STATISTICAL DETECTION AND SURVIVAL ANALYSIS WITH APPLICATIONS IN SENSOR NETWORKS AND HEALTHCARE

A Dissertation
Presented to
The Academic Faculty

By

Xi He

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology

August 2020

STATISTICAL DETECTION AND SURVIVAL ANALYSIS WITH APPLICATIONS IN SENSOR NETWORKS AND HEALTHCARE

Approved by:

Dr. Yao Xie, Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Pinar Keskinocak, Co-Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Joel Sokol
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Kamran Paynabar
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Brian M. Gurbaxani
Office of Science
Centers for Disease Control and Prevention

Date Approved: June 30, 2020

All models are wrong, but some are useful.

George E. P. Box

Break the cycle, Morty. Rise above. Focus on science.

Rick Sanchez

To Yumi and Vinnie.

ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude towards my advisors Dr. Yao Xie and Dr. Pinar Keskinocak, without whom this thesis would not have been possible. I'm grateful for my committee members, Dr. Joel Sokol, Dr. Kamran Paynabar, and Dr. Brian Gurbaxani, for giving feedbacks on my research and the thesis. My thanks also go to Liyan Xie, Sin-Mei Wu, and Fan-Chi Lin, who worked in collaboration with me, and to Ethan Mark, Ana Maria Estrada Gomez, Kirthana Hampapur, Dr. Jefferson Jones, and Dr. David Goldsman, for inspirational discussions and suggestions on my research.

The studies in the thesis are partially supported by NSF CAREER CCF-1650913, the Carlos and Marguerite Mason Trust Foundation, and the Laura and John Arnold Foundation. The support is greatly appreciated.

Lastly, I would like to thank my friends and families. Your support and encouragement made it possible for me to complete this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

In this thesis, we present novel statistical methods for detecting abnormalities in a sequence of observations. We focus on two topics in statistics: change-point detection and survival analysis, and we demonstrate the application of our new methods in real data problems in the healthcare and the sensor network domains. We are particularly interested in cases in which the observations or predictors are related, and we summarize the relations graphically to develop new methodologies based on the graphs.

CHAPTER 1

INTRODUCTION AND BACKGROUND

The thesis consists of three major studies. The first is on sequential graph scan statistics in sensor networks. Graph change-point detection problems have wide applications in graphical data types, such as social networks and sensor networks. Given a sequence of random graphs with fixed vertices and changing edges, we are interested in detecting a change that causes a shift in the distribution of a subgraph. We present two graph scanning statistics that can detect local changes in the distribution of edges in a subset of the graph. The first statistic assumes a parametric model, i.e., the observations on the edges are Gaussian random variables, and the change shifts the mean of a subgraph. We derive the scan statistic and present a theoretical approximation to the false alarm rate, which is verified to be accurate numerically. The second statistic adopts a nonparametric approach based on k-Nearest Neighbors (k-NN). We demonstrate the efficiency of our detection statistics for ambient noise imaging, using a real dataset that records real-time seismic signals around the Old Faithful Geyser in the Yellowstone National Park.

The second is on the application of survival analysis in a healthcare problem. Survival prediction is key to making efficient organ allocation decisions and optimizing patient outcomes. In this paper, we develop a statistical machine learning model that accurately predicts the post-transplant survival curves for pediatric recipients of kidney transplantation. The prediction is made based on statistically selected risk factors. We develop a new predicting model with higher concordance index than the existing models.

The last is on graph based variable selection in survival analysis. Variable selection is a fundamental problem in survival analysis. When developing an accurate survival predicting model, identifying the proper variables to include in the model is often essential. In many applications, there exists an underlying graphical structure for the predictors. For example, some predictors may have strong correlations or interactions. When predicting the survival probability of a transplant recipient, it is important to consider the compatibility of the recipient and the organ donor. In such cases, incorporating the predictor graph into the penalty function for variable selection would allow more accurate inference and prediction. In this section, we propose to incorporate a fused lasso type of constraint in the Cox proportional hazard model, which takes advantage of the predictor graph generated by the relations among the predicting variables. We derive theoretical performance guarantees to the model and demonstrate the benefits of it using simulations and real data examples.

