Darmstadt University of Applied Sciences

Department of Mathematics and Computer Science

Availability and Visualization of Predictions and their Uncertainty from Kidney Transplant Survival Models

Submitted in partial fulfilment of the requirements for the degree of

Master of Science (M.Sc.)

Patrick Eidemüller

Matriculation number: 768465

Examiner	:	Prof. Dr. Antje Jahn
Korreferent	:	Prof. Dr. Gunter Grieser

DECLARATION

I hereby declare that I have independently written the present work and have not used any sources other than those listed in the bibliography.

All passages taken verbatim or paraphrased from published or unpublished sources are indicated as such.

The drawings or illustrations in this work have been created by myself or are accompanied by the appropriate source reference.

During the preparation of this thesis, the author has partially used Grammarly to correct grammatical errors and spelling mistakes and to improve sentence structure and the OpenAI gpt-4-turbo API to support the literature research. After using this tool, the author reviewed and edited the content as necessary and assumes full responsibility for the content of the thesis.

In this study, transplant medicine data submitted to the national transplant registry from the years 2006 to 2016 (so-called legacy data) were analyzed. The data were provided in anonymized form by the transplant registry office for research purposes in accordance with § 15g, Paragraph 1 of the Transplantation Act (TPG).

Darmstadt, 27.02.2025

Patrick Eidemüller

ABSTRACT

This study provides a comprehensive analysis of machine learning based survival models in kidney transplantation, identifying methodological challenges. Although significant research efforts continue to introduce new models with enhanced flexibility in modeling complex survival patterns, their applicability to different datasets and research contexts remains a challenge. Leveraging insights from the existing literature to address similar problems significantly fosters broader knowledge transfer.

A critical finding of this study is that, despite the increasing number of survival models being developed, research remains primarily focused on evaluating predictive performance using standard survival metrics, while predictive uncertainty quantification receives little attention. This is particularly concerning in a medical domain where treatment decisions directly impact human lives. Without a clear understanding of predictive uncertainty, clinical decision making risks becoming overly dependent on potentially misleading point estimates.

Furthermore, there is a notable gap in user-centered tools that translate research findings into practical clinical applications. Although survival modeling techniques continue to advance, their integration into clinical workflows remains limited, hindering their potential impact in kidney transplant medicine. Developing accessible and interpretable decision support tools could bridge the gap between machine learning research and real world medical applications.

By evaluating the feasibility of existing prognostic models, assessing their transferability to the German Transplant Registry (TxReg), and applying uncertainty quantification methods such as Monte Carlo Dropout and Bootstrap, this study examines the applicability of ML-based survival analysis methods and assesses their strengths and limitations in the context of kidney transplantation.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

This chapter provides the background, relevance, and objectives of the study.

1.1 BACKGROUND AND RELEVANCE OF THE RESEARCH

In modern medical practice, accurately forecasting transplant outcomes is vital [14, 47], especially for kidney transplants, where the gap between organ demand and supply remains a pressing challenge [38]. Integrating machine learning (ML) promises new possibilities by incorporating a variety of variables and their complex interactions, enabling the discovery of insights that lead to more accurate and personalized predictions [52]. Precise predictions are vital in the field of transplantation, where maximizing the success of each operation is essential to optimize the use of limited organ resources. However, despite its potential, the integration of ML into clinical practice presents significant challenges, particularly in terms of interpretation of the results and quantification of the predictive uncertainties underlying the models while effectively communicating the findings to users [5].

A key element in transplant research is event-time data, which record the time until events such as graft rejection or patient death. This data is often right-censored, meaning that for some patients, the event has not occurred at the time of data collection. This requires specialized statistical approaches capable of handling censored data. In addition, the growing number and complexity of prognostic models raise important questions about their benefit and applicability to different patient populations, such as those represented in national registry such as the German Transplant Registry (TxReg). This leads to the question: can we leverage the knowledge gained from the literature to solve our specific research challenges, which share the same objectives but involve different data?

1.2 OBJECTIVES OF THE STUDY

To address this question, this thesis examines three key aspects.

First, it evaluates the feasibility of applying existing prognostic models to the TxReg data. This includes examining whether these models can theoretically be adapted or transferred to other settings by assessing how well data pre-processing, model architectures, and hyperparameters are reported in the literature. Focus on identifying models, which are completely described and therefore capable of generating individualized predictions. In order to assess the usefulness of the published model when modified for a new but comparable context, this thesis applies and compares a chosen A1 model

on TxReg data. This provides information about the model's relevance and transferability to TxReg data.

Second, this study explores the application of uncertainty quantification, which is especially important when using machine learning models to make predictions in domains where prediction could save lives.

Third, the literature is used to take a closer look at the visualization techniques used to communicate results, predictions and associated uncertainties. Are there dashboards or calculators that could be both useful and easily understood by clinicians and even patients?

1. **Evaluating Prognostic Models for TxReg Data:**
   a) Objective 1: Identify and evaluate prognostic models A, which are capable of generating individualized predictions for new patients. This involves evaluating their applicability by verifying the availability of essential information, including variables, data preprocessing steps, model architecture, and hyperparameters. Models that satisfy these criteria and have all required covariates available within the TxReg dataset will be categorized as Prognostic Models A1.
   b) Objective 2: Investigate the existence and functionality of prognostic calculators (Prognostic Models B). In addition, examine how predictions are reported and evaluates whether prognostic uncertainties are appropriately addressed in these tools.
2. **Applying and Benchmarking a Selected A1 Model on TxReg Data:**
   a) Objective 3: Select a prognostic model from the literature that meets the A1 criteria and apply it to the TxReg dataset. Benchmark the performance of the model on the TxReg data by comparing its predictive accuracy with the original article for the model development.
   b) Objective 4: Quantify the uncertainties associated with the predictions made by the selected model A1 on TxReg data, optimally using the methods identified in the literature.
   c) Objective 5: Visualize the results and predictions motivated by best practices identified in the literature, ensuring that the visualizations are user-friendly.

The results of this study help clarify how the existing literature presents a comprehensive report on methodology and investigates the utility of propagated models.

1.3 STRUCTURE OF THE THESIS

- **Chapter 1: Introduction**
  This chapter provides the background and relevance of the study, outlines its objectives, and presents the structures of the thesis.

- **Chapter 2: Theoretical Background**
  This chapter sets the foundational concepts and theories, it explores key topics including survival analysis principles, machine learning in transplant medicine, and uncertainty quantification methods.
- **Chapter 3: Research Design and Literature Review**
  This chapter outlines the methodological approach to conduct the literature review, focusing on how existing studies report their methodologies to address research questions. It describes the process of identifying relevant articles and categorizing them into groups (A, A1, and B) according to the objectives of the study. In addition, it presents the findings of the literature review, including the reporting of uncertainty quantification methods, visualizations, and approaches such as dashboards and calculators. This chapter provides the groundwork for selecting a suitable A1 model for implementation to TxReg data.
- **Chapter 4: Implementation and Evaluation of Prognostic Model A1 with TxReg data** This chapter focuses on the practical implementation of the study, detailing the process of applying the selected A1 model to TxReg data. It includes a description of the data, variable matching, preprocessing, model training, evaluation, uncertainties, and visualizations.
- **Chapter 5: Results** This chapter presents the results. The findings of the evaluation of the performance of the model, the outcomes of uncertainty quantification and the visualization used. Comparisons are made between the original A1 model and trained on the TxReg data, highlighting differences in predictive accuracy and communication of the results
- **Chapter 6: Discussion** An interpretation of the results in the context of existing literature, discussing the limitations of the study
- **Chapter 7: Conclusions** A summary of key findings, highlighting contributions to existing research, and suggesting potential directions for future studies.

2

THEORETICAL BACKGROUND

This chapter lays the foundational concepts and theories of this research. It presents the mathematical foundation of survival analysis and explores key topics including the principles of event-time data analysis, more profound techniques using deep learning, and the reason for the importance of considering uncertainty when making predictions and the theoretical principles of uncertainty and its quantification.

2.1 INTRODUCTION TO SURVIVAL ANALYSIS

Survival analysis, often referred to as time-to-event analysis, is a set of statistical methods used to model the time until a specific event occurs [29]. This involves analyzing the prognostic factors influencing the timing of this event, these events can vary widely across different fields, this study focuses on the survival time of patients after kidney transplantation.

In the context of this thesis, survival analysis is the foundational framework for modeling and predicting outcomes in data from kidney transplantation event time data. The primary objective is to predict the survival probabilities of patients with transplanted organs over time, providing important information for patient survival.

A unique aspect of survival analysis is its ability to handle censored data, where the exact event time is not observed within the study period. This occurs when the observation time ends before the event occurs or the event is not recorded for other reasons such as patients leaving the study, which leads to incomplete information about the time to the event. Survival analysis enables, by incorporating these incomplete observations, meaningful insights while avoiding the loss of information rooted in the data [31].

The foundational concepts of survival analysis, such as survival function, risk rates, and censorship, will be discussed in detail in Section 2.1.1 to establish a mathematical understanding for survival analysis and the subsequent chapters of the study.

2.1.1 *Censoring and the Fundamental of Survival Analysis*

Time-to-event data form the foundation of survival analysis. It consists of two key elements: the time until the event of interest, denoted as:

$$T^* = \text{time between start time and event time} \tag{2.1}$$

and an indicator variable, δ, which specifies whether the event was observed ($\delta = 1$) or censored ($\delta = 0$) [24]. Censoring occurs when the exact event time T^* is not observed, it can take multiple forms, including right censoring, left censoring, and interval censoring [29]. Since the data analyzed in this work exclusively involves cases of right-censoring, where the event time T^* is greater than the censoring time C the terms censoring and right-censoring will be used synonymously. For example, in clinical studies, censoring occurs when the observation period ends before the event is observed. In survival analysis, we assume that censoring is non informative, which means that the censoring time C is independent of the event time T^*, conditional on the covariates [24].

The endpoint in the survival data consists of the two components T and δ, where:

$$T = \min(T^*, C), \quad \delta = \begin{cases} 1 & \text{if } T^* \leq C, \\ 0 & \text{if } T^* > C. \end{cases}$$

where C is the censoring time and T^* is the event time. Taking into account the covariate vector X, observations can be represented as triplets (X, T, δ). To develop predictive models, classical statistical methods or machine learning approaches must be adapted to handle this specific type of data.

2.1.2 *Important Functions in Survival Analysis*

A central concept in survival analysis is the survival function, which describes the time-to-event distribution. The survival function $S(t)$ is mathematically defined as:

$$S(t) = P(T^* > t) \tag{2.2}$$

It represents the probability that the time until the event of interest T^* has not occurred by time t. The survival function is a non increasing function, satisfying $S(0) = 1$, since all individuals are event-free at the beginning, and $\lim_{t\to\infty} S(t) = 0$, as the event will occur for all individuals at some infinite time point. The survival function is closely related to the cumulative distribution function, which is also called the failure function of T^* through the complementarity relationship:

$$F(t) = P(T^* \leq t) = 1 - S(t), \tag{2.3}$$

where the failure function $F(t)$ represents the probability that the event has occurred in time t.

From this relationship we further derive the probability density function of failure times, $f(t)$, describes the probability density of the event occurring at t, which represents the likelihood per unit time that the event occurs at exactly time t: The density function is defined as:

$$f(t) = \frac{d}{dt}F(t) = -\frac{d}{dt}S(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T^* < t + \Delta t)}{\Delta t}. \tag{2.4}$$

While the density function $f(t)$ describes the rate of events in general, it is often more practical to consider the rate of events for individuals who have not yet experienced the event up to time t. This leads to the concept of the hazard function $h(t)$, which quantifies the instantaneous risk of experiencing the event at a given time t, conditional on having survived up to t [29]. It is mathematically defined as :.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T^* < t + \Delta t \mid T^* > t)}{\Delta t}.$$

Using the definition of conditional probability,

$$P(t \leq T^* < t + \Delta t \mid T^* > t) = \frac{P(t \leq T^* < t + \Delta t)}{P(T^* > t)},$$

The hazard function can be rewritten as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T^* < t + \Delta t)}{\Delta t \cdot P(T^* > t)} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} \tag{2.5}$$

Since the term $P(t \leq T^* < t + \Delta t)$ represents the probability that an event occurs in the interval $[t, t + \Delta t)$, which is approximated by $f(t) \cdot \Delta t$ and the density function can be expressed as a negative derivative of the survival function $f(t) = -\frac{dS(t)}{dt}$. This highlights that the hazard function is the ratio of the instantaneous probability density of events to the probability of survival at time t, describing the dynamics of risk over time [29].

Often it is useful to consider the total accumulated risk over a time interval, since the hazard function describes the instantaneous risk of

an event occurring at time t the cumulative hazard function, $H(t)$, aggregates the hazard over time and is defined as:

$$H(t) = \int_0^t h(u)\,du = \int_0^t -\frac{S'(u)}{S(u)}\,du = -\ln(S(t)) \tag{2.6}$$

By substituting and simplifying the cumulative Hazard function using the properties of logarithms, the survival function can be expressed in terms of the cumulative hazard function as:

$$S(t) = e^{-H(t)} \tag{2.7}$$

This section established the mathematical foundation for survival analysis, which also includes the explanation of key functions $h(t)$, $H(t)$, and their relationships with the survival function $S(t)$ and with each other. These concepts are important for understanding survival analysis and concepts built upon this, such as the Kaplan-Meier estimator, Cox proportional hazards model, as well as deep learning approaches, which estimate survival probabilities differently.

2.1.3 *Key Survival Models*

Non-parametric and semi-parametric approaches are the two primary types of traditional survival models that are examined in the subsequent sections. This work does not specifically address parametric models such as the Weibull regression [53] and non parametric methods such as the Nelson-Aalen estimator [39].

2.1.3.1 *Non-parametric Methods: Kaplan-Meier Estimator*

The Kaplan-Meier estimator, introduced by Kaplan and Meier [26], is a non parametric method used to estimate the survival function $S(t)$ from observed survival times. It does not assume a specific distribution for the event times, which makes it particularly suitable for data sets with censored observations.

The survival function is estimated as:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{2.8}$$

where t represents the time of the distinct events sorted, d_i is the number of events at time t_i, and n_i is the number of individuals at risk just

prior to t_i. At risk is a set of data points where wether an event or censoring does not occur by a specific time point. The Kaplan-Meier function provides a step function that represents the probability of survival over time, remaining constant between successive event times.

The Kaplan–Meier estimator is derived from the concept of the hazard function equation 2.5). At discrete time points t_j, the hazard function is estimated using observed event data. Here, d_j denotes the number of events that occur at time t_j, while n_j represents the number of individuals at risk immediately before t_j. Since the hazard function expresses the instantaneous failure rate, it is defined as the ratio [26]:

$$\hat{h}(t_j) = \frac{d_j}{n_j} \tag{2.9}$$

provides a natural empirical estimate of the hazard, capturing the proportion of failures relative to those still at risk at t_j. This formulation follows directly from the definition of $h(t)$, where we replace the theoretical probability terms with their empirical counterparts in the data set.

The Kaplan-Meier estimator provides the possibility to evaluate the uncertainty of estimates with the estimated survival function's variance to calculate the confidence intervals, which reveal the range of the survival probability. The variance can be estimated using the Greenwood formula [26]:

$$\mathrm{Var}(\hat{S}(t)) = \left(\hat{S}(t)\right)^2 \sum_{t_j \le t} \frac{d_j}{n_j(n_j - d_j)}. \tag{2.10}$$

There are no assumptions regarding the distribution of survival times when using the nonparametric Kaplan-Meier method. The Kaplan-Meier estimator's use for individual-level predictions is limited because it is unable to account for covariates, despite its great use for population-level analysis. Semi-parametric models, such as the Cox proportional hazards model, address this limitation. [11]

2.1.3.2 *Semi-parametric Models: Cox Proportional Hazards Model*

Developed by Cox [11], the Cox proportional hazards model (CoxPH) is a popular semi parametric technique that models the hazard function $h(t)$ as follows:

$$h(t \mid \mathbf{x}) = h_0(t) \exp(\beta^\top \mathbf{x}), \tag{2.11}$$

where $h_0(t)$ is the baseline hazard, which describes the underlying risk of the event occurring at time t while representing the hazard function for an individual in the absence of covariate effects, which means with all covariates set to zero. The covariate are represented by the vector $\mathbf{x}$ and β represents the regression coefficients. The proportional hazards assumption states that the hazard functions for two individuals differ only by a constant proportional factor, independent of time [11]:

$$\frac{h(t \mid \mathbf{x})}{h(t \mid \tilde{\mathbf{x}})} = \exp\left(\beta^\top (\mathbf{x} - \tilde{\mathbf{x}})\right).$$

This property allows the effects of covariates on event times to be estimated without making assumptions about the baseline hazard $h_0(t)$. The estimated regression coefficients β are interpretable through the hazard ratio, which quantifies the multiplicative change in hazard due to a change of one unit in a covariate.

Although the Cox model is highly flexible, it assumes proportional hazards over time and does not estimate $h_0(t)$ directly. Methods such as the Breslow estimator [8] can approximate $h_0(t)$.

