely decision-making.

Specifically, we seek to characterize the top-performing donor types for any randomly sampled recipient based on our ranking model. First, we compute our ranking score between this recipient and every donor across the entire dataset, selecting the top 5% of the highest donor candidates. This subset represents those donors who are most compatible or “ideal” for the recipient, as assessed by our score. Once we have identified these top 5% donors, we apply the *K-Prototypes* clustering algorithm to this subset to group these high-ranked donors into three clusters, representing the best three donor prototypes.

Table 7.5: Cluster 1 Donor Features (Mean and Std)

Feature	Mean	Std
HEMATOCRIT_DON	31.58	6.51
SGOT_DON	70.18	93.76
AGE_DON	28.76	6.91
COLD_ISCH	4.50	1.35
HGT_CM_DON_CALC	181.06	6.96
WGT_KG_DON_CALC	107.00	15.77
BMI_DON_CALC	32.70	4.96
CREAT_DON	1.17	0.50

Table 7.6: Cluster 2 Donor Features (Mean and Std)

Feature	Mean	Std
HEMATOCRIT_DON	29.64	6.59
SGOT_DON	83.26	104.18
AGE_DON	24.61	5.67
COLD_ISCH	4.09	1.31
HGT_CM_DON_CALC	181.66	5.99
WGT_KG_DON_CALC	81.42	8.75
BMI_DON_CALC	24.70	2.67
CREAT_DON	1.09	0.41

7.2.1 Interpreting the Cluster Centers

The result of the K-Prototypes clustering is three clusters of top donors. Each cluster is associated with a cluster center, representing an approximate “average profile” of donors within that cluster. For instance, features such as age, weight, serum creatinine, or BMI can offer insights into the typical clinical status or organ quality of donors in each group. Alongside the mean values, standard deviations reveal how much variability exists for each feature within a cluster. A low standard deviation suggests that most donors in that cluster share similar values for that attribute, whereas a higher standard deviation indicates greater heterogeneity.

Tables 7.5, 7.6, and 7.7 illustrate three optimal donor types for a given recipient, shown in Table 7.8. We also perform a t-SNE (t-distributed stochastic neighbor embedding) [56] analysis in Figure 7.1, providing a visualization of each cluster. Each cluster thus represents a distinct optimal “average” donor type, to drive compatibility based on clinical priorities. For instance, Table 7.7 might represent younger donors with a moderate BMI and lower

Table 7.7: Cluster 3 Donor Features (Mean and Std)

Feature	Mean	Std
HEMATOCRIT_DON	29.56	6.57
SGOT_DON	317.73	520.68
AGE_DON	26.97	5.14
COLD_ISCH	3.88	1.27
HGT_CM_DON_CALC	182.52	5.70
WGT_KG_DON_CALC	94.85	17.39
BMI_DON_CALC	28.54	5.55
CREAT_DON	4.36	1.84

Table 7.8: Sampled Recipient Features

Feature	Mean
INIT_WGT_KG	101.20
DISTANCE	66.00
BMI_CALC	31.70
HGT_CM_CALC	180.34
AGE	46.00
WGT_KG_CALC	103.10
DAYSWAIT_CHRON	7.00

cold ischemia time, whereas Table 7.5 might depict older donors who nonetheless have favorable attributes such as ideal height or manageable serum creatinine levels. When the sampled patient gets paired with any of the three optimal donor types, it will get an estimated 252 months post-transplant draft survival time predicted by our system.

By examining these “average donor profiles” and estimated optimal survival times, transplant teams and researchers can better understand how the model’s highest-ranked donors group into particular phenotypes and expected survival outcomes. This will enable more precise, swift organ allocation for patients on the waiting list.

Chapter 8

Conclusion

In this study, we addressed the challenge of donor-recipient matching in liver transplantation by developing a machine-learning-based framework that predicts post-transplant survival probabilities and produces an interpretable compatibility score. We demonstrate that our method provides more accurate and meaningful patient rankings through various discriminative metrics than traditional approaches. Beyond conventional performance measures, we applied rigorous statistical tests to confirm that the top-ranked donors identified by our model significantly differ from those prioritized by the existing scoring system.