CHAPTER 2
SEQUENTIAL GRAPH SCANNING STATISTIC FOR CHANGE-POINT DETECTION

Change-point detection is a fundamental problem in social networks [1], sensor networks, and power networks. In this paper, we use graph scanning techniques [2], [3] to study the question of how to detect a change in the distribution of the graphs. In particular, we are interested in detecting a *local* change in the graph.

This means, when the change happens, only a subset of the graph, or a subgraph, of known size is affected by the change and acquires a different distribution. The observed change in distribution for the graphs are caused by a local change, while the distribution for the rest of the graph remains the same. The problem of local change-point detection is challenging in that, first, we do not know whether there is a change, and second, if there is a change at some unknown time, it is not clear which subgraph contains the change.

A motivating application of our study is monitoring ambient noises in seismic sensor networks. In ambient noise imaging, because the signals are weak, it is difficult to observe any signal using observations from a single sensor. Fortunately, when we construct the pairwise cross-correlation between the sensors, there will be coherent signals between affected sensors who observe changes in the subsurface structures. Specifically, at the time of the change, the cross-correlation function between the sensors affected by the change will have a significant peak. Between the affected sensors and the unaffected sensors, and among the unaffected sensors, such a waveform of the cross-correlation function does not exist. Therefore, this problem, mathematically, becomes detecting a local change in a sequence of graphs.

We present two approaches for constructing *scan statistics* to detect a local change in a sequence of graphs, the parametric and the non-parametric approach. For the parametric approach, we assume Gaussian graphs and apply a scan statistic based on counting the maximum number of edges in a subgraph of fixed size. We derive an accurate theoretical approximation to the false alarm rate of the scan statistic based on selective inference [4], which can be used to set the threshold for the false alarm rate without large scale simulation. For the non-parametric approach, the scan statistic is constructed using similarity measures on the subgraphs and k-Nearest Neighbors (k-NN). We demonstrate the efficiency of the non-parametric approach on real data for the seismic sensor network in Yellowstone [5].

This work is related to change-point detection, graph scan statistics, and community detection. Graph scan statistic for the stochastic block model, which counts the maximum number of edges in the subgraphs of an Erods-Renyi graph, has been considered in [6]. A likelihood ratio test for detecting communities in the Erdos-Renyi graph is studied in [7]. A non-parametric graph scan statistic based on k-NN is discussed in [8] and [9].

2.1 Problem Formulation

Suppose we observe a sequence of undirected graphs $G_1, \dots, G_N$, where N is the time horizon. For $t = 1, \dots, N$, let $G_t = \{V, E_t\}$, with V and E_t being the set of vertices and the set of edges respectively. Let V^i be a size-m subset of the nodes V, $i = 1, \dots, d$, where $d = \binom{N}{m}$ if all possible subsets are considered. In networks, usually $d \ll \binom{N}{m}$. Let $S^i = \{V^i, E^i\}$ be the subgraph containing V^i and the edges between them, which change over time. Denote $\mathcal{S}$ as the set of all possible subgraphs, then $\mathcal{S} = \{S^i, \dots, S^d\}$. Assume a change-point happening at an unknown time τ and the change is contained in the graph $S^* = \{V^*, E^*\}$, such that before and after τ, the distribution of the edges in E^* are different. At time t, denote $S^i(t) = \{V^i, E^i_t\} \subset G_t$.

When there is a change, we assume $E^*_1, \dots, E^*_{\tau-1}$ are i.i.d. distributed according to some distribution P, and $E^*_\tau \dots, E^*_T$ are i.i.d. distributed according to another distribution Q. The problem of detecting a local change becomes the following hypothesis testing problem.

$$
\begin{aligned}
H_0:&\quad E_t^i \sim P, \quad t = 1, \ldots, N, \ \forall\, S^i \in \mathcal{S}; \\
H_1:&\quad E_t^i \sim Q, \quad t \geq \tau, \ S^i = S^*, \\
&\quad E_t^i \sim P, \quad \text{otherwise}.
\end{aligned}
\tag{2.1}
$$

E_t^i is also the adjacency matrix of the subgraph S^i at time t. The hypothesis testing problem is illustrated in Fig. 2.1.