2.1.3.3 *Parameter Estimation: The Partial Likelihood in the Cox Model*

The Cox model estimates the parameter vector β using a partial likelihood approach, without explicitly estimating the baseline hazard function $h_0(t)$ first introduced by Cox [11] and later further elaborated by Klein and Moeschberger [29] and Moeschberger. At each event time t_i, the risk set $R(t_i)$ consists of all individuals still under observation just before t_i. When an individual i experiences an event in t_i, their contribution to the partial likelihood represents the relative likelihood that this specific individual, rather than any other in the risk set, experiences the event in t_i. Mathematically, the partial likelihood function is given by:

$$L_{\text{partial}}(\beta) = \prod_{i \in \text{Events}} \frac{\exp(\beta^\top X_i)}{\sum_{j \in R(t_i)} \exp(\beta^\top X_j)}. \tag{2.12}$$

Here, $\beta^\top X_i$ represents the linear predictor for individual i, while the denominator accounts for the summed contribution of all individuals in the risk sett_i. This fraction expresses the conditional probability that the individual i experiences the event at t_i, given that an event occurs within the risk set $R(t_i)$ [31]. The partial likelihood simplifies the computation by canceling $h_0(t)$, which is common across all individuals in $R(t_i)$. This simplifies the estimation.

Taking the logarithm for numerical optimization:

$$\ell(\beta) = \sum_{i \in \text{Events}} \Big(\beta^\top X_i - \ln \big(\sum_{j \in R(t_i)} \exp(\beta^\top X_j) \big) \Big).$$

Numerical methods are used to maximize $\ell(\beta)$, resulting in estimates of β. The partial likelihood effectively incorporates censored observations by considering the risk set $R(t_i)$ at each event time t_i, the model ensures that the censored data contribute meaningfully to the estimation process.

2.1.3.4 *Advantages and Limitations of the Cox Model*

Interpretability is a strength of the Cox model. By estimating hazard ratios $\exp(\beta_j)$, the model allows a straightforward calculation of the effects of each covariate on the hazard [11]. The Cox model is easy to interpret and useful for its applications, since these hazard ratios explain how the event's risk varies directly with the change in the covariates. The Cox model, does not need to explicitly specify the baseline hazard, which makes it applicable even though the underlying hazard is unknown or difficult to estimate directly. [11].

However, the Cox model is not without limitations. The proportional hazards assumption is a key assumption underlying the model, which requires that the ratio of hazards between individuals remains constant over time. This can limit the applicability of the Cox model in real world application, particularly in cases where hazard ratios vary over time or the underlying relationships are non-linear [19, 29].

2.2 DEEP SURVIVAL ANALYSIS: DEEPHIT

Although traditional survival analysis methods, such as the Kaplan-Meier estimator and the Cox proportional hazards model, rely on certain statistical assumptions, machine learning approaches have emerged as more flexible alternatives. Models such as Random Survival Forests [23] and deep learning frameworks, including DeepSurv [27] and DeepHit [32], are capable of capturing complex non linear interactions.

DeepHit is a deep learning-based model, which further extends the capabilities of survival analysis by directly estimating survival probabilities over discrete time intervals extending neural networks to model complex, non linear relationships in survival data[54].

2.2.1 *DeepHit: A Deep Learning Approach for Survival Analysis*

DeepHit, developed by Lee et al. [32], is a deep learning model for survival analysis that estimates the probability mass function (PMF) of event times, conditioned on covariates, denoted as $P(T = t \mid X)$. Unlike traditional models such as Cox Proportional Hazards, which assume proportional hazards, or Kaplan-Meier, which does not use covariates, DeepHit directly predicts event probabilities over discrete time intervals without restrictive assumptions about the underlying hazard function. DeepHit discretize the time horizon into intervals $\{1, \ldots, K\}$, where K is the total number of time intervals. [32] This work focuses on the single risk setting, where only one type of event is considered.

2.2.2 *Model Architecture of DeepHit*

The DeepHit architecture consists of three main components: an input layer, a shared feature extraction network, and an output layer that produces time-specific event probabilities. The input layer processes covariates X and feeds them into a fully connected deep neural network. This shared sub network, denoted as $f_s(X)$, captures complex and nonlinear dependencies across all time points using multiple hidden layers, ReLU activation functions, and dropout regularization [32]. The output of the Softmax activation function is a vector of size K, where each element represents the probability that an event will occur at a specific time k. For single risk settings, the probabilities that the event occurs at time point t for covariates X are calculated as:

$$\hat{P}(T = k \mid X) = \frac{\exp(f_k(X;\theta))}{\sum_{k'=1}^{K} \exp(f_{k'}(X;\theta))}, \tag{2.13}$$

where $f_k(X;\theta)$ is the output of the neural network for time k, parameterized by the weights of the network θ. Softmax normalization ensures that the output probabilities form a valid distribution $\sum_{k=1}^{K} \hat{P}(T = k \mid X) = 1$.

Although DeepHit directly estimates the probability distribution of event times, it does not explicitly model the hazard rate $h(k \mid X)$ or the cumulative hazard function $H(k \mid X)$.

To approximate the hazard rates proposed by DeepHit Lee et al. [32], first derive the survival function $S(k \mid X)$, which is calculated as the cumulative sum of the probabilities of the predicted event for all future time points $k' > k$:

$$\hat{S}(k \mid X) = \sum_{k'=k+1}^{K} \hat{P}(T = k' \mid X). \tag{2.14}$$

and then compute:

$$\hat{h}(k \mid X) = \frac{\hat{P}(T = k \mid X)}{\hat{S}(k-1 \mid X)}. \tag{2.15}$$

However, this estimation can be numerically unstable, especially for small event probabilities $\hat{P}(T = k \mid X)$. Unlike CoxPH, where hazard ratios provide an interpretable measure of relative risk, DeepHit produces the absolute probability of events [32].

2.2.3 *Loss Function of DeepHit*

To effectively train DeepHit, an appropriate loss function is required. The model optimizes a combination of two key loss components: the log-likelihood loss, which maximizes the probability of observed events, and the ranking loss, which ensures the correct temporal ordering of survival probabilities.

2.2.3.1 *Log-Likelihood Loss*

The concept of partial likelihood was first introduced by Cox [11] and is fully explained by Lawless [31] in the context of semiparametric hazard models. Lawless [31] justifies the partial likelihood approach by decomposition of the full likelihood into baseline and regression components, highlighting that the baseline hazard cancels out. Specifically, he emphasizes that the partial likelihood represents the probability that the observed event belongs to a specific individual within the risk set $R(t_i)$.

Lawless [31] further demonstrates how the full likelihood for survival models is decomposed into components for censored and observed events. For an uncensored observation i ($\delta_i = 1$), DeepHit maximizes the probability of the event occurring at time $T_i = k$:

$$\mathcal{L}_{\text{obs},i} = \log \hat{P}(T_i = k \mid X_i) = \log \frac{\exp(f_k(X_i; \theta))}{\sum_{k'=1}^{K} \exp(f_{k'}(X_i; \theta))}. \tag{2.16}$$

For censored data ($\delta_i = 0$), we do not observe the exact time of the event. Instead, we only know that the event occurred after the cen-

soring time k the loss encourages the survival probability beyond the censoring time:

$$\mathcal{L}_{\text{cens},i} = \log S(k_i \mid X_i) = \log \sum_{k'=K_i+1}^{K} \hat{P}(T = k' \mid X_i). \tag{2.17}$$

To simultaneously handle both uncensored ($\delta_i = 1$) and censored ($\delta_i = 0$) data, the total log-likelihood loss is defined as:

$$\mathcal{L}_{\text{likelihood}} = \sum_{i=1}^{n} \left[\delta_i \log \hat{P}(k_i \mid X_i) + (1 - \delta_i) \log S(k_i \mid X_i) \right]. \tag{2.18}$$

For uncensored cases, the model maximizes the likelihood of the observed event time k_i. For censored cases, it maximizes the probability that the event has not yet occurred by time k_i, i.e., $\hat{S}(k_i \mid X_i)$.

Log-likelihood loss maximizes the probability of observed events and appropriately handles incomplete observations while modeling survival probabilities directly [32].

2.2.3.2 *Ranking Loss*

To maintain the correct temporal ordering, DeepHit penalizes violations where an individual experiencing an earlier event ($k_i < k_j$) has a higher predicted cumulative incidence function value compared to another individual experiencing a later event, as proposed by Lee et al. [32]. The ranking loss is inspired by the idea of Harrell's concordance index (2.3.1) and is defined as:

$$\mathcal{L}_{\text{ranking}} = \sum_{k=1}^{K} \alpha_k \sum_{i \neq j} A_{k,i,j} \exp\left(-\frac{\hat{F}_k(k_i^{\text{obs}} \mid \mathbf{x}^{(i)}) - \hat{F}_k(k_i^{\text{obs}} \mid \mathbf{x}^{(j)})}{\sigma} \right). \tag{2.19}$$

where:

- $A_{k,i,j} = 1$ if $k_i^{\text{obs}} < k_j^{\text{obs}}$, and $A_{k,i,j} = 0$ otherwise, marking acceptable event pairs for event k.
- $\hat{F}_k(k^{\text{obs}} \mid \mathbf{x})$ is the predicted cumulative incidence function for event k at the observed event time k^{obs}.
- α_k is the weighting factor for event k.
- σ is a scaling parameter.
- $\exp\left(-\frac{x-y}{\sigma}\right)$ is a convex loss function that penalizes incorrect orderings.

Incorporating this ranking loss into the total loss function encourages the model to predict a correct temporal ordering of pairs. This ensures that individuals with shorter observed survival times are assigned higher predicted risk with lower CIF values compared to those with longer observed survival times.

2.2.3.3 *Combined Loss Function*

The final loss function in DeepHit combines the log-likelihood loss and the ranking loss, balancing them using a hyperparameter λ [32]:

$$\mathcal{L}_{\text{DeepHit}} = -\mathcal{L}_{\text{likelihood}} + \lambda \mathcal{L}_{\text{ranking}}.$$

where $L_{\text{likelihood}}$ maximizes the likelihood of observed event times and survival probabilities for censored data and L_{ranking} ensures the proper pairwise ordering of survival probabilities between individuals based on their observed event times. The hyperparameter λ controls the relative importance of the ranking loss. A higher λ emphasizes the consistency of the pairwise ranking, while a lower λ prioritizes the probability of observed events and survival probabilities [32].

This combined loss formulation enables DeepHit to model survival data effectively while maintaining both individual-level likelihood and population-level ordering consistency.

2.2.3.4 *Advantages and Limitations of the DeepHit Model*

DeepHit provides a highly flexible framework for survival analysis by directly estimating event-time probabilities, rather than relying on hazard ratios or restrictive parametric assumptions [32]. This allows DeepHit to capture complex, time-dependent relationships between covariates and survival probabilities, without assuming proportional hazards, which is a key limitation of traditional models like the Cox model. Unlike the Cox model, which assumes a fixed relationship between covariates and survival risk over time, DeepHit models survival probabilities across discrete time intervals.

Another key advantage of DeepHit is its ability to incorporate censored and uncensored data. By optimizing a log-likelihood loss for observed events and a ranking loss for survival probability ordering, DeepHit ensures that predictions remain temporally consistent and reliable [32].

However, this flexibility comes at a cost: unlike CoxPH, DeepHit does not produce interpretable hazard ratios, making direct comparisons between risk factors more difficult. Additionally, since DeepHit discretize time, its predictions depend on the chosen time resolution, which may introduce biases depending on the discretization.

2.3 EVALUATION METRICS

The evaluation of survival analysis models is a critical step in evaluating their predictive performance and reliability. Various metrics and techniques are used to validate models and their output, while special methods are required for survival analysis to address the challenges inherent in time-to-event data, such as censoring and time-dependent results [19, 21, 50]. For example, these evaluation methods provide insight into the model's ability to discriminate between survival times and predict risks, while handling censored observations effectively.

2.3.1 *Concordance Index (C-Index)*

The Concordance Index (C-Index), introduced by Harrell et al. [20], is a widely used metric in survival analysis to assess the discriminative ability of the model. Measures the model's ability to correctly rank predicted risk scores or survival probabilities relative to the actual event times.

The C-Index is formally defined as:

$$C = P(\hat{\eta}_i > \hat{\eta}_j \mid T_i^* < T_j^*),$$

where $\hat{\eta}_i$ and $\hat{\eta}_j$ are the predicted risk scores for individuals i and j, and T_i^*, T_j^* are their true event times.

To compute the C-Index empirically, we compare all possible pairs of individuals. The model predictions are considered concordant if the predicted risk scores correctly reflect the order of their event times:

$$\hat{\eta}_i > \hat{\eta}_j \iff \hat{S}(t_i) < \hat{S}(t_j), \quad \forall t > 0,$$

Higher risk scores correspond to lower survival probabilities.

The C-Index, as defined by Harrell, is calculated as:

$$\hat{C}_{\text{Harrell}} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I(\hat{\eta}_i > \hat{\eta}_j) I(t_i < t_j) \delta_i}{\sum_{i=1}^{n} \sum_{j=1}^{n} I(t_i < t_j) \delta_i}, \tag{2.20}$$

To account for censoring, Harrell's C index excludes pairs where the order of the events cannot be determined:

- If neither individual is censored, the pair is concordant if $\hat{\eta}_i > \hat{\eta}_j$ and $t_i < t_j$.
- If both individuals are censored, the pair is excluded.

- If one individual is censored at t_i and the other experiences an event at t_j:
  - The pair is excluded if $t_i < t_j$.
  - The pair is concordant if $t_i > t_j$ and $\hat{\eta}_j > \hat{\eta}_i$, and discordant otherwise.

Only concordant and discordant pairs are considered in the calculation:

$$C = \frac{|\text{concordant pairs}|}{|\text{concordant pairs}| + |\text{discordant pairs}|}.$$

Although Harrell's C-Index is widely used, it introduces bias by systematically excluding censored pairs [50]. This exclusion affects the metric, as censored pairs may not be randomly distributed across the data set [19, 50]. In datasets with heavy censoring, the C-Index may overestimate or underestimate the model's discriminative ability.

To address these limitations, different modified versions, such as Antolini's C index, have been developed. These approaches adjust for censoring and incorporate time-dependent predictive accuracy, which is particularly relevant for DeepHit as the model explicitly estimates survival probabilities at different time intervals.

2.3.2 *Antolini's Time-Dependent C-Index*

Antolini's C index extends the standard Concordance Index by incorporating a time-dependent predictive accuracy [2]. It evaluates a model's ability to correctly rank survival probabilities at specific time points, focusing on the temporal ordering of events. The C-Index is defined as:

$$C_{\text{Antolini}} = \frac{\sum_{i,j} I(T_i < T_j)\delta_i I(\hat{S}_j(T_i) > \hat{S}_i(T_i))}{\sum_{i,j} I(T_i < T_j)\delta_i}, \tag{2.21}$$

where $\hat{S}_i(T_i)$ and $\hat{S}_j(T_i)$ denote the predicted survival probabilities at time T_i, where $\hat{S}_i(T_i)$ corresponds to the individual experiencing the event and $\hat{S}_j(T_i)$ represents the survival probability of the paired individual at the same time.

Antolini's C index evaluates pairs of individuals (i, j) where:

- $T_i < T_j$ (individual i experiences the event first),
- $\delta_i = 1$ (the event for i is observed),
- $\hat{S}_j(T_i) > \hat{S}_i(T_i)$, meaning individual i has a higher predicted risk than j, consistent with the earlier observed event.

The numerator counts the number of concordant pairs, where the model correctly assigns a lower survival probability to the individual who experiences the event first (i), ensuring consistency with the order of the observed event. The denominator represents the total number of permissible pairs, where i has an observed event before j.

Antolini's C index accounts for censored data by excluding pairs where the event for i is not observed ($\delta_i = 0$). This exclusion prevents ambiguous comparisons where the true event time of i is unknown, ensuring a robust evaluation of time-dependent discriminative performance [2].

Algorithm 1 Computation of Antolini's Time-Dependent C-Index

Require: Predicted survival probabilities $\hat{S}_i(T_i)$, event times T_i, event indicators δ_i for n individuals
Ensure: Antolini's C-Index C_{Antolini}
Initialize concordant pairs count $C \leftarrow 0$
Initialize comparable pairs count $P \leftarrow 0$
for $i = 1$ to n **do**
  for $j = i + 1$ to n **do**
    if $T_i < T_j$ **and** $\delta_i = 1$ **then** ▷ Check if i has an observed event before j
      $P \leftarrow P + 1$ ▷ Increment comparable pairs count
      Find time index t_i in predicted survival times:

$$t_i \leftarrow \arg\min_t |T_i - \text{times}[t]|$$

      Retrieve survival probabilities at T_i:

$$\hat{S}_i(T_i), \quad \hat{S}_j(T_i)$$

      if $\hat{S}_j(T_i) > \hat{S}_i(T_i)$ **then** ▷ Check if model correctly ranks survival probabilities
        $C \leftarrow C + 1$ ▷ Increment concordant pairs count
      end if
    end if
  end for
end for
  return $C_{\text{Antolini}} = \frac{C}{P}$

2.3.3 *Brier Score*

Although the Concordance Index measures the ability of a model to discriminate between individuals at risk, it does not evaluate the accuracy of predicted survival probabilities [50]. A model can achieve a high C-Index by correctly ranking risks while still producing poorly predicted survival probabilities. The Brier score addresses this limita-

tion by directly measuring the accuracy of survival probability predictions at a given time t. It is defined as:

$$\mathrm{BS}(t) = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{S}_i(t) - I(T_i > t) \right)^2, \tag{2.22}$$

The Brier score can take values between 0 and 1, where smaller values indicate better predictive performance. A score of 0.25 corresponds to predictions that are as good as random guessing, where all individuals are assigned a survival probability of 0.5 at time t. For a good model, the Brier Score should fall below this threshold.