In addition to strong predictive performance, we ensured that the model’s underlying features were both statistically valid and clinically relevant, thereby enhancing transparency, credibility, and eventual clinical utility. Notably, this work is among the first to propose an “optimal donor profile” for each recipient, offering clinicians the ability to allocate patient-specific ideal donors before the actual donor arrives. We also integrated our framework with the MELD score, demonstrating particular utility for high-MELD patients who constitute a large segment of the waitlist population.

Together, these contributions lay the foundation for more efficient and clinically informed organ allocation strategies. This work paves the way for increasingly sophisticated, transparent, and impactful decision-support tools in real clinical applications.

8.1 Future Work

There are several directions that could further enhance this transplant matching framework. First, incorporating a ranking-based loss function into the survival objective may

better capture the direct comparisons needed when ranking multiple candidates.

Second, leveraging longitudinal data for patients on the waitlist could yield more accurate and dynamic predictions. Many clinical indicators evolve over time, and recurrent or transformer-based models that process sequential measurements may reveal patterns missed by a single snapshot.

In addition, our results do not reveal a significant difference among the three types of optimal donor profiles in the examples we tested, both here and in our offline trials. Looking ahead, rather than continuing to generate three donor types, it may be more effective to produce a single donor profile while enhancing how we determine the optimal donor range. Currently, we use a fixed 5% threshold for top donors; future efforts could instead rely on expected survival time to guide the selection threshold more dynamically.

Finally, data limitations and concerns about privacy and fairness in the UNOS dataset point to the value of privacy-preserving synthetic data generation. By releasing synthetic datasets that mirror key statistical properties without exposing protected health information or perpetuating disparities, more researchers can contribute to improving donor-recipient matching without compromising patient confidentiality.

References

- [1] Caterina Accardo, Ivan Vella, Duilio Pagano, Fabrizio di Francesco, Sergio Li Petri, Sergio Calamia, Pasquale Bonsignore, Alessandro Tropea, and Salvatore Gruttadauria. Donor-recipient matching in adult liver transplantation: Current status and advances. *BioScience Trends*, 17(3):203–210, 2023.
- [2] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, 2020.
- [3] Naseem Asghar, Umair Khalil, Basheer Ahmad, Huda M Alshanbari, Muhammad Hamraz, Bakhtiyar Ahmad, and Dost Muhammad Khan. Improved nonparametric survival prediction using coxph, random survival forest & deephit neural network. *BMC Medical Informatics and Decision Making*, 24(1):120, 2024.
- [4] Sumeet K Asrani, Giovanna Saracino, Anji Wall, James F Trotter, Giuliano Testa, Ruben Hernaez, Pratima Sharma, Allison Kwong, Srikanta Banerjee, and Gregory McKenna. Assessment of donor quality and risk of graft failure after liver transplantation: The id2eal score. *American Journal of Transplantation*, 22(12):2921–2930, 2022.
- [5] M. D. Ayllón et al. Validation of artificial neural networks as a methodology for donor-recipient matching for liver transplantation. *Liver Transpl.*, 24:192–203, 2018.
- [6] Jeroen Berrevoets, Ahmed Alaa, Zhaozhi Qian, James Jordon, Alexander ES Gimson, and Mihaela Van Der Schaar. Learning queueing policies for organ transplantation allocation using interpretable counterfactual survival analysis. In *International Conference on Machine Learning*, pages 792–802. PMLR, 2021.
- [7] N. Börner et al. A novel deep learning model as a donor-recipient matching tool to predict survival after liver transplantation. *J. Clin. Med.*, 11:6422, 2022.