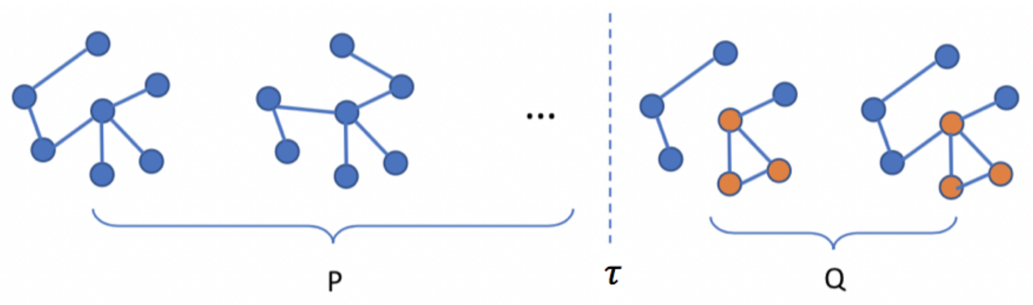


Figure 2.1: Graphs prior to the change-point in time τ follow the distribution P, and graphs after time τ follow the distribution Q. We are particularly interested in detecting the local change in a subgraph (showed in highlight).

Assuming that the change happens at τ, at each time t, for each subgraph S^i, we form a test statistic $R(t, \tau, S^i)$. The change is detected when the test statistic exceeds a given threshold γ. Let w be a small sliding window, the test scheme can be formulated as

$$
T = \inf\{t : \max_{t-w<\tau<t} \max_{S^i \in \mathcal{S}} R(t, \tau, S^i) > \gamma\}. \tag{2.2}
$$

We are further interested in knowing which subgraph causes the change in the graph structure. The test statistic $R(t, \tau, S^i)$ is useful in localizing the change, as the subgraph S^* that maximizes $R(t, \tau, S^i)$ is the subgraph containing the change,

$$
S^* = \arg\max_{S^i \in \mathcal{S}} R(t, \tau, S^i).
$$

We present two possible approaches to this problem based on *scan statistic* in the next sections, a parametric approach and a non-parametric approach. Moreover, we will study real data for this problem in the numerical example section.

2.2 Parametric Approach

First, we consider a parametric approach to form the scan statistic $R(t, \tau, S^i)$ in (2.2) by introducing a probability model to the sequence of graphs. In particular, we assume that the entries of the adjacency matrices are Gaussian random variables. Before the change, the edges have smaller means (e.g., zero mean) to represent that there is no significant correlation between the sensors. After the change, a subset of the nodes, i.e. sensors containing the change, will have higher means on the edges between them. For any subgraph $S^i \in \mathcal{S}$, at time t, let $W_{u,v}(t)$ denote the probability of the edge formation between the vertices u and v, where $u, v \in V^i$, then $E_t^i = \{W_{u,v}(t) : u, v \in V^i\}$. In this case, in the hypothesis testing problem (2.1), P represents $\mathcal{N}(\mu_0, \sigma_0^2)$, and Q represents $\mathcal{N}(\mu_1, \sigma_0^2)$, where μ_0, μ_1, σ_0^2 are constants, and $\mu_1 > \mu_0$. We can re-write (2.1) as

$$
\begin{aligned}
H_0 : \quad & W_{u,v}(t) \sim \mathcal{N}(\mu_0, \sigma_0^2), \quad && t = 1, \ldots, N, \ \forall\, u, v \in V; \\
H_1 : \quad & W_{u,v}(t) \sim \mathcal{N}(\mu_1, \sigma_0^2), \quad && t \geq \tau, \ \mu_1 > \mu_0, \ u, v \in S^*, \\
& W_{u,v}(t) \sim \mathcal{N}(\mu_0, \sigma_0^2), \quad && \text{otherwise.}
\end{aligned}
$$

In this section, we first set aside the time dimension and focus on detecting the subgraph S^* affected by the change. Once we formulate the subgraph detection scheme, we can repeatedly apply the test to the sequence of graphs as a Shewhart chart procedure.