However, in survival analysis, censoring introduces a challenge, as survival status $I(T_i > t)$ can be unknown to censored individuals. To account for censoring, Graf et al. introduced an inverse probability of censoring weighting (IPCW) approach [18]. This method weights each observation based on the probability of remaining uncensored, ensuring that censored individuals contribute meaningfully to the estimation. The IPCW-adjusted Brier score is defined as

$$\mathrm{BS}_{\mathrm{IPCW}}(t) = \frac{1}{N} \sum_{i=1}^{N} w_i(t) \cdot \left(I(T_i > t) - \hat{S}_i(t) \right)^2. \tag{2.23}$$

where $w_i(t)$ represents the inverse probability of remaining uncensored, typically estimated using the Kaplan-Meier estimator of the censoring distribution $G(t)$.

2.3.4 *Integrated Brier Score*

To evaluate the overall accuracy of the prediction throughout the observation period, the Integrated Brier Score (IBS) is used. The IBS is defined as the integral of the Brier score over time:

$$\mathrm{IBS} = \frac{1}{t_{\max}} \int_0^{t_{\max}} \mathrm{BS}(t)\, dt,$$

where $t_{\max}$ is the maximum observed follow-up time in the data set. The IBS provides a single summary measure of the model's predictive accuracy over the full time horizon. The IBS is particularly useful for assessing model performance throughout the follow-up period, as it accounts for both calibration and discrimination.

2.4 MACHINE LEARNING IN TRANSPLANT MEDICINE

Machine learning has the potential to transform the field of transplant medicine by providing advanced tools for predictive analytics for decision support. With the ability to process high-dimensional data and model complex relationships, ML enables new ways to predict critical outcomes such as graft failure, patient mortality, and post-transplant complications [18, 30]. Techniques such as Random Forests, Support Vector Machines and neural networks have been applied to identify patterns in large transplant datasets, outperforming traditional statistical methods in both accuracy and robustness. [23, 27, 32]

The main advantage of ML is the ability to process and capture non linear relationships between clinical, demographic, and other variables that impact transplant outcomes [52]. Models like Random Forests not only predict outcomes, but also provide insight into the relative importance of features [23]. Neural networks, including deep learning architectures, can further leverage large unstructured data sets to obtain meaningful information [27, 32]. These tools can significantly improve transplant decision making processes, allowing for the early identification of high-risk patients and advancing the organ matching process [30].

Despite these advances, integrating ML into clinical workflows in transplant medicine is not without challenges. First, the interpretability of ML models remains a key concern. Many algorithms, especially deep neural networks, function as "black boxes," making it difficult to understand the reasoning behind predictions [44] This lack of interpretability can weaken trust and clinical adoption, particularly in important decisions such as organ transplantation [5]. Second, censoring in survival analysis is an additional difficulty. Survival data sets often contain censored observations, which can bias predictions if not handled appropriately [10].

Another critical challenge is the communication of uncertainty in critical real-world setting ML predictions [42]. In clinical practice, decisions based on predictive models must account for inherent uncertainties. As Kompa, Snoek, and Beam [30] state: "four of the most widely cited medical ML models published since 2016 do not have a mechanism for abstention when uncertain" [30]. The authors argue that the integration of uncertainty quantification techniques to confront the user with uncertainties in predictions fosters the natural human reflex of seeking a second opinion from colleagues when confronted with an unusual clinical case [30]. It is essential to guide clinicians in determining when model output can be trusted or when additional human expertise is needed. Furthermore, Kompa, Snoek, and Beam [30] state: "Medical ML models will be increasingly integrated into clinical practice, and incorporation of predictive uncertainty estimates should become a re-

quired part of this integration. With the ability to say: I don't know, based on predictive uncertainty estimates, that models will be able to alert physicians for a second opinion" [30].

2.5 UNCERTAINTY QUANTIFICATION IN PROGNOSTIC MODELS

One of the main focus of machine learning is the extraction of models from data for the purpose of prediction. The learning task is to build a model that generalizes beyond the data with which it was trained. Machine learning models are approximations of the real-world data distribution, which means that their predictions are inherently uncertain. Gal summarizes this in his work: "In analyzing data or making decisions, it is often necessary to be able to tell whether a model is certain about its output, being able to ask "maybe I need to use more diverse data? or change the model? or perhaps be careful when making a decision?" [16]. Uncertainty quantification (UQ) is a critical component of machine learning, enabling robust predictions and informed decision-making. At its core, UQ addresses the variability in model predictions due to inherent data noise and model limitations. From the steps of the raw information to the prediction, **Gawlikowski** identify five factors contributing to uncertainty in their work:

a) **Variability in Real-World Situations:** This refers to the inherent complexity and variability in the environment from which the data are obtained [17]. In survival analysis, factors such as patient populations are heterogeneous and disease progression varies due to unknown factors. [22].

b) **Error and Noise in Measurement Systems:** Measurement systems may introduce noise or errors, such as imprecise sensor readings. These uncertainties are intrinsic to the data collection process and cannot be reduced by additional training [17]. Medical data often suffer from missing values and imprecise diagnostic tests, which introduces uncertainty in the data [29].

c) **Errors in the Model Structure:** The architecture and design of the models and the parameters used for training can lead to errors. Poorly chosen model structures might fail to capture essential features of the data, contributing to uncertainty [17]. Some survival models, such as the Cox model, have assumptions about the data. If these assumptions are violated, the model can produce biased estimates, introducing the uncertainty of the model [32].

d) **Errors in the Training Procedure:** Training-induced uncertainties arise from factors like initialization randomness, small sample sizes, or insufficient coverage of the training data, high-dimensional covariates, and poor hyperparameter choices. This can lead to models with instability in survival predictions and poor generalization to unseen data [17].

e) **Errors Caused by Unknown Data:** When survival models are applied to new data that are significantly different from their training data (out-of-distribution samples), they might struggle to provide reliable predictions [17].

Although these five factors contribute to uncertainty, they can be categorized into two main types: aleatoric and epistemic uncertainty [37].

2.5.1 *Aleatoric and Epistemic Uncertainty*

Aleatoric uncertainty represents random variability in observations that cannot be eliminated, even with infinite training data. This type of uncertainty is often modeled by incorporating a noise term ϵ [28]:

$$y = f(X) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2(x)), \tag{2.24}$$

where $\sigma^2(x)$ represents the input-dependent noise variance. This uncertainty captures variability in the observations themselves and remains even if the true function $f(x)$ is perfectly known.

Although aleatoric uncertainty is important in modeling data noise, this study focuses on epistemic uncertainty, as it captures model uncertainty and plays a crucial role in decision-making.

2.5.1.1 *Epistemic Uncertainty*

Epistemic uncertainty arises from limited knowledge about the model or its parameters and is often referred to as model uncertainty. Unlike aleatoric uncertainty, epistemic uncertainty is reducible as more data become available. This type of uncertainty quantifies how much individual predictions deviate between different possible model configurations [16].

Let $\hat{f}_\theta(X)$ be the predictions of the model, where θ represents the learned parameters. The variance in predictions across different models defines epistemic uncertainty.

$$\mathrm{Var}[\hat{f}_\theta(X)] = \mathbb{E}_{\theta \sim p(\theta|\mathcal{D})}[(\hat{f}_\theta(X) - \mathbb{E}[\hat{f}_\theta(X)])^2]. \tag{2.25}$$

where the expectation is over the posterior distribution of the model parameters $p(\theta|\mathcal{D})$, which captures the uncertainty in the estimated parameters.

2.5.1.2 *Bayesian Inference for Epistemic Uncertainty*

A fundamental approach to addressing model uncertainty is Bayesian inference, which provides a probabilistic framework to learn from data by treating model parameters as random variables with an associated probability distribution. Unlike frequentest approaches, which assume that the parameters have fixed but unknown values, Bayesian inference represents these parameters using probability distributions that encode the uncertainty about their true values [7].

Let $f(x, \theta)$ be a neural network with parameters θ. Bayesian inference seeks to estimate the posterior distribution over model parameters:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \tag{2.26}$$

where $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ is the training dataset. The predictive distribution for a new input x^* is obtained by marginalizing over θ [7]:

$$p(y^*|x^*, \mathcal{D}) = \int p(y^*|x^*, \theta)p(\theta|\mathcal{D})d\theta. \tag{2.27}$$

However, for deep neural networks, computing this posterior $p(\theta \mid \mathcal{D})$is analytically intractable. Therefore, approximate methods are necessary. One such approach is Monte Carlo Dropout, which leverages stochastic forward passes during inference to approximate the posterior distribution [16].

2.5.2 *Monte Carlo Dropout*

According to Gal and Ghahramani [16], MC Dropout can be interpreted as a variational approximation to Bayesian inference. Gal showed that applying dropout during inference approximates the posterior over neural network weights, making it a practical and scalable method for estimating model uncertainty without explicitly computing the full posterior distribution.

Using the Monte Carlo approximation, the predictive distribution becomes an average over multiple stochastic forward passes with different model configurations obtained through dropout:

$$p(y^* \mid x^*, \mathcal{D}) \approx \frac{1}{T} \sum_{t=1}^{T} p(y^* \mid f(x^*, \theta_t)),$$

where θ_t represents different sampled model parameters from dropout masks during T stochastic forward passes.

When the model outputs direct predictions such as regression values or class probabilities, the predictive distribution simplifies. Instead of integrating over probabilities, we average the outputs from multiple stochastic forward passes [16]:

$$\mathbb{E}[y^*] = \frac{1}{T} \sum_{t=1}^{T} f(x^*, \theta_t),$$

Epistemic uncertainty arises from variability across different parameter samples. Since each forward pass corresponds to a sample from the approximate posterior, the predictive variance measures how spread out the predictions are across these sampled models [16]:

The predictive variance can be estimated as:

$$\mathbb{V}[y^*] = \frac{1}{T} \sum_{t=1}^{T} f(x^*, \theta_t)^2 - \left(\frac{1}{T} \sum_{t=1}^{T} f(x^*, \theta_t) \right)^2.$$

The variance $\mathbb{V}[y^*]$ reflects epistemic uncertainty, which can be reduced by collecting more data. Unlike aleatoric uncertainty, which arises from data noise, epistemic uncertainty stems from limited knowledge of model parameters [16].

2.6 BOOTSTRAPPING FOR UNCERTAINTY QUANTIFICATION

Bootstrapping is a nonparametric resampling technique introduced by Efron [12] and is a fundamental method for quantification of uncertainty in machine learning models. It enables the estimation of the distribution of a parameter or model prediction by repeatedly drawing samples that are replaced with the observed data. The key advantage of bootstrapping is its ability to derive confidence intervals and standard errors without requiring strong assumptions about the underlying data distribution [13].

2.6.1 *Theoretical Foundation of the Bootstrap*

Bootstrapping is a resampling method that is used to estimate the uncertainty of a statistic, model prediction, or performance metric. Instead of deriving confidence intervals analytically, bootstrapping generates multiple datasets from the observed data and computes the variation in the estimated quantities.

Given an observed dataset $x = (x_1, x_2, \ldots, x_n)$, where each x_i is an individual observation, According to Efron [12] the bootstrapping proceeds as follows:

a) **Resampling:** Construct a bootstrap sample $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ by randomly drawing n values with replacement from x, preserving the sample size.

   Since sampling is done with replacement, some observations may appear multiple times, while others may not be selected.

b) **Recomputing the Statistic:** Apply an estimation function $s(x)$ to compute the desired statistic, such as the mean, median, survival function, or a performance metric:

$$s(x^*) = s(x_1^*, x_2^*, \ldots, x_n^*).$$

c) **Repeating the Process:** Generate B bootstrap samples $x_1^*, x_2^*, \ldots, x_B^*$ and compute the corresponding bootstrap replications:

$$s(x_1^*), s(x_2^*), \ldots, s(x_B^*).$$

d) **Estimating Uncertainty:** The bootstrap estimate of the standard error of $s(x)$ is given by:

$$\mathrm{SE}_{\mathrm{boot}}(s) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(s(x_b^*) - \bar{s}^*\right)^2},$$

   where the bootstrap mean is:

$$\bar{s}^* = \frac{1}{B} \sum_{b=1}^{B} s(x_b^*).$$

   As Efron demonstrated, this estimate closely approximates the standard error obtained through classical statistical techniques in many practical cases [12].

Efron [12] emphasize that bootstrap is particularly useful for estimating the variability of statistics such as median, hazard ratio, or concordance index, especially when theoretical standard error formulas are difficult to derive. However, they also highlight that bootstrap estimates can be biased when applied to highly skewed data or small samples.

Recent research further distinguishes the role of bootstrapping in uncertainty quantification. When bootstrapping is used to train multiple models on different bootstrapped datasets, it captures epistemic uncertainty, as the variability across models reflects uncertainty in the learned function due to limited data. In contrast, when bootstrapping

is applied within a fixed model to estimate confidence intervals or standard errors, it primarily quantifies aleatoric uncertainty, as it captures the statistical variability inherent in the data [48]. This dual perspective highlights the flexibility of bootstrapping as a tool for robust uncertainty quantification in machine learning.

3

RESEARCH DESIGN AND LITERATURE REVIEW

This chapter outlines the research questions and describes the methodological approach of the literature review, identifying type A models, focusing on the feasibility assessment of the A1 models, their methods of uncertainty quantification, and the investigation of Model B calculators, including dashboards and user interfaces. By evaluating the current literature, this chapter establishes the foundation for addressing the research questions and the objectives of the study.

3.1 RESEARCH QUESTIONS

This study is guided by objectives, as outlined in Section 1.2, the following research questions aim to explore the practical application of a feasible A1 model and assess the results of prognostic models for kidney transplantation reported in the literature when applied to data from the German Transplant Registry:

- **RQ1:** To what extent can a selected prognostic model (Type A1) accurately predict individual outcomes for new patients based on all relevant covariates when applied to TxReg data? How does the predictive accuracy of the selected model compare with its performance on the original datasets used for its development?
- **RQ2:** How are predictions and uncertainties reported and communicated in a selected prognostic model (Type A1 or B), and can the reporting and communication approaches of this model be meaningfully abstracted and implemented for TxReg?
- **RQ3:** How can methods and concepts from the literature, particularly those related to uncertainty quantification and result visualization through dashboards or user interfaces, be effectively adapted and applied to enhance communication and interpretation of the output of the selected model used on TxReg data?

Research questions arise from the core challenge of assessing whether existing prognostic models can be applied practically to the TxReg registry and whether their predictive performance aligns with the expectation of achieving similar results as the original source when transferred to these new registry data. Furthermore, the study investigates whether uncertainty quantification methods from the literature can be effectively implemented when applied to TxReg data. In addition, it evaluates whether visualization methods, such as dashboards or user

interfaces, can efficiently communicate the results of the applied models and enhance the usability of the literature. The literature review is motivated by these questions, while they address the study's aim of exploring the feasibility and utility of the reported approaches to demonstrate their practical application to TxReg data.

3.2 Methodology: Systematic Review in Survival Analysis

This chapter outlines the systematic approach used to review 50 articles on machine learning for survival analysis with time-to-event data, with a primary focus on kidney transplant medicine. To ensure a comprehensive and structured review, a detailed criteria table was constructed which can be found in the Appendix 6.1. This table represents the defined criteria to extract the information from the articles, including details about the variables, data preprocessing methods, information about the model including the architecture and hyperparameters and evaluation procedures, as well as aspects of uncertainty quantification and visual representation of the predictions or the presence of dashboards or user-interfaces.

The review process used a hybrid methodology that combined manual review and the assistance of a large language model (LLM), specifically the gpt-4-turbo model [41]. The methodology was divided into three distinct steps, as outlined below.

3.2.1 *Text Processing and Hybrid Review*

The first step involved analyzing the pre-selected articles through preprocessing with GROBID (GeneRation Of Bibliographic Data) [15] to extract structured textual data from the PDFs and reducing the words in the text. This processed text was then analyzed using the criteria table to extract information according to the defined criteria, leveraging a large language model combined with human-based research.

a) **PDF Processing with GROBID:** PDF's were processed through the GROBID API. The full text of each paper was extracted and parsed to isolate the body text, removing irrelevant sections like headers and footers.

b) **LLM-Based Text Analysis:** The preprocessed text was analyzed using the gpt-4-turbo model. The model was configured and the message 6.4 was set with detailed search criteria for data sources, variables, preprocessing methods, models, uncertainty quantification, and visualizations 6.1, additional instructions to ensure consistency 6.2, and a JSON response format 6.3 for structured responses.

c) **Hybrid Review Process:** While the large language model generated responses for numerous criteria, all results were manually cross-checked against the articles to ensure accuracy. The revised review served as the basis for subsequent steps.