- [8] Nikolaus Börner, Markus B Schoenberg, Philipp Pöschke, Christian Heiliger, Sven Jacob, Dominik Koch, Benedikt Pöhlmann, Moritz Drefs, Dionysios Koliogiannis, Christian Böhm, et al. A novel deep learning model as a donor–recipient matching tool to predict survival after liver transplantation. *Journal of Clinical Medicine*, 11(21):6422, 2022.
- [9] J. Briceño et al. Use of artificial intelligence as an innovative donor-recipient matching model for liver transplantation: results from a multicenter spanish study. *J. Hepatol.*, 61:1020–1028, 2014.
- [10] Javier Briceño, María Dolores Ayllón, and Rubén Ciria. Machine-learning algorithms for predicting results in liver transplantation: The problem of donor–recipient matching. *Current Opinion in Organ Transplantation*, 25(4):406–411, 2020.
- [11] J. M. Cecka. The unos scientific renal transplant registry–2000. In *Clinical Transplants*, pages 1–18. 2000.
- [12] Kumardeep Chaudhary, Olivier B Poirion, Liangqun Lu, and Lana X Garmire. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, 24(6):1248–1259, 2018.
- [13] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- [14] P. Dutkowski et al. Are there better guidelines for allocation in liver transplantation? a novel score targeting justice and utility in the model for end-stage liver disease era. *Ann. Surg.*, 254:745–753, 2011.
- [15] B. D. Ershoff et al. Training and validation of deep neural networks for the prediction of 90-day post-liver transplant mortality using unos registry data. *Transplant. Proc.*, 52:246–258, 2020.
- [16] Sandy Feng, Nathan P Goodrich, Jennifer L Bragg-Gresham, Dawn M Dykstra, Jeffery D Punch, MA DebRoy, Stuart M Greenstein, and Robert M Merion. Characteristics associated with liver graft failure: the concept of a donor risk index. *American journal of transplantation*, 6(4):783–790, 2006.
- [17] Alberto Ferrarese, Giuseppe Sartori, Graziella Orrù, Anna Chiara Frigo, Filippo Pelizzaro, Patrizia Burra, and Marco Senzolo. Machine learning in liver transplantation: a tool for some unsolved questions? *Transplant International*, 34(3):398–411, 2021.

- [18] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [19] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore. Understanding survival analysis: Kaplan-meier estimate. *International journal of Ayurveda research*, 1(4):274, 2010.
- [20] N. Goldaracena, J. M. Cullen, D.-S. Kim, B. Ekser, and K. J. Halazun. Expanding the donor pool for liver transplantation with marginal donors. *Int. J. Surg.*, 82S:30–35, 2020.
- [21] David Guijo-Rubio, Javier Briceño, Pedro Antonio Gutiérrez, Maria Dolores Ayllón, Rubén Ciria, and César Hervás-Martínez. Statistical methods versus machine learning techniques for donor-recipient matching in liver transplantation. *PLOS ONE*, 16(5):e0252068, May 2021.
- [22] Florian Hahne, Wolfgang Huber, Robert Gentleman, Seth Falcon, S Falcon, and R Gentleman. Hypergeometric testing used for gene set enrichment analysis. *Bioconductor case studies*, pages 207–220, 2008.
- [23] J. B. Halldorson, R. Bakthavatsalam, O. Fix, J. D. Reyes, and J. D. Perkins. D-meld, a simple predictor of post liver transplant mortality for optimization of donor/recipient matching. *Am. J. Transplant*, 9:318–326, 2009.
- [24] Frank E Harrell, Jr and Frank E Harrell. Cox proportional hazards regression model. *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*, pages 475–519, 2015.
- [25] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.
- [26] D. Q. Huang, P. Mathurin, H. Cortez-Pinto, and R. Loomba. Global epidemiology of alcohol-associated cirrhosis and hcc: trends, projections and risk factors. *Nat. Rev. Gastroenterol. Hepatol.*, 20:37–49, 2023.
- [27] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. 2008.
- [28] Jinchao Ji, Tian Bai, Chunguang Zhou, Chao Ma, and Zhe Wang. An improved k-prototypes clustering algorithm for mixed numeric and categorical data. *Neurocomputing*, 120:590–596, 2013.

- [29] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2002.
- [30] Adina Najwa Kamarudin, Trevor Cox, and Ruwanthi Kolamunnage-Dona. Time-dependent roc curve analysis in medical research: current methods and applications. *BMC medical research methodology*, 17:1–19, 2017.
- [31] Patrick S Kamath and W Ray Kim. The model for end-stage liver disease (meld). *Hepatology*, 45(3):797–805, 2007.
- [32] Georgios Kantidakis, Hein Putter, Carlo Lancia, Jacob de Boer, Andries E Braat, and Marta Fiocco. Survival prediction models since liver transplantation-comparisons between cox models and machine learning techniques. *BMC medical research methodology*, 20:1–14, 2020.
- [33] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18:1–12, 2018.
- [34] Tae Kyun Kim. T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6):540–546, 2015.
- [35] W Ray Kim, Ajitha Mannalithara, Julie K Heimbach, Patrick S Kamath, Sumeet K Asrani, Scott W Biggins, Nicholas L Wood, Sommer E Gentry, and Allison J Kwong. Meld 3.0: the model for end-stage liver disease updated for the modern era. *Gastroenterology*, 161(6):1887–1895, 2021.
- [36] Ilari Kuitunen, Ville T Ponkilainen, Mikko M Uimonen, Antti Eskelinen, and Aleks Reito. Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review. *BMC musculoskeletal disorders*, 22(1):489, 2021.
- [37] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [38] W. H. Lim et al. Natural history of nash cirrhosis in liver transplant waitlist registrants. *J. Hepatol.*, 79:1015–1024, 2023.