3.2.2 *Aligning Variables with TxReg Data*

A critical challenge in the review process was mapping the variables used in the reviewed articles to those of the TxReg. Inconsistencies in variable naming across articles and the TxReg database as well as data context variability, leading to missing variables, such as the presence of "race" in international studies but its absence in German datasets and the divergent meanings and categories for similarly named variables, contributed to a challenging process. The matching process involved the following steps:

a) **LLM-Assisted Matching:** A list of TxReg variables was provided to the LLM to streamline the matching. The matching process employed the following methodology:

- Each paper's variables were compared against the TxReg variables using gpt-4-turbo initialized with the message in the Appendix 6.5.
- The function incorporated rules for direct matching and adjusted matches such as BMI calculations from weight and height.

b) **Hybrid Matching Process:** While the LLM provided an initial matching of the variables, the results need to be manually reviewed to minimize errors caused by semantic differences, knowledge gaps and lack of context.

The following figure 3.1 illustrates the systematic review process used in the study. It outlines the sequence of tasks involved in processing the pre-selected articles through a workflow that integrates tools such as GROBID and LLM-based text analysis, followed by manual interventions to ensure accuracy and consistency, ultimately resulting in the final data set of reviewed articles.

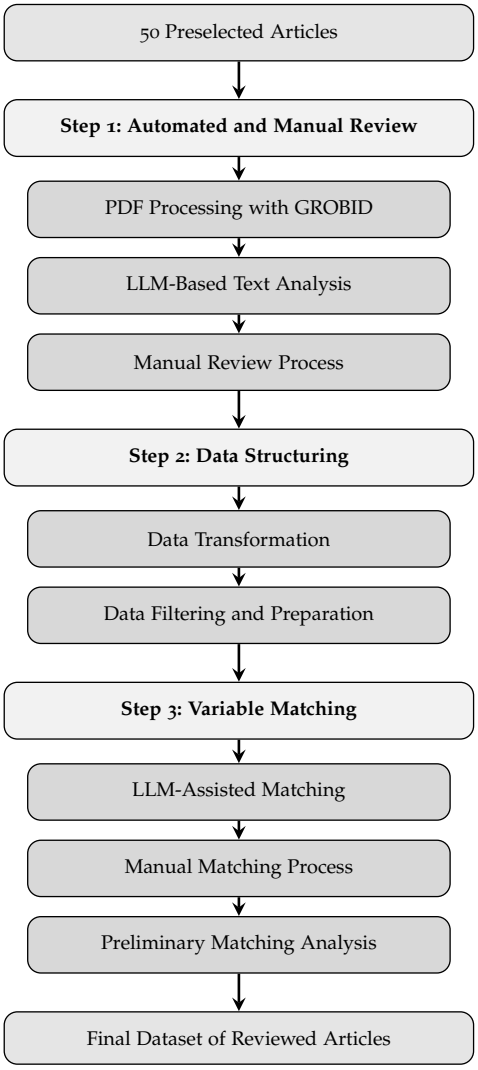

Figure 3.1: Systematic Literature Review process

3.3 FINDINGS FROM THE LITERATURE REVIEW

The systematic literature review initially assessed a total of 50 articles. Among these articles, only three included a dashboard or a graphical user interface as part of their findings. In addition, 6 articles featured dashboard that could be used to visualize predictions for individual patients. In particular, 17 papers mentioned uncertainty-related keywords, indicating some level of consideration for uncertainty quantification within their methodologies. Only 10 papers could be classified as potentially applicable for model reconstruction. Reproducible articles mean that they were characterized by a detailed description of their data, model architecture, hyperparameters, and preprocessing methods. This level of detail is critical for ensuring reproducibility and practical applicability.

Figure 3.2 highlights the systematic filtering process, categorizing the initial 50 articles based on uncertainty, calculators, dashboards and reproducibility.

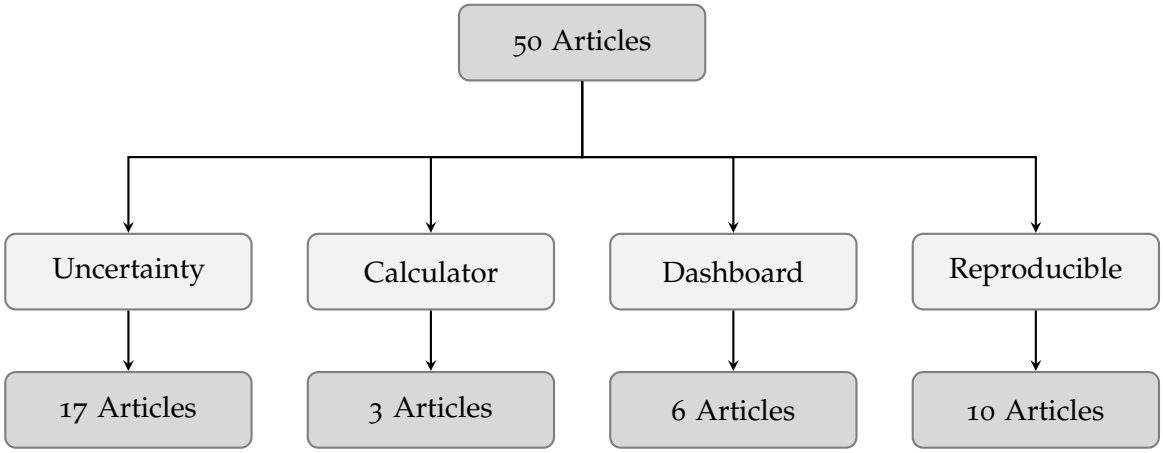

Figure 3.2: Filtering Process of the Literature Review

Based on this, Figure 3.3 provides a detailed visualization of the evaluation results for various criteria in the reviewed articles. This stacked bar chart highlights the distribution of reported criteria, providing insight into how frequently specific aspects were addressed in the literature.

Categories such as variables, data cleaning, metrics, and detailed description of data or visual representation of results have a significantly high proportion of the information provided. This indicates that these aspects are frequently addressed in the reviewed articles. Several categories, such as imputation, scaling, data transformation, and variable selection, which describe data preprocessing steps, show a more balanced distribution of available details. This implies that the reporting of these aspects is inconsistently addressed in the reviewed articles. Categories such as detailed descriptions of model architecture and hyperparameter settings contain only limited information in the articles, demonstrating that model architecture is rarely addressed in the literature. This lack of detail implies significant challenges for implementing most of the models discussed in the reviewed articles. Categories such as dashboard implementations and calculator links have very few reported entries, suggesting that these elements are underrepresented in the literature.

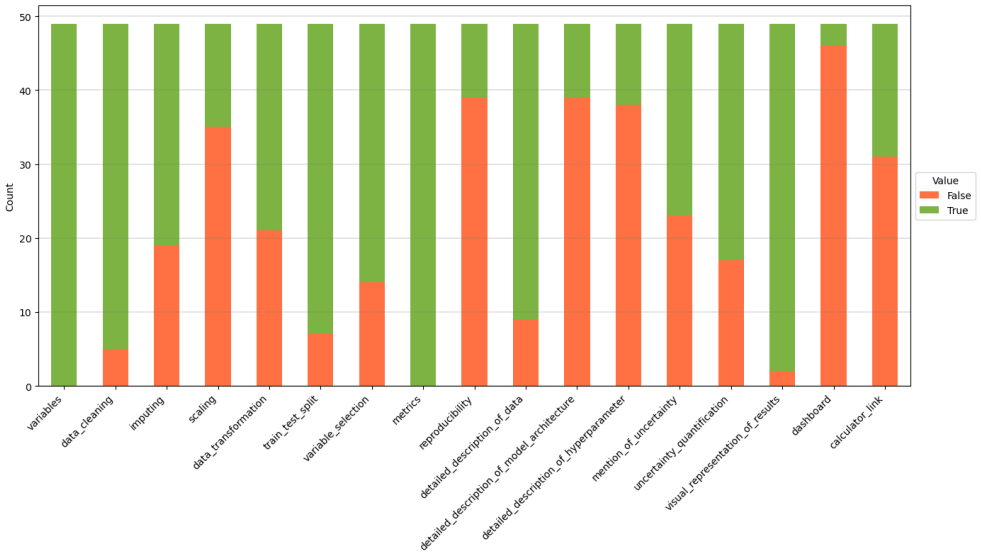

Figure 3.3: Fulfillment of Search Criteria

3.3.1 *Feasibility of Prognostic Models for TxReg*

In this section, the findings of the literature review are presented regarding the feasibility of applying A1 models to the TxReg data set. The analysis identifies models (referred to as Prognostic Models A) that are capable of generating individualized predictions for new patients based on all relevant covariates available in the TxReg data set. The review evaluates whether the articles provide a clearly stated methodology, including comprehensive details on the variables used, preprocessing methods, model architecture, and hyperparameters. Furthermore, the review examines which of these models have matching variables with the TxReg dataset, allowing us to identify potential A1 models.

Furthermore, the findings highlight the existence and functionality of prognostic calculators (Prognostic Models B), with an analysis of how these tools report predictions and uncertainties. Particular attention is paid to whether these models incorporate dashboards or user interfaces to communicate results effectively.

The process involves:

- **Identifying Prognostic Models A:** Reviewing existing models to determine which are capable of producing personalized predictions.
- **Assessing Prognostic Calculators Models B:** Investigating the availability of prognostic calculators and evaluating how predictions are reported and whether uncertainties are adequately addressed.

- **Feasibility Analysis of potential A1 models:** Determining which Prognostic Models A contain all the necessary covariates captured in the TxReg data and evaluating their applicability (prognostic models A1).

IDENTIFYING PROGNOSTIC MODELS A Analyzing objective 1 in Section 1.2, the systematic literature review identified several prognostic models classified as type A, which are capable of generating personalized predictions based on covariates. To assess their applicability, we analyzed the methodological aspects reported in each study, focusing on key factors such as preprocessing techniques, model architecture, and hyperparameters. The table 3.1 presents an overview of type A models. The following criteria were evaluated to determine whether a model is reproducible and qualifies as Type A:

- **Variables:** Indicates whether all relevant covariates required for individualized predictions are specified and adequately documented.
- **Data Cleaning:** Assesses whether data cleaning procedures were applied and clearly reported in the study.
- **Imputing:** Evaluates whether missing data imputation methods were applied and adequately described.
- **Scaling:** Determines whether data scaling or normalization techniques were used and explicitly documented.
- **Transformation:** Indicates whether any transformations were applied to the data and reported properly.
- **Train-Test Split:** Verifies whether the data set was appropriately split into training and test subsets for model evaluation and whether this was clearly reported.
- **Model Architecture:** Evaluates whether the study provides a comprehensive description of the model's structure
- **Hyperparameter:** Assesses whether all hyperparameters used in the model are specified.

FEASIBILITY ANALYSIS MODELS A1 While various models were identified in the literature, not all of them can be directly applied to the TxReg dataset. The feasibility analysis assesses whether the identified Type A models meet the criteria for variable compatibility. Furthermore, these articles were analyzed for their incorporation of uncertainty quantification and the availability of dashboards or user interfaces.

Table 3.1 summarizes the reproducible models and the number of missing variables in each study.

Paper ID	Reference	Missing Variables	Reproducible	Uncertainty	Dashboard
34414609	Ayers et al. [3]	22	Yes	Yes	No
32419922	Senanayake et al. [47]	not available	Yes	Yes	No
34448704	Naqvi et al. [38]	7	Yes	Yes	No
35700006	Paquette et al. [43]	3	Yes	Yes	No
34822363	Thongprayoon et al. [49]	3	Yes	No	No
30625130	Mark et al. [35]	5	Yes	Yes	No
36938431	Linse et al. [33]	55	Yes	Yes	No
31926745	Ershoff et al. [14]	80	Yes	Yes	No
36388342	Roller et al. [46]	5	Yes	Yes	Yes
33198650	Kantidakis et al. [25]	6	Yes	Yes	No

Table 3.1: Potential A1 Models in the Literature Review

Table 3.1 demonstrates that the majority of the reviewed articles exhibit a considerable number of missing variables. Notably, there is no single article in which all variables align perfectly with the TxReg data.

However, four articles with the ID 34448704 by Naqvi et al. [38], 34822363 by Thongprayoon et al. [49], 30625130 by Mark et al. [35], and 35700006 by Paquette et al. [43] stand out as particularly noteworthy. Although none of these articles achieve a perfect match with the registry, they present significant potential for further investigation. These studies could serve as a valuable foundation for adapting and applying their methodologies to the TxReg.

The paper *34448704* provides a detailed description of the variables used in the UNOS registry, including a table detailing the data types and descriptions for transparency. In summary, the paper offers a robust methodology with a clear description of variables, preprocessing, and hyperparameters, making it suitable for reproducibility and application on similar datasets. Logistic Regression, Random Forest, Support Vector Machines, Artificial Neural Networks, and AdaBoost were evaluated for binary classification across temporal cohorts (0-1, 1-5, and 5-17 years). Feature importance and AUC score plots were included, but uncertainty quantification was not discussed or applied. Although the paper is generally well constructed, it lacks the use of specialized metrics such as the C-index or Brier score and the authors do not specify how censored patients were handled, suggesting that censoring may have been ignored [38].

The paper *34822363* focuses specifically on black kidney transplant recipients with data from the UNOS registry. Furthermore, there is no focus on predictions or uncertainty quantification; instead the authors' research focus is variable importance for black kidney recipients [49].

The paper *30625130* presents a strong methodology and comprehensible results. The data source is the UNOS registry, all variables are detailed in the appendix, including type, description, and categories. The

preprocessing steps are described. The most relevant model parameters and hyperparameters are described, though some are excluded. The model used is a combination of Random Survival Forests for cohort 1 (age under 50) and a Cox Proportional Hazard model for cohort 2 (age over 50). The metrics used include the C index and the Brier score. Permutation importance plots and survival plots with comparisons to Kaplan-Meier estimates were provided. However, there is no focus on uncertainty quantification or alternative visualizations of the results [35].

The paper *35700006* investigates five models for predicting survival after kidney transplantation, namely Cox proportional hazards, random survival forests, DeepSurv, DeepHit, and a Recurrent Neural Network. Among these, DeepHit was identified as the best-performing model. The probability of graft survival was predicted at fixed time points ranging from 0 to 15 years post transplantation, with 3-month intervals between each time point [43].

The study is distinguished by its well-documented methodology, particularly data sources, preprocessing, and feature selection. The data set used in this study originates from the Scientific Registry of Transplant Recipients, with the variables detailed in the Appendix of the articles. Data preparation is described in detail, and feature selection was based on a combination of expert knowledge, data completeness, and findings from previously published studies. The model evaluation was conducted using the concordance index, the Brier score, and calibration metrics, including the Integrated Calibration Index score and calibration plots. Furthermore, the paper is notable for the development of a dashboard, designed to facilitate the practical application of the model and enhance clinical decision-making through an interactive user interface.

ASSESSING PROGNOSTIC CALCULATORS MODELS B Examining objective 2 1.2 in the literature review reveals that beyond standard prognostic models, few studies have also introduced prognostic calculators aimed at providing individualized risk predictions through formulas, dashboards, or interfaces. These calculators vary in their accessibility; some are available as online tools, others are not available. Table 3.2 presents an overview of the prognostic calculators identified in the reviewed literature with a link and a short description. The accessible user interfaces or dashboards are provided in the Appendix 6.1 6.2 6.4 6.3.

Paper ID	Reference	Details	Remarks
29543895	Andres et al. [1]	http://webdocs.cs.ualberta.ca/~uhlich/Liver_2002/html/pssp_2002.html	Online tool available
29483521	Medved et al. [36]	http://ihtsa.cs.lth.se	Offline
28921876	Ayllón et al. [4]	Formula-based	Derived results based on mathematical formulas
		Mentioned interface (DISPO)	Interface discussed, but not included in the paper
35389371	Zafar et al. [55]	https://lungscore.research.cchmc.org/96b53228-f7b6-4cb0-bbdf-59bf733d7056	Online calculator available
35700006	Paquette et al. [43]	On request, no answer	Authors requested for tool but no response yet
33858815	Nitski et al. [40]	Working on calculator	Example Dashboard Plot
36388342	Roller et al. [46]	No Online tool	Example Dashboard Plot

Table 3.2: Prognostic Calculators in the Literature Review

The Patient-Specific Survival Prediction Tool for Liver Transplant Survival 29543895 [1] describes a tool that uses a machine learning algorithm to generate individualized survival curves based on recipient data. The model demonstrated excellent calibration, outperforming traditional Cox models, particularly in long-term predictions. The online calculator provides survival predictions through a user-friendly interface and is accessible.

The International Heart Transplantation Survival Algorithm calculator in article 29483521 [36] describes deep learning based techniques to predict survival after heart transplantation. This tool integrates recipient and donor variables, using artificial neural networks for survival estimates for one, five, and ten years. The web based calculator is no longer available online and could not be evaluated as part of this review.

Article 28921876 [4] provides a formula derived from an artificial neural network model. This model predicts 3-month and 1-year graft survival probabilities based on multiple variables with assigned weights. Variables include pretransplant status, MELD score at transplantation, days on waiting list, liver disease etiology, donor's cause of death, and cold ischemia time, among others. The formula offers personalized risk predictions by combining these variables into logistic regression-based probability functions.

The Lung Transplantation Advanced Prediction Tool in article 35389371 [55] presents a machine learning approach, specifically Cox-Lasso regression, to predict survival outcomes for lung transplant recipients. This web based calculator estimates 1-, 5-, and 10-year survival probabilities and assigns risk scores based on recipient, donor, and transplant factors. The interactive tool is designed for clinical use, providing survival times and risk classification to assist in recipient-donor matching.

The prognostic calculator of paper 35700006 [43] is described as a prototype tool to predict the survival of kidney transplants. It is based on a recurrent neural network model trained on the data of the Scientific Registry of Transplant recipients. The tool estimates graft survival probabilities for 1 to 15 years post transplant and provides comparative survival rates using different donor and recipient variables. The authors mention that the code and model can be made available upon request, but no response was received when requested for this review.

Article 33858815 [40] is a liver transplant survival prediction tool that applies deep learning algorithms to patient data. The tool is currently in development, with the authors working on a web-based interface for clinical use.

The Clinical Decision Support System for kidney transplantation described in article 36388342 [46] is designed to detect patients at risk of rejection and death-censored graft failure within 90 days. This system is not a fully operational prognostic calculator yet, but rather a machine learning model integrated into a dashboard for clinical decision-making.

3.3.2 *Uncertainty Quantification Methods in Literature*

The reviewed studies reveal various approaches to uncertainty quantification, with differences in the methods used and their reporting as summarized in 3.3. Primarily through confidence interval estimation. Bootstrap confidence intervals were the most frequently employed method across multiple studies, particularly for survival function estimates and model validation. In several cases, bootstrapping was performed without further methodological details or visual representation of the uncertainty.

Another commonly used approach was the calculation of confidence intervals for hazard ratios, which in some studies was based on the standard error (SE) of regression parameters. However, not all papers specified whether this approximation was used or provided further details on the computation method.

Kaplan-Meier confidence intervals, cross-validation-based confidence intervals were also reported in some studies, though their derivation was not always described. Additionally, cross-validation techniques, such as 5-fold or 10-fold cross-validation, were employed in several studies to estimate variability in performance metrics like the concordance index (C-Index) or AUROC but were not always explicitly linked to uncertainty quantification in model predictions.

Overall, the analysis of the literature indicates that while confidence interval estimation primarily via bootstrapping was widely used, explicit discussions of uncertainty quantification in prognostic predictions remained limited. The methodologies applied for uncertainty estimation were often not the primary focus of the studies but were rather used as standard statistical techniques for performance evaluation, often without a specific focus on uncertainty quantification itself.

Paper ID	Uncertainty/CI Method	Additional Features
MLinTrans34271025	Bootstrap CI for survival function estimates (1-year mortality prediction)	SHAP, Feature Importance (RF)
MLinTrans29543895	Confidence Interval of Standard Error of regression parameters for Hazard Ratios	None
MLinTrans29483521	CI (No Method Specified)	None
MLinTrans32221367	10-Fold-Cross Validation C-Index and AUROC	Random Forest Feature Importance
MLinTrans36572246	Bootstrap CI for survival function estimates	Permutation Importance
MLinTrans33113221	Standard Error of regression parameters for Hazard Ratios, Kaplan-Meier CI	SHAP
MLinTrans34756569	Bootstrap CI for AUROC	None
MLinTrans35389371	Kaplan Meier CI	None
MLinTrans30738152	Bootstrap Calibration	None
MLinTrans32996170	Confidence Interval (CI) for predictions (method not specified)(No Method Specified)	None
MLinTrans36315983	Bootstrap CI (no plot provided) for survival function estimates	SHAP
MLinTrans36938431	Bootstrap CI (no plot provided) for survival function estimates	None
MLinTrans32922997	Confidence Interval (CI) (method not specified)	None
MLinTrans32383068	DeLong Method 95% CI (AUROC)	None
MLinTrans33198650	SE of regression parameters for Hazard Ratios	None
MLinTrans29590219	5-fold Cross Validation for Auroc (Mean + SD)	None
MLinTrans33858815	Bootstrap CI for survival function estimates using deep learning models	None

Table 3.3: Uncertainty and Feature Importance in the Literature Review

Furthermore, a feature importance analysis was incorporated, as it is a highly relevant aspect that enhances the interpretability of the model and provides valuable insights to users. Understanding which variables influence the model predictions helps to assess the reliability of the results, with some articles explicitly integrating advanced feature importance techniques such as SHAP values or other model-agnostic methods.

3.3.3 *Visualization of Model Predictions and Uncertainty*

Most of the reviewed articles, as summarized in Table 3.3, primarily rely on standalone visualizations to present their results, incorporating various types of visualizations, including survival curves, calibration plots, and performance metrics. However, the primary focus is on the visualization of model predictions, particularly the representation of uncertainties in these predictions, as well as the identification of suitable user-facing visualizations, online tools, or dashboards for effectively communicating these uncertainties.

A subset of studies explicitly visualizes model predictions, primarily through survival function plots with confidence intervals, often displayed as shaded regions or error bars. In contrast, other studies report survival probabilities numerically without a corresponding graphical representation. Calibration plots, which assess the alignment of predicted and observed survival probabilities, are also present in some articles, but they rarely include uncertainty measures.

Feature importance plots, while useful for model interpretability, do not directly contribute to the visualization of individual predictions. Performance metrics such as the concordance index and the Brier score are frequently reported in tables, but these typically lack confidence intervals or any representation of predictive uncertainty.

While visual representations of results are common, dedicated tools for interactive exploration of individual predictions remain rare. Two studies (29543895 35389371; see Table 3.2) provide online tools for model outputs, whereas two additional dashboards (33858815 36388342) were identified, though without interactive features. In particular, none of these dashboards integrates uncertainty quantification in their prediction visualizations. Table 3.4 provides an overview of articles incorporating dashboards for presenting model predictions.

A closer examination of the reviewed literature reveals that only a minority of four articles are incorporating some sort of dashboard or user interface for displaying predictions in their study. Although visualizations of results are frequently presented across the literature, explicit development of user-facing dashboards remains uncommon. Among the reviewed articles, only the work of Andres et al. [1] (29543895) and Zafar et al. [55] (35389371) as shown in Table 3.2) present an online tool designed for end-users. Nitski et al. [40] and Roller et al. [46] study included dashboards that featured a design without interactive elements 3.2). In particular, none of the dashboards incorporate uncertainty quantification in their prediction visualizations. Table 3.4 presents an overview of articles that incorporate some form of a dashboard described in the literature.

Paper ID	Dashboard Description	Visualizations and Plots Used
33858815	Dashboard 6.3 to predict 1-year and 5-year survival outcomes for transplant patients, integrating dynamic risk factors over time and providing information for causes like survival, cardiovascular events, graft failure, cancer, and infection.	Pie charts visualize survival probabilities and competing risks (cardiovascular events, graft failure) at different time points before death. However, the dashboard does not explicitly quantify uncertainty in predictions
36388342	Clinical Decision Support Dashboard 6.4 displaying dynamic risk scores over time using a traffic light system (green, yellow, red zones) and detailed feature explanations for both individual and overall decisions.	A line plot tracks risk scores over time, overlaid with traffic light zones to indicate high, medium, and low risk thresholds. Additionally, the top 5 feature importance values are displayed. However, uncertainty in predictions is not explicitly visualized .
35389371	The Lung Transplantation Advanced Prediction Tool (LAPT) 6.2 allows users to input a wide range of patient, donor, and transplant variables to generate individualized post-transplant survival probability predictions using a scoring model	The tool predicts post-lung transplant survival probabilities at 1, 5, and 10 years, calculates a risk score, classifies recipients into Low, Medium, or High risk, and provides the posterior probability for each risk category. However, no graphical survival curve visualization or explicit uncertainty representation is included.
29543895	The Liver Transplant Survival Prediction Dashboard 6.1 allows users to input patient-specific variables (e.g., age, albumin levels, hospitalization status, diabetes) to generate individualized survival predictions. It also enables a comparison between personal survival estimates and population-based Kaplan-Meier survival curves.	Kaplan-Meier survival curves are used to display patient-specific survival probabilities for 1, 2, 3, 5, and 10 years, but no confidence intervals (CI) or other uncertainty measures are included. The model predictions are shown as deterministic survival curves without explicit uncertainty quantification.

Table 3.4: Dashboards in the Literature Review

4

RESULTS

This chapter presents the findings of the systematic review of the literature, model evaluation, uncertainty quantification, and the development of an interactive dashboard.

4.1 RESULTS FROM THE SYSTEMATIC LITERATURE REVIEW

The systematic review of the literature initially examined 50 studies on machine learning for survival analysis, focusing mainly on kidney transplantation. These studies were evaluated with the criteria table 6.1 based on the process detailed in Chapter 3 Section 3.2 to identify reproducible models, assess the availability of tools such as dashboards or calculators, and determine how the literature addresses uncertainty quantification. The Results are briefly described in section 3.3 and summarized in the figure 3.2

Figure 3.3 provides a structured visualization of the frequency with which different methodological aspects were reported. Specific categories, such as variable documentation, data cleaning, performance metrics, and visual representation of results, were covered extensively in most studies, indicating that these aspects were well documented by researchers.

In contrast, several preprocessing steps were inconsistently reported, including imputation, scaling, data transformation, and variable selection. Although these techniques were frequently applied, their documentation was often incomplete, limiting reproducibility and comparability between studies.

A major gap was observed in the reporting model architecture and hyperparameter settings, which were rarely described completely in detail. The lack of this information presents challenges for replicating or adapting existing models, as key methodological components remain unclear.

4.1.1 *Feasibility of Prognostic Models for TxReg Implementation*

A core objective of the review objective 1 described in Chapter Section 1.2) was to assess the feasibility of applying prognostic models (Type A) to the TxReg data described in Chapter 3 Section 3.3.1. The literature review identified 10 articles with models that meet the criteria for

type A models. However, their direct applicability to the TxReg dataset is limited due to missing variables, measurement equivalence issues, or categorization discrepancies. Many type A models exhibited missing variables, as summarized in Table 3.1. Among the reviewed Type A models, four articles with minor missing variables could be considered potential Type A1 models, Paquette et al. [43] article *35700006* was identified as relevant for further analysis and potential adaptation to TxReg data. The study provided comprehensive methodological documentation detailing the variables used, preprocessing steps, feature selection methodology, model architecture and hyperparameters making it reproducible with the constraint that not all variables could be matched.

4.1.2 *Availability of Prognostic Calculators and Dashboards*

Objective 2 described in Section (1.2) focused on assessing the availability of prognostic calculators, which was addressed in the Chapter 3 section 3.3.1. As summarized in Table 3.2, only eight studies presented some sort of calculator or dashboard for survival prediction. Andres et al. [1] and Zafar et al. [55] featured fully accessible online tools, enabling real-time survival predictions. Four studies included dashboards, but none incorporated uncertainty quantification. The identified calculators varied in accessibility, with some tools freely available online, while others were not accessible.

4.1.3 *Uncertainty Reporting in Reviewed Studies*

In chapter 3 section 3.3.2 17 studies mentioned uncertainty quantification as shown in table 3.3 were evaluated. The most commonly used approach was bootstrap confidence intervals, reported in several studies. Other methods included: Kaplan-Meier confidence intervals, cross-validation-based confidence intervals and feature importance-based methods such as SHAP values and permutation importance. The methodology for computing confidence intervals in many study's was not specified. Additionally, no study explicitly focused on uncertainty quantification modeling techniques beyond calculating confidence intervals without further reference and evaluation of uncertainties.

4.2 Comparison of model evaluation on TxReg vs. original study

To address objective 3 (1.2), Chapter applied the in chapter 3 identified A1 model to the TxReg dataset. The selected model, DeepHit, from Paquette et al. [43] with article ID 35700006, was reconstructed following

the methodology of the original study, including data preprocessing, model architecture and evaluation metrics.

The evaluation of DeepHit on TxReg data was compared with the original performance reported by Paquette et al. [43]. Additionally, a Cox proportional hazards model trained on the TxReg dataset was included as a benchmark model. Table 4.1 summarizes the results of the performance.

Evaluation Metric (Test Data)	**35700006 DeepHit**	**35700006 CoxPH**	**TxReg DeepHit**	**TxReg Cox**
Harrel's C-Index	0.661	0.646	0.6433	0.6757
Antolini C-Index	-	-	0.6429	0.6813
Integrated Brier Score (IBS)	0.1528	0.1543	0.1811	0.1791

Table 4.1: Comparison of Evaluation Metrics: 35700006 vs. TxReg Results

The C-Index values reported in Table 4.1 were calculated using the method of Harrell et al. [20], which evaluates discrimination among 1,311,976 comparable pairs in the test data set. Assesses whether the model correctly ranks individuals by risk based on all event times.

The Antolini C-index, on the contrary, was calculated as a time-dependent concordance index, evaluating 650,050 comparable pairs by assessing ranking correctness at the exact failure time of each individual. Unlike Harrell's C-Index, which considers all event times, this approach focuses on ranking performance only at observed event times.

The evaluation of model performance based on Harrell's C-index, Antolini's C-index, and the Integrated Brier Score (IBS) reveals differences between the original DeepHit and CoxPH implementations from article 35700006 and their application to the TxReg dataset.

The Harrell's C-index for DeepHit on TxReg (0.6433) is slightly lower than in the original study (0.661), suggesting a small reduction in discriminative ability. Similarly, the CoxPH model in 35700006 (0.646) performed slightly worse than DeepHit in the original data set but outperformed DeepHit when applied to TxReg data. The TxReg CoxPH model (0.6757) achieved the highest C-index, indicating better ranking performance in this data set.

A similar trend is observed for the Antolini C-index, where DeepHit on TxReg (0.6429) performed worse than the TxReg Cox model (0.6813), reinforcing the observation that the Cox model demonstrates better discrimination ability in this dataset.

Regarding calibration, the Integrated Brier Score (IBS) indicates that DeepHit on TxReg (0.1811) had a higher error rate compared to its original version (0.1528), reflecting a decline in predictive accuracy. The CoxPH model in 35700006 (0.1543) performed similarly to the original DeepHit model, while the TxReg Cox model (0.1791) had a slightly lower IBS than DeepHit on TxReg but still performed worse than the original CoxPH implementation.

To address objective 4 (1.2), uncertainty quantification was implemented using Monte Carlo dropout **??** and bootstrap resampling **??**. The original study did not incorporate uncertainty quantification, which prevented a direct comparison with its results.

Monte Carlo dropout was performed using 1000 forward passes, resulting in very narrow confidence intervals, indicating minimal estimated predictive uncertainty. Bootstrapping was applied with 100 resampled datasets, producing wider confidence intervals, capturing a greater degree of uncertainty in the model predictions. Figures 4.1 and 4.2 illustrate the survival function of the DeepHit model whose implementation is described in section **??** with parameters specified in **??**. The visualizations present survival estimates with uncertainty confidence intervals for both methods, following the methodologies outlined in this study.