- [39] Chien-Liang Liu, Ruey-Shyang Soong, Wei-Chen Lee, Guo-Wei Jiang, and Yun-Chun Lin. Predicting short-term survival after liver transplantation using machine learning. *Scientific reports*, 10(1):5654, 2020.
- [40] J. S. Malamon et al. A model for calculating the long-term estimated post-transplant survival of deceased donor liver transplant patients. *EBioMedicine*, 90:104505, 2023.
- [41] L. E. Matevish, P. A. Vagefi, and M. S. Patel. Trends in current use of machine perfusion for donation after cardiac death donors in the us. *J. Hepatol.*, 81:e187–e188, 2024.
- [42] P. Elliott Miller, Sumeet Pawar, Benjamin Vaccaro, Megan McCullough, Pooja Rao, Rohit Ghosh, Prashant Warier, Nihar R. Desai, and Tariq Ahmad. Predictive abilities of machine learning techniques may be limited by dataset characteristics: Insights from the unos database. *Journal of Cardiac Failure*, 25(6):479–483, May 2019.
- [43] Suresh H Moolgavkar, Ellen T Chang, Heather N Watson, and Edmund C Lau. An assessment of the cox proportional hazards regression model for epidemiologic studies. *Risk Analysis*, 38(4):777–794, 2018.
- [44] W. Ni, J. Shi, and J. Li. Epidemiology, natural history, and management of patients with chb concurrent with masld. *Clin. Liver Dis. (Hoboken)*, 23:e0171, 2024.
- [45] A. Rana et al. Survival outcomes following liver transplantation (soft) score: a novel method to predict patient survival following liver transplantation. *Am. J. Transplant*, 8:2537–2546, 2008.
- [46] Kaushik K Rana et al. A survey on decision tree algorithm for classification. *International journal of Engineering development and research*, 2(1):1–5, 2014.
- [47] Benedikt Reichert¹ABCDEF, Thomas Becker²ADFE, Tobias J Weismüller³ACFE, Moritz Kleine¹ACDE, Lea Zachau¹BCDF, Kai Johanning⁴ADF, Frank Lehner¹AD, Jürgen Klempnauer¹AD, and Harald Schrem¹ABCDEF. Value of the preoperative soft-score, p-soft-score, salt-score and labmeld-score for the prediction of short-term patient and graft survival of high-risk liver transplant recipients with a pre-transplant labmeld-score 30. *Ann Transplant*, 17(2):11–17, 2012.
- [48] Jason T Rich, J Gail Neely, Randal C Paniello, Courtney CJ Voelker, Brian Nussenbaum, and Eric W Wang. A practical guide to understanding kaplan-meier curves. *Otolaryngology—Head and Neck Surgery*, 143(3):331–336, 2010.

- [49] D.E. Schaubel, M.K. Guidinger, S.W. Biggins, J.D. Kalbfleisch, E.A. Pomfret, P. Sharma, and R.M. Merion. Survival benefit-based deceased-donor liver allocation. *American Journal of Transplantation*, 9(4):970–981, Mar 2009.
- [50] M. T. Serrano et al. Mortality and causes of death after liver transplantation: Analysis of sex differences in a large nationwide cohort. *Transpl. Int.*, 35:10263, 2022.
- [51] Lauren Shaffer, Samir Abu-Gazala, Douglas E Schaubel, Peter Abt, and Nadim Mahmud. Performance of risk prediction models for post-liver transplant patient and graft survival over time. *Liver Transplantation*, 30(7):689–698, 2024.
- [52] Ronald J Tallarida, Rodney B Murray, Ronald J Tallarida, and Rodney B Murray. Chi-square test. *Manual of pharmacologic calculations: with computer programs*, pages 140–142, 1987.
- [53] Zeyang Tang, Chengke Zhou, Wei Jiang, Wenjun Zhou, Xiaoping Jing, Jianhui Yu, Babakalli Alkali, and Bojie Sheng. Analysis of significant factors on cable failure using the cox proportional hazard model. *IEEE Transactions on Power Delivery*, 29(2):951–957, 2013.
- [54] Norah A Terrault, Claire Francoz, Marina Berenguer, Michael Charlton, and Julie Heimbach. Liver transplantation 2023: status report, current and future challenges. *Clinical Gastroenterology and Hepatology*, 21(8):2150–2166, 2023.
- [55] Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D’Agostino, and Lee-Jen Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.
- [56] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