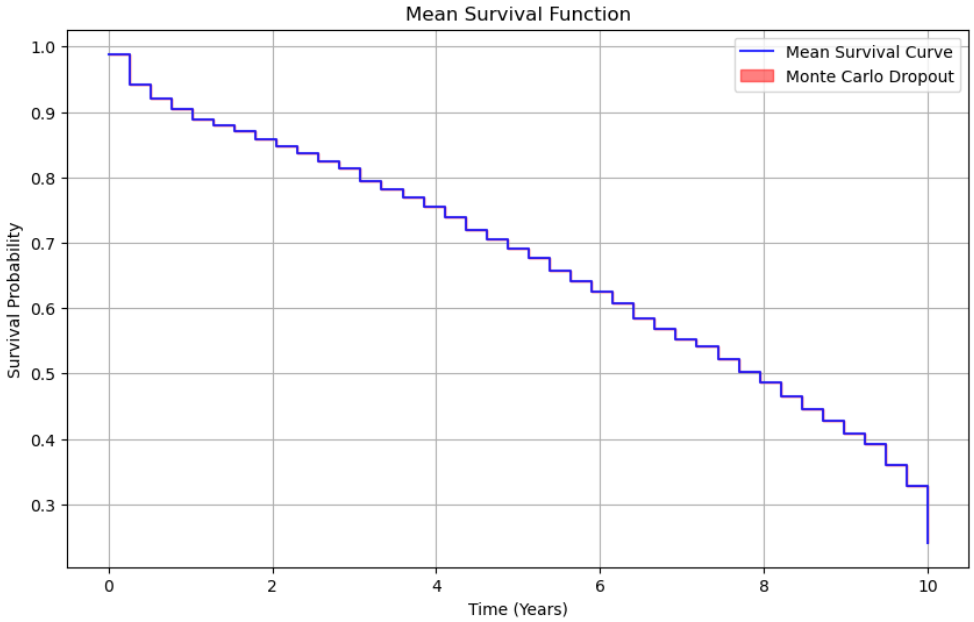

Figure 4.1: Survival Function with Monte Carlo Dropout

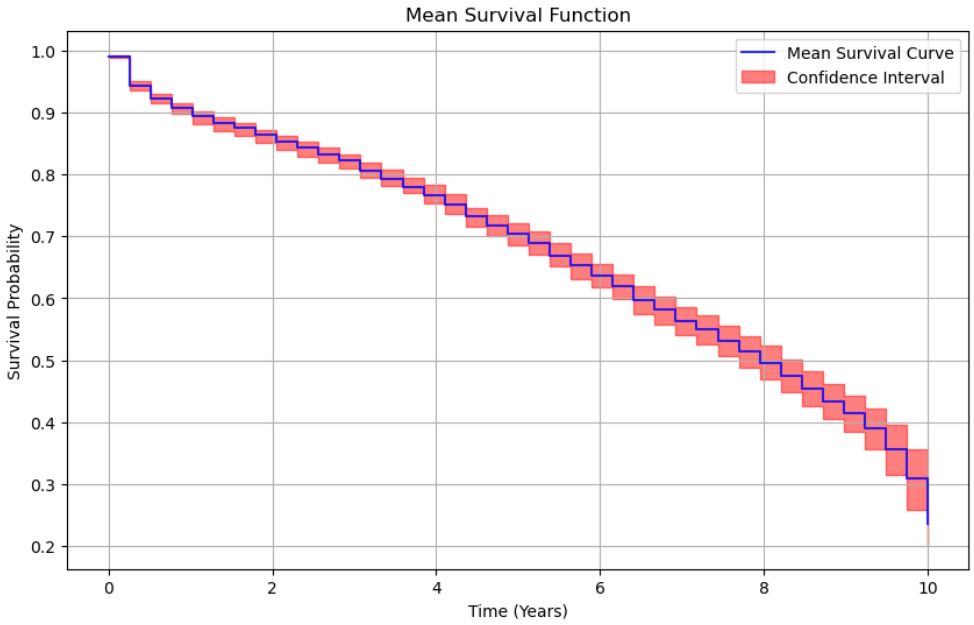

Figure 4.2: Survival Function with Bootstrapping

The results in Figures 4.1 and 4.2 demonstrate clear differences in the uncertainty estimates obtained through Monte Carlo dropout and bootstrapping.

4.3.1 *Dashboard Visualization Outcomes*

To address objective 5 (1.2), an interactive dashboard was developed to provide a comprehensive visualization of individual patient outcomes. The dashboard incorporates individualized survival curves, following the approach proposed by Andres et al. [1] (article 29543895, see 6.1), allowing for patient-specific survival probability estimations over time. To account for predictive uncertainty, Monte Carlo dropout based confidence intervals were integrated, as recommended by Kompa, Snoek, and Beam [30], allowing a more robust interpretation of survival predictions. However, bootstrapping is not applicable in this case, as it requires multiple resampled datasets from an existing data set, while for a new patient, only a single observation is available, making resampling infeasible. Furthermore, a visualization of the importance of the characteristics was implemented based on the article's in the literature review Roller et al. [46], highlighting the most influential predictors in survival estimation. To improve the interpretability of the model, SHAP-based feature analysis proposed by Lundberg and Lee [34], was integrated into the dashboard, allowing for an individualized assessment of the factors contributing to survival predictions. Although permutation importance is commonly used for feature relevance analysis, it could not be applied in this case as it requires repeated permuta-

tions of input features over multiple data points. Given that only a single patient dataset is available for prediction, the necessary resampling process for permutation importance was not feasible.

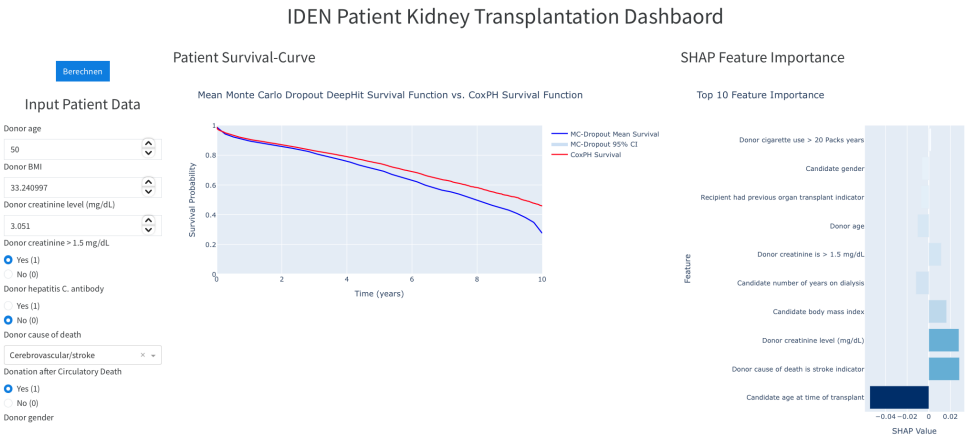

Figure 4.3: Survival Dashboard

5

DISCUSSION AND CONCLUSION

5.1 DISCUSSION OF LITERATURE REVIEW

5.1.1 *Review Methodology and Evaluation Framework*

The hybrid approach detailed in 3.2.1 provided a efficient mechanism for systematically reviewing a large number of articles while ensuring accuracy. The use of a large language model significantly reduced the time required for initial data extraction and matching, while manual verification and refinement ensured the validity of the results.

To further systematize the literature review and ensure methodological transparency, a structured evaluation framework was implemented, as outlined in Table 6.1. This framework enabled a rigorous and reproducible assessment of each study by categorizing key aspects such as the variables used (Criterion 1), model reproducibility, which means a clear description of preprocessing methods, model architectures, hyperparameters (Criterion 4-8) and uncertainty evaluation and methods (Criteria 12–14). By applying these predefined criteria, the review process ensured a standardized and unbiased evaluation of the methodological and practical feasibility of each prognostic model.

Furthermore, the integration of visualization strategies (Criteria 15–16) and applicability considerations (Criterion 17) including tools such as prediction calculators and dashboards offered valuable insights into the interpretability and practical usability of each model. These aspects were particularly relevant for assessing the usability of implementing the models in real-world clinical settings.

The hybrid automated manual review process streamlined the evaluation of relevant literature in combination with the criteria table to ensure a systematic, reproducible, and objective evaluation of each study. This structured methodology strengthened the validity of the literature review.

5.1.2 *Core Findings and Missing Variables*

The findings of this review of the literature highlight the strengths and limitations of the methodology and communication of current machine learning approaches in survival analysis. Although certain aspects such as variable documentation and performance evaluation are

well covered, the lack of transparency in model architecture, hyperparameters, and preprocessing methods presents major obstacles to reproducibility, therefore the review identified just 10 studies with proposed prognostic models, which could be identified as type A. However, their direct applicability to the TxReg was limited due to missing variables, measurement equivalence issues, or discrepancies in variable categorization. The lack of variable compatibility raises concerns about the feasibility of existing survival models to different patient cohorts such as the TxReg.

Paper ID	**Reference**	**Missing Variables**
34414609	Ayers et al. [3]	22
32419922	Senanayake et al. [47]	not available
34448704	Naqvi et al. [38]	7
35700006	Paquette et al. [43]	3
34822363	Thongprayoon et al. [49]	3
30625130	Mark et al. [35]	5
36938431	Linse et al. [33]	55
31926745	Ershoff et al. [14]	80
36388342	Roller et al. [46]	5
33198650	Kantidakis et al. [25]	6

Table 5.1: Missing Variables in Selected Papers

Among the reviewed models, four studies were identified as potential Type A1 with only minor missing variables. The study 35700006 by Paquette et al. [43] was selected for further analysis due to its methodological transparency, including detailed documentation of data sources, preprocessing steps, feature selection, model architecture and hyperparameters. However, this model required adaptations before it could be implemented and applied to the TxReg dataset, as some variables were not available. This highlights the broader challenge of transferring ML models across different registry datasets, where differences in data collection practices and clinical standards limit direct model applicability.

A key finding from the literature review is that many studies do not provide sufficient methodological details on model architectures and hyperparameter configurations, as summarized in table 3.3. This lack of transparency makes it difficult to reproduce the model. Future research should not only emphasize standardized reporting of variables, data preprocessing and evaluation metrics but also focus on appropriate description of the models architecture and hyperparameters to enhance the comparability and reproducibility of ML models in survival analysis.

A secondary finding of the review was to assess the availability of interactive prognostic calculators and dashboards for survival prediction. Only eight studies provided some form of predictive tool or dashboard, and only two of them Andres et al. [1] and Zafar et al. [55] offered fully accessible online calculators. The limited number of publicly available tools suggests that, although survival models are widely studied, their practical implementation in clinical settings remains insufficient researched.

One major limitation observed in all identified calculators and dashboards was the absence of uncertainty quantification. None of the reviewed tools provided uncertainty quantification methods. This is a significant shortcoming, as uncertainty estimation is crucial in clinical decision making [45]. Physicians need to understand not only the predicted survival probabilities but also the degree of confidence in those predictions. Previous studies [30] have emphasized that models without uncertainty estimates may lead to overconfident and potentially misleading predictions, reducing their clinical utility.

5.1.3 *Uncertainty Quantification in Reviewed Studies*

The literature review (3.3.2) reveals that uncertainty quantification for prediction remains under explored or inconsistently addressed in many of the reviewed studies. Although confidence intervals are frequently reported, their methodological derivation is often unclear. Some studies state the presence of confidence intervals without specifying the computational approach, making it difficult to assess their reliability. This lack of transparency in uncertainty reporting raises concerns about the trust and robustness of model predictions and their real world applicability [**JÃűckel:2023** , 9].

These findings directly relate to **RQ2** 3.1, as they highlight significant gaps in how uncertainty is communicated in prognostic models. While bootstrapping is commonly employed for confidence interval estimation, its implementation and visualization are often inadequately documented. Several studies that utilize bootstrap-based uncertainty estimation fail to provide explicit methodological details, making it challenging to compare or reproduce their results. Furthermore, the majority of reviewed articles do not mention uncertainty quantification at all, indicating that this aspect is is often not prioritized in survival model evaluation in the literature.

This implies a broader trend in the researched literature, where model predictions are assessed primarily through performance metrics such as discrimination and calibration while their associated uncertainties remain largely unexplored. As uncertainty quantification plays a crucial role in model trustworthiness and interpretability, it represents a fundamental limitation in the existing literature.

The findings indicate that DeepHit performs worse on TxReg than in its original dataset, while the Cox model serves as a strong benchmark. The lack of uncertainty quantification in the original study was addressed through Monte Carlo dropout and bootstrapping, with notable differences in uncertainty estimation. The dashboard visualization follows best practices from prior research while ensuring improved transparency and interpretability.

5.2.1 *Model Integration and Performance on TxReg*

The evaluation results highlight a substantial discrepancy between the predictive performance of DeepHit on TxReg and its original dataset, as reported by Paquette et al. [43]. The following metrics are summarized in table 4.1. While DeepHit previously demonstrated competitive results, its performance on TxReg is notably lower, with a Harrell's C-index of 0.6433 compared to 0.661 in the original study. In contrast, the Cox proportional hazards model serves as a strong benchmark, achieving a C-index of 0.6757 on TxReg, outperforming DeepHit across all metrics. Additionally, the Integrated Brier Score (IBS) suggests that DeepHit exhibits lower calibration on TxReg than in its initial application, with an IBS of 0.1811 compared to 0.1528 in the original dataset.

These findings directly address **RQ1** 3.1, as they indicate that while DeepHit was designed to model complex survival distributions, its generalization to new datasets, such as TxReg, is suboptimal. The drop in predictive accuracy suggests that model performance is highly dependent on the underlying data distribution and feature representation [6]. This raises concerns regarding the model's robustness and applicability to independent survival registry data, particularly when domain or regional specific characteristics differ .

Several factors may contribute to this observed performance gap. First, differences in feature distributions, censoring patterns, and patient demographics between TxReg and the original dataset may have led to a deterioration in DeepHit's discriminative ability. Machine learning models, particularly deep neural networks, are known to be sensitive to shifts in input distributions, which could explain the observed discrepancy. Additionally, missing variables in the TxReg dataset may have weakened DeepHit's ability to capture relevant temporal dependencies, thereby limiting its predictive power.

Another critical factor is hyperparameter tuning. The hyperparameters used in the original study were optimized for a different dataset and may not be well-suited for TxReg. While re-optimizing hyperparameters could potentially improve performance, the observed results em-

phasize the importance of validating survival models on independent datasets before clinical application.

5.2.2 *Uncertainty Quantification and Its Implications*

Uncertainty quantification was introduced using Monte Carlo dropout **??** and bootstrapping **??**, revealing differences in confidence interval estimates. Monte Carlo dropout resulted in narrower confidence intervals, indicating lower estimated uncertainty, while Bootstrapping produced wider intervals, capturing a broader range of variability in the model's predictions.

Another critical observation is the absence of uncertainty quantification in existing prognostic calculators and survival dashboards. Despite the increasing adoption of machine learning in clinical decision support, none of the reviewed tools incorporated measures of predictive uncertainty. This represents a problematic gap, as reliable uncertainty estimates could improve trust in model predictions and enable more informed decision-making. The survival function plots presented in Figures 4.1 and 4.2 demonstrate how uncertainty can be visualized, yet further improvements are needed. User-centered dashboards with appropriate visualizations and the integration of uncertainty measures could enhance the interpretability and usability of predictive models.

5.2.3 *Dashboard Development and Visual Communication*

The proposed dashboard is inspired by the work of Andres et al. [1] (article 29543895, see 6.1) to present individual patient survival curves to enhance interpretability. Following the principles outlined by Barda et al. [5], the visualization seeks to minimize cognitive overload and support clinical decision-making by integrating model-agnostic explanations. Representing survival functions in this way provides an effective means of visualizing survival rates across different time points.

These design considerations directly address **RQ3**(3.1), as it explores how established methods from the literature, particularly in uncertainty quantification and dashboard based result visualization, can be adapted to improve communication and interpretation of model outputs on TxReg data. To ensure interpretability, the proposed dashboard incorporates uncertainty quantification through Monte Carlo dropout confidence intervals, as recommended by Kompa, Snoek, and Beam [30].

Furthermore, the dashboard incorporates information on the importance of the variables similar to the dashboard of Roller et al. [46], allowing users to understand which factors contribute most to the survival predictions. This follows recommendations from Barda et al. [5],

emphasizing the necessity of transparent feature attributions to mitigate bias and supports the understanding behind the model decisions. The dashboard proposed by Zafar et al. [55] includes a scoring metric, which, according to Barda et al. [5], can lead to cognitive biases relying on incomplete information. Additionally, the lack of interpretability poses a significant challenge user could interpret scores without a deeper understanding of the underlying models. This can result in misconceptions or unjustified confidence in the model's predictions [5, 30].

To mitigate these risks, our dashboard deliberately excludes a scoring metric. Instead, it prioritizes transparency through direct survival curve visualization with uncertainty quantification. By adapting best practices from the literature, the proposed dashboard ensures that predictive outputs are presented in a way that is both interpretable and clinically meaningful. This approach aligns with prior research while addressing gaps in uncertainty reporting and result visualization, ultimately contributing to more informed decision-making.

5.3 Limitations and Future Directions

While this study provides valuable insights into survival modeling and uncertainty quantification, several limitations must be acknowledged. These limitations primarily relate to the methodological constraints in the literature review.

5.3.1 *Challenges in Model Reproducibility and Transferability*

One of the primary limitations of this study concerns the reproducibility of the DeepHit model. The original implementation was reconstructed as faithfully as possible based on the details provided in Paquette et al. [43] article, yet minor differences in preprocessing, feature encoding, and hyperparameter settings may have influenced the observed performance. This issue reflects a broader challenge in the reviewed literature: the lack of standardized reporting practices for model architectures and training procedures.

Additionally, this study underscores the inherent challenges of generalizing survival models across different datasets, since generalizability itself constitutes a substantial research challenge. Although DeepHit demonstrated strong performance in the original study, its predictive accuracy declined when applied to the TxReg dataset, underscoring the limitations of direct model transfer. The presence of missing variables, differences in data distributions, and dataset-specific biases all contribute to this challenge.

5.3.2 *Gaps in Uncertainty Quantification Methodologies*

Another key limitation of this study is its focus on only two uncertainty quantification methods: Monte Carlo Dropout and Bootstrapping. While these approaches provide valuable insights, they represent only a subset of the available techniques. Alternative methods, such as Bayesian neural networks, conformal prediction, and ensemble-based uncertainty estimation, could be explored to further enhance confidence estimation in survival models.

Moreover, while the study demonstrated that bootstrapping captures greater predictive uncertainty than Monte Carlo dropout, it remains unclear which method provides the most reliable uncertainty estimates for clinical decision making. Future research should conduct systematic comparisons of different uncertainty quantification techniques and assess their impact on real-world transplant predictions. This study has barely scratched the surface of uncertainty quantification, underscoring the need for deeper investigations to obtain more reliable results and foster interpretable uncertainty assessment.

5.3.3 *Dashboard and Clinical Applicability*

The practical integration of ML-based survival models into clinical workflows remains a major challenge. Although this study developed an interactive dashboard for individualized survival predictions, it is important to recognize that clinical decision support systems require extensive validation before they can be deployed in practice. The lack of uncertainty quantification in existing prognostic calculators further highlights a gap in current predictive tools, as confidence estimates are essential for informed medical decision making. To enhance the practical utility of survival models, future research should prioritize the development of user-centric dashboards tailored to clinical needs **Barda:2020**, Wang et al. [51]. The proposed dashboard was developed without direct user feedback, and therefore it lacks empirical validation regarding its usability and effectiveness.

Future work should focus on improving the interpretability and usability of survival models, incorporating interactive components, refined uncertainty visualizations, and adaptive features that support clinical decision making. Refinement of visual representations of uncertainty and incorporating explanations of the importance of characteristics could significantly enhance the clinical relevance of predictive models. Furthermore, interdisciplinary collaboration between machine learning researchers and users will be essential to ensure that survival models are interpretable in real-world medical settings. Incorporating insights from medical professionals will be crucial in optimizing dash-

board functionality, improving transparency, and facilitating seamless integration into clinical workflows.

5.4 CONTRIBUTIONS TO SURVIVAL ANALYSIS RESEARCH

This study contributes to the field of survival analysis in several key areas:

- **Methodological Transparency:** The literature review highlights the lack of standardized documentation in the current literature on kidney transplant survival analysis, emphasizing the need for improved reporting practices to enhance reproducibility and comparability across studies.
- **Enhanced Understanding of Model Transferability:** By applying an A1 model to the TxReg dataset and benchmarking it against the Cox model, this study demonstrates the challenges of transferring survival models to new registries with similar settings.
- **Uncertainty Quantification in Survival Models:** This study evaluates uncertainty quantification methods used in the literature and existing online tools, such as calculators and dashboards, identifying significant gaps in their availability, methodologies used, and usability for end users.
- **Bridging the Gap Between Research and Clinical Application:** The developed dashboard enhances the interpretability of survival predictions by visualizing survival probabilities and uncertainty, providing a more practical, transparent, and user-friendly tool.

5.5 CONCLUDING REFLECTIONS AND IMPLICATIONS FOR PRACTICE

This study provides a comprehensive analysis of machine learning-based survival models in kidney transplantation, identifying key challenges, and proposing methodological advances. While significant research efforts continue to introduce new models with enhanced flexibility in modeling complex survival patterns, their applicability to similar research contexts with different datasets remains a challenge. Leveraging insights from the existing literature to address similar problems significantly fosters broader knowledge transfer.

A critical finding of this study is that, despite the increasing number of survival models being developed, the focus remains on evaluating predictive performance using standard survival metrics, while uncertainty quantification in model predictions receives little attention. This is particularly concerning in a medical domain where treatment decisions

directly impact human lives. Without a clear understanding of predictive uncertainty, clinical decision-making risks becoming too reliant or biased on potentially misleading point estimates.

Furthermore, there is a notable gap in user-centered tools that translate research findings into practical clinical applications. Although survival modeling techniques continue to advance, their integration into clinical workflows remains limited, hindering their potential to make a real impact in kidney transplant medicine. Developing accessible, interactive, and interpretable decision support tools is essential to bridge this gap between machine learning research and real-world medical applications.

6

APPENDIX

Table 6.1: Search Criteria and Details

Criterion	**Details**
1. Variables	List all variables; separate lists for different models if applicable.
2. Data Source	National register, international register, local data source, not mentioned.
3. Data Availability	Categories: publicly accessible, available on request, not accessible.
5. Data Preparation	Includes: cleaning, imputing, scaling, transformations, train-test split, variable selection, dimension reduction.
6. Model Type	Kaplan-Meier, CoxPH, AFT, Random Survival Forests, Neural Networks, Bayesian Models, Deep Learning, DeepSurv, DeepHit, Other.
7. Model Architecture	Details include layer types, dimensions, activations, number of layers, and other.
8. Hyperparameters	Examples: optimizer, learning rate, batch size, epochs, regularization.
9. Metrics	Examples: confusion matrix, accuracy, F1-score, ROC-AUC, C-index, Brier score, calibration plots, AIC, BIC.
10. Reproducibility	Based on criteria 1,5,7,8 is the model reproducible (Yes/No)
11. Prediction or Estimation	Focus of the model: prediction or estimation.
12. Mention of Uncertainty	Keywords like uncertainty estimation, confidence interval, Monte Carlo, prediction interval.
13. Types of Uncertainty	Aleatoric or epistemic uncertainty.
14. Uncertainty Quantification	Summary of techniques or metrics used for uncertainty quantification.
15. Visual Representation	Types include plots, diagrams, charts, graphs, and tables.
16. Visualization Concept	Strategies like dashboards, GUIs, or tools for interpretation.
17. Applicability	Tools such as prediction calculators, formulas, or online tools.
18. Individual Predictions	Support for patient-specific or individual predictions.

Table 6.2: Search Criteria and Details

SEARCH CRITERIA FOR THE LLM

```
search_criteria = """
1. Variables:
    - List all variables that are used to train the model as a
       python list
    - If necessary make another list if other models used different
       variables.
2. Data Source:
    - National Register, International Register, Local Data Source,
       Not mentioned.
3. Availability of Data:
    - Indicate the accessibility category of the data. Possible
       categories are:
      publicly accessible, available on request, not accessible.
5. Data Preparation:
    - Are the steps for data preparation and cleaning described in
       detail? If yes,
      list the methods briefly:
        - Data Cleaning, Imputing, Scaling, Data Transformation,
           Train-Test Split,
          Variable Selection, Dimensionality Reduction, Other.
6. Model Type:
    - Which models are used in the paper? List standard models, e.g
       ., Cox Proportional
      Hazards Model, Random Survival Forests, Neural Network-based
         Models (DeepSurv,
      DeepHit, etc.), Bayesian Survival Models, or others.
7. Model Architecture:
    - Details of the model architecture, including the type, layers
       , size, and connections.
8. Hyperparameters:
    - Are the hyperparameters described? Include details such as
       learning rate, batch
      size, number of epochs, optimizer, regularization.
9. Metrics:
    - Are evaluation metrics described? Examples: Confusion Matrix,
       Accuracy,
      C-index, Brier Score, Calibration Plots, etc.
10. Reproducibility:
    - Based on criteria (4-8), assess whether the model is
       reproducible. Provide
      a clear conclusion: Yes/No with reasons.
11. Prediction or Estimation:
    - Does the model focus on prediction or estimation?
12. Mention of Uncertainty:
```

```
    - Search for mentions of uncertainty in the paper: Keywords
        include "Uncertainty
      Estimation", "Confidence Interval", "Prediction Interval",
          etc.
13. Types of Uncertainty:
    - Are types of uncertainty discussed (Aleatoric, Epistemic)?
14. Uncertainty Quantification:
    - Techniques or metrics used to quantify uncertainty in the
        model.
15. Visual Representation of Results:
    - List the types of visualizations used, such as Kaplan-Meier
        Plots, ROC Curves,
      Dashboards, etc.
16. Visualization Concept:
    - Is there a discussion about visualization strategies (e.g.,
        dashboards)?
17. Applicability and Prediction Calculator:
    - Does the paper mention the use of a prediction calculator or
        formula?
18. Individual Predictions:
    - Can the model provide individual predictions for new data?
"""
```

Listing 6.1: Search Criteria for Academic Paper Analysis

ADDITIONAL INSTRUCTIONS FOR THE LLM

```
additional_instructions = """
1. Ensure Consistency and Order: Follow the order of the search
    criteria exactly
   as given. Each response should match the sequence of the search
       criteria.
2. Brief and Precise Answers: Provide concise and precise answers.
3. Use Delimiter ";": When a Yes/No or category question is
    followed by an explanation,
   use ";" to separate them.
4. Python Lists: Output Python lists in one line.
5. Answer "None" or "not mentioned": If a question does not fit the
     context of
   the paper or if the information is not available, answer with "
       None" or
   "not mentioned" respectively.
6. JSON Format: Structure the output in JSON format for easy
    parsing and iteration.
7. Maintain Consistency: The response should be parsable and
    consistent in structure.
8. Adaptation to Model Types: Use the given model list for
    consistent naming
   conventions, and ensure explanations match model-specific
       requirements.
```

```
9. Each model type has a different architecture. Adapt accordingly
    but maintain
   consistency in sentence structure.
10. Each model type has different hyperparameters. Adapt
    accordingly but maintain
    consistency in sentence structure.
"""
```

Listing 6.2: Additional Instructions for Responses

ANSWER FORMAT FOR THE LLM

```
{
    "variables": {
        "Cox Regression": ["age", "sex", "weight", "height", "
            ethnicity", "donation type", "creatinine level", "
            history of diabetes", "hypertension diagnosis", "
            hepatitis C diagnosis", "smoking habit", "diagnosis", "
            years on dialysis", "angina", "BMI"],
        "Random Survival Forest": ["Donor age", "DR locus 1", "A
            locus 2", "Height", "Donor diabetes", "Donor
            hypertension", "Cause of death", "Creatinine
            terminal", "Oliguria, Race", "Age at transplant", "HLA-
            DR mismatch", "Pre-emptive transplant", "Duration of
            peritoneal dialysis", "Duration of haemodialysis", "
            Primary renal disease", "Smoking", "Peripheral vascular
             disease", "Age at starting renal replacement therapy",
             "number of previous rejections"]
    },
    "data_source": "National Register",
    "availability_of_data": "available on request",
    "data_preparation": {
        "data_cleaning": "removed outliers and missing values;
            threshold >20% missing",
        "imputing": "used mean imputation for missing values",
        "scaling": "standardization",
        "data_transformation": "Calculated BMI",
        "train_test_split": "80-20 split",
        "variable_selection": "PCA"
    },
    "model_type": ["Cox Proportional Hazards Model","Random
        Survival Forest"],
    "model_architecture": {
        "details": [
            {
                "model": "Cox Proportional Hazards Model",
                "layers": []
            },
            {
                "model": "Random Survival Forest",
```

```
                "layers": [
                    {"type": "Tree", "count": 100, "max_depth": 25,
                         "min_samples_split": 400, "min_samples_
                        leaf": 200, "max_features": 13}
                ]
            }
        ]
    },
    "metrics": ["C-index", "Brier Score", "Calibration plots"]
}
```

Listing 6.3: Example Output Format for Responses

MESSAGE LITERATURE REVIEW

```
messages = [
    {
        "role": "system",
        "content": f"""You are a capable assistant skilled in
            extracting and analyzing information from academic
            papers based on given criteria. Answering the following
             Search Criteria:\n {search_criteria}.
                            \n Use the following instructions for
                                your response: {
                                additional_instructions}\n
                            \n Use this example format for the
                                structure of your answer: {
                                answer_format}\n
                            """
    },
    {
        "role": "user",
        "content": f"Please analyze the provided text and extract
            information according to the given criteria. Text:{text
            }"
    }
]
```

Listing 6.4: Messages for Contextual Instructions

MESSAGE VARIABLE MATCHING

```
messages = [
    {
        "role": "system",
        "content": "You are a capable assistant skilled in
            comparing variable names and finding matches based on
```

```
        their name or meaning, even if the names are different.
        "
},
{
    "role": "system",
    "content": "Consider that some variables may require
        adjustments or calculations to match. Or other
        variables are sub or top categories where you have to
        find the belonging sub or top category."
},
{
    "role": "system",
    "content": """
    Please also consider the following rules:
    1. For each input variable, find the best matching TxReg
        variable based on their name and meaning.
    2. If a direct match is found, use it.
    3. If no direct match is found, look for variables that
        could be adjusted or calculated to match. Example:
       data["Donor body mass index"] = data.apply(
           lambda row: row['spender_postmortem::weight_kg'] /
               ((row['spender_postmortem::height_cm'] / 100) **
                2)
           if row['spender_postmortem::height_cm'] != 0 else
               None,
           axis=1
       )
    4. Second example for calculations:
       data["Donor Female"] = data["empfaenger::sex"].apply(
           lambda x: "1" if x == "female" else 0)
    5. When there is a need to adjust and calculate the
        variable, structure your answer like the example in 3.
        and directly provide the calculation for the new
        variable.
    6. Only if there is no direct match and the variable is a
        subcategory, match the appropriate top category (and
        vice versa). For example, types of death like heart
        attack could belong to the top category "death reasons"
         or "death circumstances."
    7. Format the result as a Python dictionary (JSON format),
        where the input variable is the key, and the matching
        TxReg variable or calculation is the value. Example:
       "Donor age": data["spender_postmortem::age"]
    8. When doing calculations, show the calculation directly
        in the dictionary value. For example:
       "Donor body mass index": data.apply(lambda row: row['
           spender_postmortem::weight_kg'] / ((row['
           spender_postmortem::height_cm'] / 100) ** 2) if row
           ['spender_postmortem::height_cm'] != 0 else None,
           axis=1)
    9. If no match is found, set the value to None.
    10. Example output:
```

```
            {
                "Donor age": data["spender_postmortem::age"],
                "Donor history of hypertension": data["
                    spender_postmortem::hypertension"],
                "Donor history of diabetes": data["
                    spender_postmortem::diabetes_dso"],
                "Donor body mass index": data.apply(lambda row: row
                    ['spender_postmortem::weight_kg'] / ((row['
                    spender_postmortem::height_cm'] / 100) ** 2) if
                     row['spender_postmortem::height_cm'] != 0 else
                     None, axis=1),
                "Donor creatinine level (mg/dL)": data["
                    spender_postmortem_labor_klinische_chemie::
                    creatin_umol_per_l"].apply(lambda x: x * 0.0113
                     if pd.notnull(x) else None),
                "Donor creatinine is > 1.5 mg/dL": data["Donor
                    creatinine level (mg/dL)"] > 1.5
            }
        """
    },
    {
        "role": "system",
        "content": "Here are the variables included in the TxReg
            Database as a Python list: \n{txreg_variables_list}"
    },
    {
        "role": "user",
        "content": "Here is the input list of variables as a Python
             list: \n{input_variables}"
    }
]
```