APPENDICES

Appendix A

Proof of Ranking Score as Survival Time Lower bound

Since the survival function $\hat{S}(t)$ is non-increasing and satisfies $0 \leq \hat{S}(t) \leq 1$, we can express the expected survival time as:

$$E[T] = \int_0^\infty \hat{S}(t) dt = \int_0^{T_{\max}} \hat{S}(t) dt + \int_{T_{\max}}^\infty \hat{S}(t) dt. \quad (\text{A.1})$$

Since $\hat{S}(t) \geq 0$ for all t , it follows that:

$$E[T] \geq \int_0^{T_{\max}} \hat{S}(t) dt = A. \quad (\text{A.2})$$

Thus, the area A under the predicted survival curve up to $T_{\max}$, the maximum possible survival time, is a lower bound for the expected survival time $E[T]$.

A.0.1 Discrete Approximation

Given that we have predictions at discrete time points, we can approximate A using numerical integration methods, such as the left Riemann sum:

$$A_{\text{approx}} = \sum_{i=1}^{n-1} \hat{S}(t_i) \cdot (t_{i+1} - t_i). \quad (\text{A.3})$$

Since $\hat{S}(t)$ is non-increasing, this approximation underestimates the true area A :

$$A_{\text{approx}} \leq A \leq E[T]. \quad (\text{A.4})$$

Therefore, A_{approx} is also a lower bound for the expected survival time $E[T]$.

Appendix B

Study Samples and Features Selection of UNOS Dataset

In this section, we list all pipelines that are used to preprocess the samples and features of the UNOS dataset.