Listing 6.5: Messages for Contextual Instructions

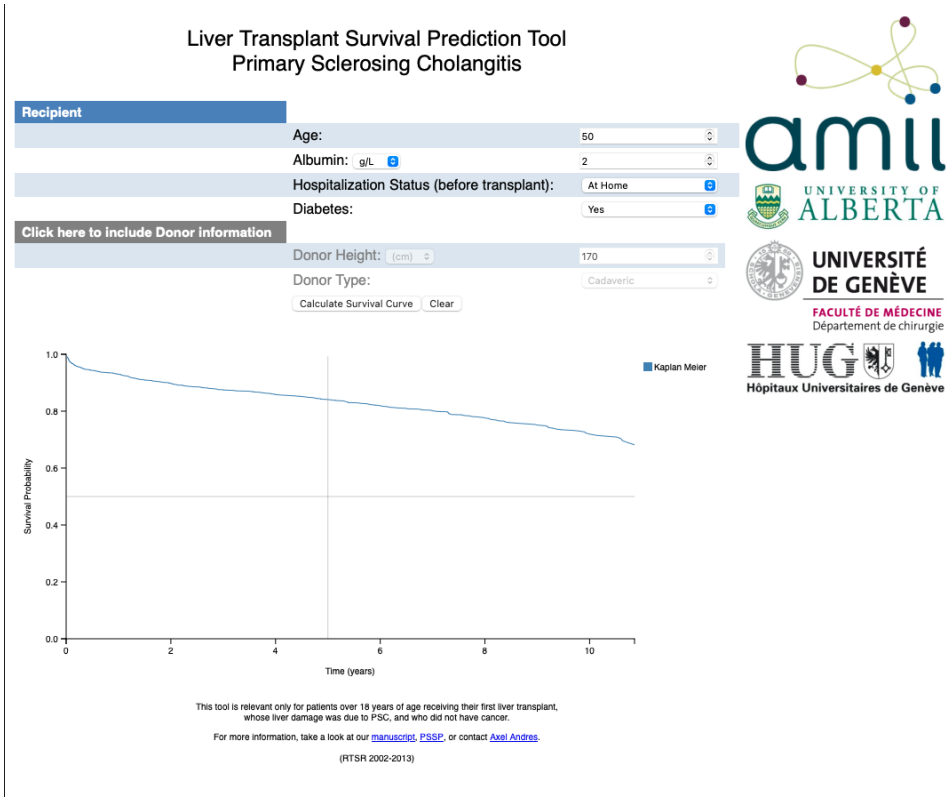

Figure 6.1: Dashbaord 29543895

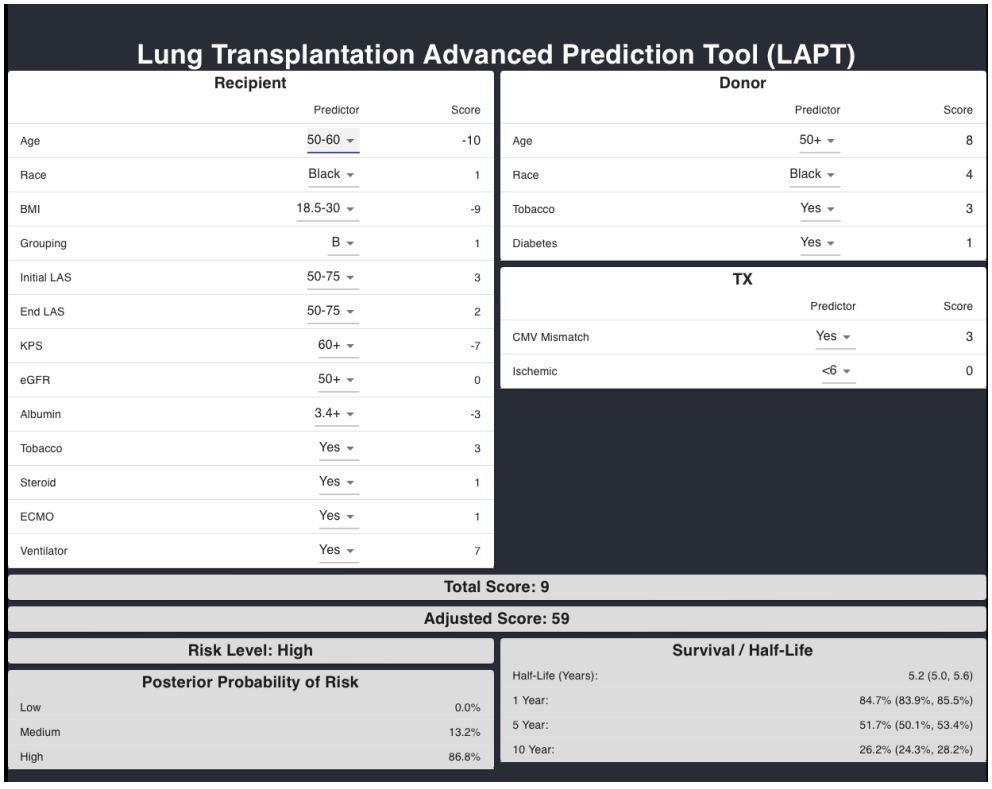

Figure 6.2: Dashboard 35389371

Figure 6.3: Dashboard 33858815

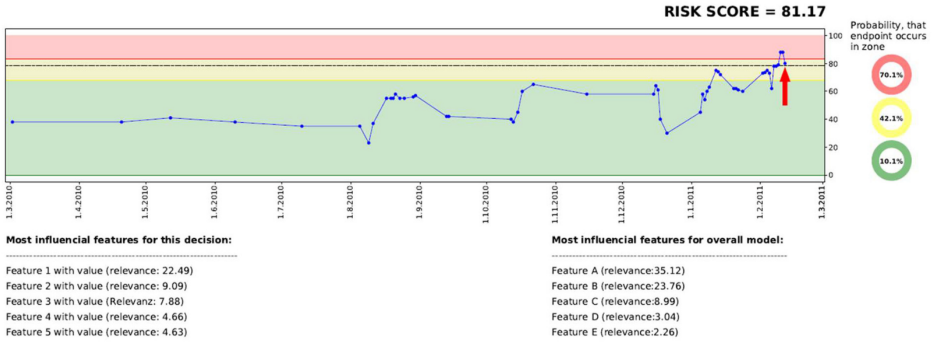

Figure 6.4: Dashboard 36388342

BIBLIOGRAPHY

[1] A. Andres, A. Montano-Loza, R. Greiner, M. Uhlich, P. Jin, B. Hoehn, D. Bigam, J.A.M. Shapiro, and N.M. Kneteman. "A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis." In: *PLOS ONE* 13.3 (2018), e0193523. DOI: 10.1371/journal.pone.0193523. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0193523.

[2] Laura Antolini, Mihaela van der Schaar, and Riccardo D. Ambrogi. "A Time-Dependent Discrimination Index for Survival Data." In: *Statistics in Medicine* 24.24 (2005), pp. 3927–3944. DOI: 10.1002/sim.2420.

[3] Brian Ayers, Tuomas Sandholm, Igor Gosev, Sunil Prasad, and Arman Kilic. "Using machine learning to improve survival prediction after heart transplantation." In: *Journal of Cardiac Surgery* 36.11 (2021), pp. 4113–4120. DOI: 10.1111/jocs.15917. URL: https://onlinelibrary.wiley.com/doi/10.1111/jocs.15917.

[4] María Dolores Ayllón et al. "Validation of Artificial Neural Networks as a Methodology for Donor-Recipient Matching for Liver Transplantation." In: *Liver Transplantation* 24.2 (2018), pp. 192–203. DOI: 10.1002/lt.24870. URL: https://pubmed.ncbi.nlm.nih.gov/28921876/.

[5] Noa Barda, Mareike Seiffert, Sergio Wollenstein-Betech, Gal Yona, Jonathan Moss, Uri Shalit, and Ran D. Balicer. "A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare." In: *BMC Medical Informatics and Decision Making* 20.1 (2020), p. 257. DOI: 10.1186/s12911-020-01273-4.

[6] Savannah Bergquist, Gabriel A. Brooks, Mary Beth Landrum, Nancy L. Keating, and Sherri Rose. "Uncertainty in lung cancer stage for survival estimation via set-valued classification." In: *Statistics in Medicine* 41.19 (2022), pp. 3772–3788. DOI: 10.1002/sim.9448. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9540678/.

[7] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN: 978-0387310732.

[8] Norman E. Breslow. "Discussion on Professor Cox's Paper." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–220. ISSN: 0035-9246. URL: https://www.jstor.org/stable/2985049.

[9] Tianyi Chen, Yingzhou Lu, Nan Hao, Yuanyuan Zhang, Capucine Van Rechem, Jintai Chen, and Tianfan Fu. "Uncertainty Quantification on Clinical Trial Outcome Prediction." In: *arXiv preprint arXiv:2401.03482* (2024). URL: https://arxiv.org/abs/2401.03482.

[10] Tracy G. Clark, Michael J. Bradburn, Stephen B. Love, and David G. Altman. "Survival Analysis Part I: Basic Concepts and First Analyses." In: *British Journal of Cancer* 89.2 (2003), pp. 232–238. DOI: 10.1038/sj.bjc.6601118. URL: https://www.researchgate.net/publication/10657786_Survival_Analysis_Part_I_Basic_Concepts_and_First_Analyses.

[11] David R. Cox. "Regression models and life-tables." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202.

[12] Bradley Efron. "Bootstrap Methods: Another Look at the Jackknife." In: *The Annals of Statistics* 7.1 (1979), pp. 1–26. DOI: 10.1214/aos/1176344552.

[13] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Boca Raton, FL, USA: Chapman & Hall/CRC Monographs on Statistics and Applied Probability, 1993.

[14] Brent D. Ershoff, Christine K. Lee, Christopher L. Wray, Vatche G. Agopian, Gregor Urban, Pierre Baldi, and Maxime Cannesson. "Training and Validation of Deep Neural Networks for the Prediction of 90-Day Post-Liver Transplant Mortality Using UNOS Registry Data." In: *Transplantation Proceedings* 52.1 (2020), pp. 246–258. DOI: 10.1016/j.transproceed.2019.10.019. URL: https://escholarship.org/uc/item/4rj1b0m3.

[15] GROBID Developers. *GROBID: GeneRation Of Bibliographic Data*. Accessed: 2025-01-18. 2025. URL: https://grobid.readthedocs.io/en/latest/Introduction/.

[16] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." In: *Proceedings of the International Conference on Machine Learning (ICML)* 48 (2016), pp. 1050–1059.

[17] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, et al. "A Survey of Uncertainty in Deep Neural Networks." In: *arXiv preprint arXiv:2107.03342* (2021).

[18] Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. "Assessment and comparison of prognostic classification schemes for survival data." In: *Statistics in Medicine* 18.17-18 (1999), pp. 2529–2545. DOI: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5.

[19] Frank E. Jr. Harrell, Kerry L. Lee, and Daniel B. Mark. "Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors." In: *Statistics in Medicine* 15.4 (1996), pp. 361–387.

[20] Frank E. Harrell, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Rolando A. Rosati. "Evaluating the Yield of Medical Tests." In: *Journal of the American Medical Association (JAMA)* 247.18 (1982). Introduces the Concordance Index (C-Index) as a measure for evaluating predictive accuracy in survival analysis., pp. 2543–2546. DOI: 10.1001/jama.1982.03320430047030.

[21] Patrick J. Heagerty, Thomas Lumley, and Margaret S. Pepe. "Time-dependent ROC curves for censored survival data and a diagnostic marker." In: *Biometrics* 56.2 (2000), pp. 337–344.

[22] Eyke Hullermeier and Willem Waegeman. "Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods." In: *Machine Learning* 110.3 (2021), pp. 457–506. DOI: 10.1007/s10994-021-05946-3.

[23] Hemant Ishwaran, Uday B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. "Random survival forests." In: *The Annals of Applied Statistics* 2.3 (2008), pp. 841–860.

[24] John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. 2nd ed. Hoboken, NJ, USA: Wiley-Interscience, 2002. DOI: 10.1002/9781118032985.

[25] Georgios Kantidakis, Hein Putter, Claudia Lancia, and Emanuele Biganzoli. "Survival prediction models since liver transplantation - comparisons between Cox models and machine learning techniques." In: *BMC Medical Research Methodology* 20.1 (2020), p. 277. DOI: 10.1186/s12874-020-01153-1. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-01153-1.

[26] Edward L. Kaplan and Paul Meier. "Nonparametric estimation from incomplete observations." In: *Journal of the American Statistical Association* 53 (1958), pp. 457–481.

[27] 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." In: *BMC Medical Research Methodology*. Vol. 18. 1. London, United Kingdom: BioMed Central, 2018, p. 24.

[28] Armen Der Kiureghian and Ove Ditlevsen. "Aleatory or epistemic? Does it matter?" In: (2007). URL: https://www.researchgate.net/publication/245585007_Aleatory_or_epistemic_Does_it_matter.

[29] John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. New York, NY, USA: Springer Science & Business Media, 2003.

[30] Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. "Second opinion needed: communicating uncertainty in medical machine learning." In: *npj Digital Medicine* 4 (2021), p. 4. DOI: 10.1038/s41746-020-00367-3.

[31] Jerald F. Lawless. *Statistical Models and Methods for Lifetime Data*. 2nd ed. Hoboken, NJ, USA: Wiley, 2002. DOI: 10.1002/9781118033005.

[32] Changhee Lee, William R. 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*. Vol. 32. 1. New Orleans, LA, USA: AAAI Press, 2018.

[33] Björn Linse, Martin Ohlsson, Josef Stehlik, Lars H. Lund, Bozena C. Andersson, and Johan Nilsson. "A machine learning model for prediction of 30-day primary graft failure after heart transplantation." In: *Heliyon* 9.3 (2023), e14282. DOI: 10.1016/j.heliyon.2023.e14282. URL: https://pubmed.ncbi.nlm.nih.gov/36938431/.

[34] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2017, pp. 4765–4774.

[35] Ethan Mark, David Goldsman, Brian Gurbaxani, Pinar Keskinocak, and Joel Sokol. "Using machine learning and an ensemble of methods to predict kidney transplant survival." In: *PLOS ONE* 14.1 (2019), e0209068. DOI: 10.1371/journal.pone.0209068. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209068.

[36] Dennis Medved, Mattias Ohlsson, Peter Höglund, Bodil Andersson, Pierre Nugues, and Johan Nilsson. "Improving prediction of heart transplantation outcome using deep learning techniques." In: *Scientific Reports* 8.1 (2018), p. 3613. DOI: 10.1038/s41598-018-21417-7. URL: https://www.nature.com/articles/s41598-018-21417-7.

[37] Sadiq Hussain Dana Rezazadegan Li Liu Mohammad Ghavamzadeh Paul Fieguth Xiaochun Cao Abbas Khosravi U Rajendra Acharya Vladimir Makarenkov Saeid Nahavandi Moloud Abdar Farhad Pourpanah. "A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges." In: *arXiv preprint arXiv:2011.06225* (2021).

[38] Syed Asil Ali Naqvi, Karthik Tennankore, Amanda Vinson, Patrice C. Roy, and Syed Sibte Raza Abidi. "Predicting Kidney Graft Survival Using Machine Learning Methods: Prediction Model Development and Feature Significance Analysis Study." In: *Journal of Medical Internet Research* 23.8 (2021), e26843. DOI: 10.2196/26843. URL: https://www.jmir.org/2021/8/e26843/.

[39] Wayne Nelson. "Hazard plotting for incomplete failure data." In: *Journal of Quality Technology* 1.1 (1969), pp. 27–52.

[40] Osvald Nitski et al. "Long-term mortality risk stratification of liver transplant recipients: real-time application of deep learning algorithms on longitudinal data." In: *The Lancet Digital Health* 3.5 (2021), e295–e305. DOI: 10.1016/S2589-7500(21)00040-6. URL: https://pubmed.ncbi.nlm.nih.gov/33858815/.

[41] OpenAI. *GPT-4 Turbo Model*. Accessed: 2024-02-20. 2024. URL: https://openai.com/research/gpt-4.

[42] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Balaji Lakshminarayanan, Jasper Snoek, Dustin Tran, and Alexander Wiltschko. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift." In: *Advances in Neural Information Processing Systems (NeurIPS)* 32 (2019).

[43] François-Xavier Paquette, Amir Ghassemi, Olga Bukhtiyarova, Moustapha Cisse, Natanael Gagnon, Alexia Della Vecchia, Hobivola A. Rabearivelo, and Youssef Loudiyi. "Machine Learning Support for Decision-Making in Kidney Transplantation: Step-by-Step Development of a Technological Solution." In: *JMIR Medical Informatics* 10.6 (2022), e34554. DOI: 10.2196/34554. URL: https://pubmed.ncbi.nlm.nih.gov/35700006/.

[44] Mireia Ribera. "Can we do better explanations? A proposal of User-Centered Explainable AI." In: *Universitat de Barcelona - Departament de Matemàtiques i Informàtica* (2020). Institut de Matemàtica de la Universitat de Barcelona, ribera@ub.edu.

[45] Richard D Riley et al. "Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches." In: *BMJ* 388 (2024), bmj–2024–080749. DOI: 10.1136/bmj-2024-080749. URL: https://www.bmj.com/content/388/bmj-2024-080749.

[46] Roland Roller et al. "Evaluation of a Clinical Decision Support System for Detection of Patients at Risk After Kidney Transplantation." In: *Frontiers in Public Health* 10 (2022), p. 979448. DOI: 10.3389/fpubh.2022.979448. URL: https://pubmed.ncbi.nlm.nih.gov/36388342/.

[47] Sameera Senanayake, Adrian Barnett, Nicholas Graves, Helen Healy, and Sanjeewa Kularatna. "Using machine learning techniques to develop risk prediction models to predict graft failure following kidney transplantation: protocol for a retrospective cohort study." In: *F1000Research* 8 (2020), p. 1810. DOI: 10.12688/f1000research.20661.2. URL: https://f1000research.com/articles/8-1810/v2.

[48] Peyman Tavallali, Hamed Hamze Bajgiran, Danial J. Esaid, and Houman Owhadi. "Decision Theoretic Bootstrapping." In: *arXiv preprint arXiv:2103.09982* (2021). URL: https://arxiv.org/abs/2103.09982.

[49] Charat Thongprayoon, Caroline C. Jadlowiec, Napat Leeaphorn, Jackrapong Bruminhent, Prakrati C. Acharya, Chirag Acharya, Pattharawin Pattharanitima, Wisit Kaewput, Boonphiphop Boonpheng, and Wisit Cheungpasitporn. "Feature Importance of Acute Rejection among Black Kidney Transplant Recipients by Utilizing Random Forest Analysis: An Analysis of the UNOS Database." In: *Medicines* 8.11 (2021), p. 66. DOI: 10.3390/medicines8110066. URL: https://www.mdpi.com/2305-6320/8/11/66.

[50] 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." In: *Statistics in Medicine* 30.10 (2011), pp. 1105–1117.

[51] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. "Designing Theory-Driven User-Centric Explainable AI." In: (2019), pp. 1–15. DOI: 10.1145/3290605.3300831.

[52] Ping Wang, Yan Li, and Chandan K. Reddy. "Machine Learning for Survival Analysis: A Survey." In: *arXiv preprint arXiv:1708.04649* (2017).

[53] Waloddi Weibull. "A statistical distribution function of wide applicability." In: *Journal of Applied Mechanics* 18.3 (1951), pp. 293–297.

[54] Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. "Deep Learning for Survival Analysis: A Review." In: *Artificial Intelligence Review* 57.3 (2024), p. 65. DOI: 10.1007/s10462-023-10681-3.

[55] Farhan Zafar, Md Monir Hossain, Yin Zhang, Alia Dani, Marc Schecter, Don Hayes, Maurizio Macaluso, Christopher Towe, and David L. S. Morales. "Lung Transplantation Advanced Prediction Tool: Determining Recipient's Outcome for a Certain Donor." In: *Transplantation* 106.10 (2022), e463–e472. DOI: 10.1097/TP.0000000000004131. URL: https://pubmed.ncbi.nlm.nih.gov/35389371/.