B.1 Workflow of Preprocessing Study Samples

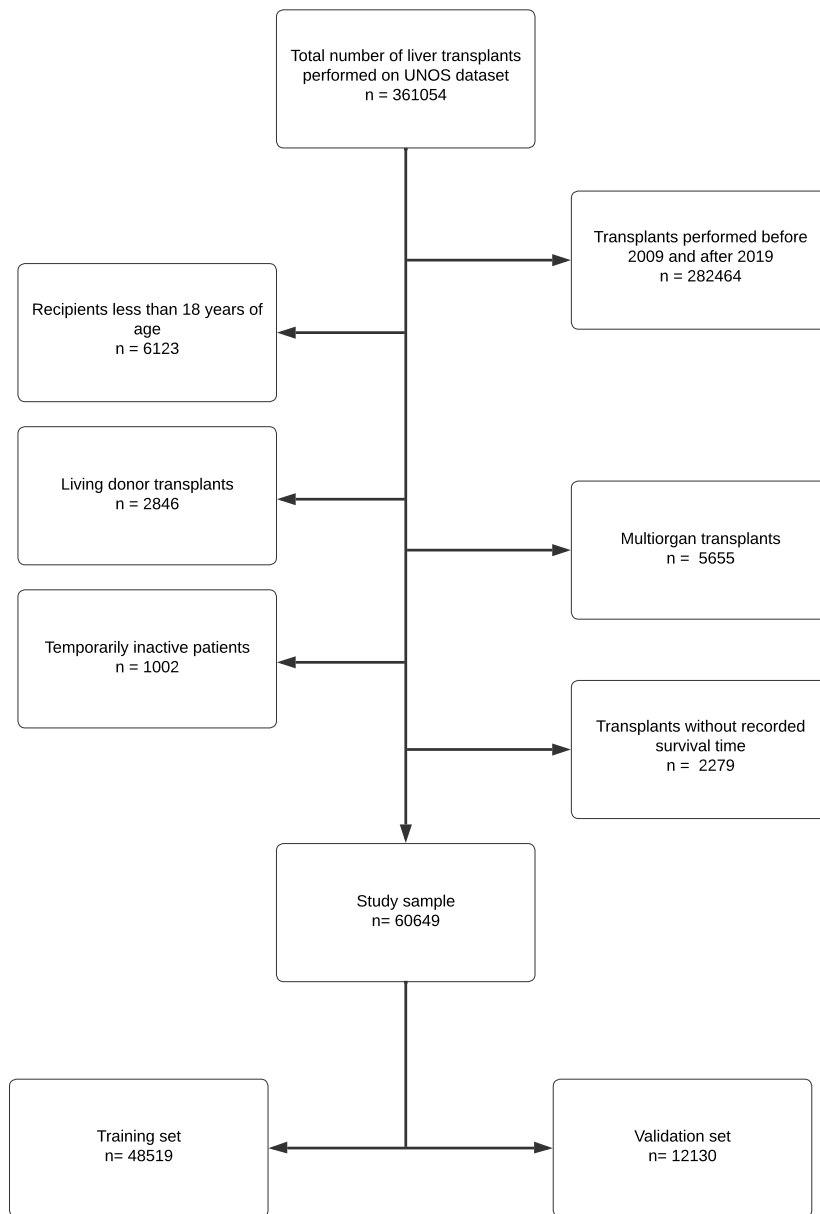


Figure B.1: Flow Chart of Processing Study Samples

B.2 Workflow of Preprocessing Transplant Features

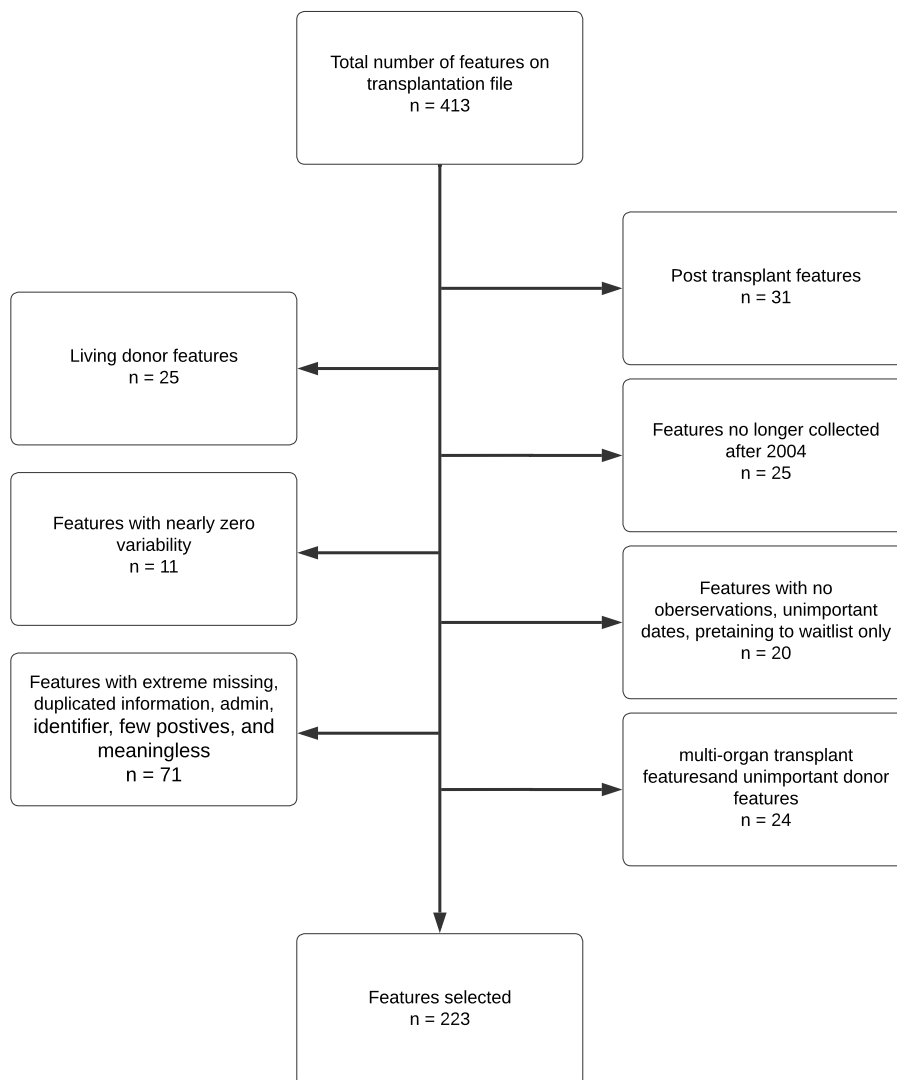


Figure B.2: Flow Chart of Transplant Feature Selection

B.3 Workflow of Deriving Final Study Features

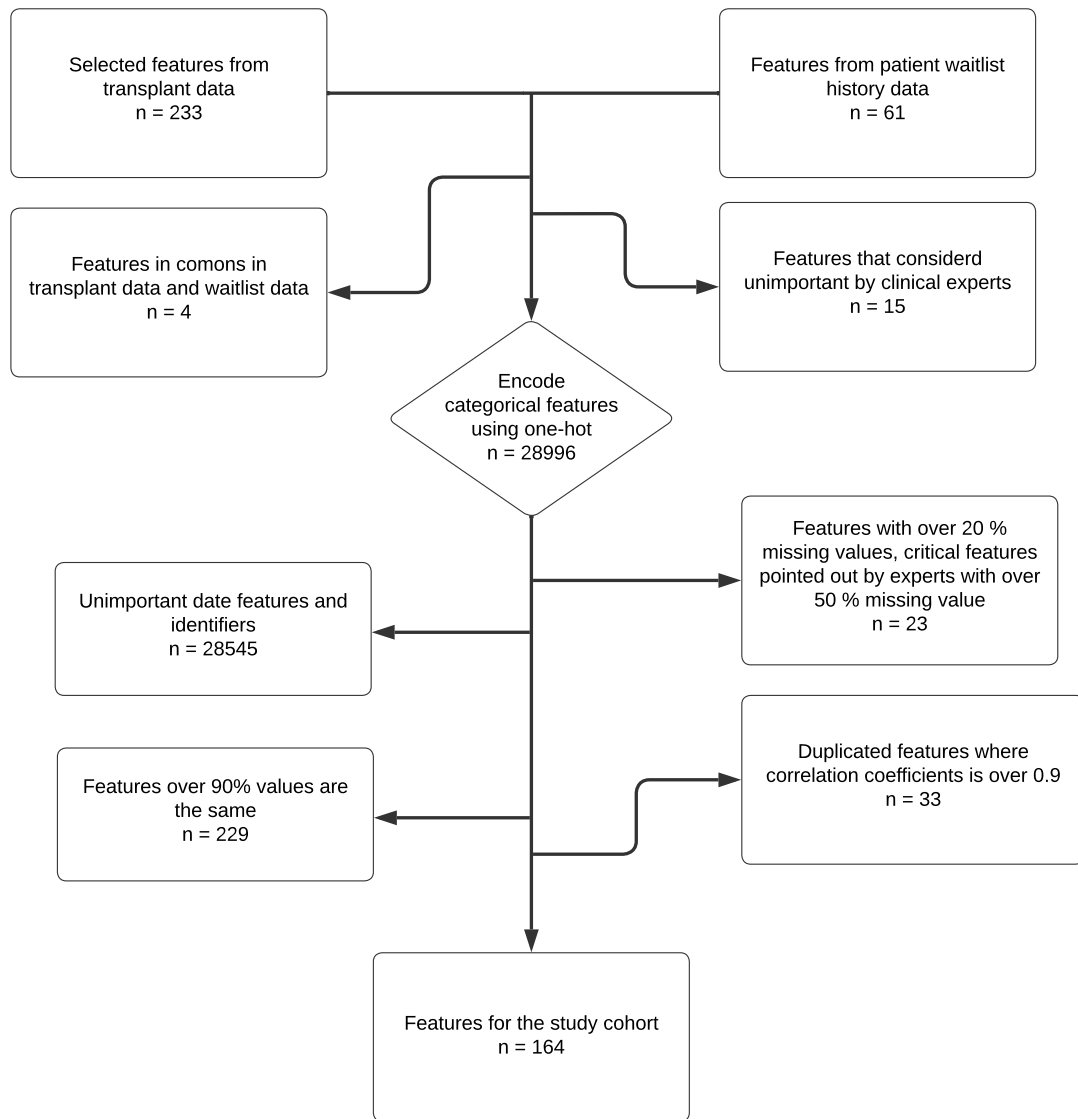


Figure B.3: Flowchart of the feature selection process Used to Derive the Final Study Features