Now we present the construction of the scan statistic in the parametric setting. Let x_i denote the number of edges in a subgraph S^i with m vertices. Then x_i follows a Gaussian distribution with mean μ_{x_i} and covariance Σ_{x_i}.

$$x_i = \sum_{u,v \in S^i} W_{u,v} \sim \mathcal{N}(\mu_i, \Sigma_i).$$

Under the null hypothesis,

$$\mu_i = \frac{m(m-1)}{2}\mu_0, \quad \Sigma_i = \frac{m(m-1)}{2}\sigma_0^2.$$

A change is detected when the maximum number of edges in a subgraph exceeds a pre-specified threshold γ, i.e.

$$\max_{S^i \in \mathcal{S}} x_i > \gamma.$$

We estimate the false alarm rate: $\mathbb{P}_0\{\max_{S^i \in \mathcal{S}} x_i > \gamma\}$. Recall $|\mathcal{S}| = d$. So the false alarm rate can also be written as

$$\mathbb{P}_0\{\max_{i=1,\ldots,d} x_i > \gamma\}. \tag{2.3}$$

2.2.1 Theoretical Threshold

We observe that (2.3) is the tail probability of the maximum of a series of correlated Gaussian random variables. In this section, we transform the false alarm rate formula using Bayes rule, and then apply the idea of selective inference [4] to estimate the probability.

Notice that we can decompose the event in (2.3) as the union of polyhedrons:

$$\begin{aligned}\left\{\max_{i=1,\ldots,d} x_i > \gamma\right\} &= \bigcup_{i=1,\ldots,d} \{x_i > \gamma, x_i \geq x_j, j \neq i\} \\ &\triangleq \bigcup_{i=1,\ldots,d} \{A_i \mathbf{x} \geq \mathbf{b}\},\end{aligned}$$

where $\mathbf{x} = [x_1, \ldots, x_d]^N \in \mathbb{R}^d$, $\mathbf{b} = [\gamma, 0, \ldots, 0]^N \in \mathbb{R}^d$, and $A_i = AP_i$. Here,

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & & & \ddots & \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix} \in \mathbb{R}^{d\times d},$$

and P_i is the permutation matrix swapping the 1^{st} and the i^{th} entry of $\mathbf{x}$. Similar decomposition appears in [10]. Thus,

$$\mathbb{P}_0\{\max_{i=1,\ldots,d} x_i > \gamma\} = \frac{\beta}{\alpha}, \tag{2.4}$$

where

$$\begin{aligned} \alpha &= \mathbb{P}_0\{x_1 > \gamma | \max_{i=1,\ldots,d} x_i > \gamma\} \\ &= \mathbb{P}_0\Big\{x_1 > \gamma | \bigcup_{i=1,\ldots,d} \{-A_i\mathbf{x} \leq -\mathbf{b}\}\Big\}, \\ \beta &= \mathbb{P}_0\{x_1 > \gamma\} \\ &= 1 - \Phi\Big(\gamma; \frac{m(m-1)}{2}\mu_0, \frac{m(m-1)}{2}\sigma_0^2\Big), \end{aligned}$$

where Φ is the CDF of the standard normal distribution, and α can be evaluated using selective inference as Theorem 5.3 in [4]. Our result is summarized in Lemma 2.2.1.

Lemma 2.2.1. *Let F^B_{μ,σ^2} denote the CDF of a normal random variable with mean μ and* *variance σ^2 truncated to the set B, and let $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$. Then*

$$\begin{aligned}
\boldsymbol{\mu} &= \frac{m(m-1)}{2}\mu_0 \mathbb{1}_d,\\
\Sigma_{(i,i)} &= \frac{m(m-1)}{2}\sigma_0^2,\\
\Sigma_{(i,i')} &= \frac{l_{i,i'}(l_{i,i'}-1)}{2}\sigma_0^2,\ i \neq i',
\end{aligned}$$

where $\mathbb{1}_d$ is the d-dimensional vector of all 1's, and $l_{i,i'}$ is the number of overlapping nodes between two subgraphs S^i and $S^{i'}$. Then we have the following conclusion.

$$F^{\bigcup_i [\mathcal{V}_i^-(\mathbf{z}),\mathcal{V}_i^+(\mathbf{z})]}_{\boldsymbol{\eta}^N\boldsymbol{\mu},\boldsymbol{\eta}^N\Sigma\boldsymbol{\eta}}(\boldsymbol{\eta}^N\mathbf{x}) | \bigcup_{i=1,\dots,d}\{-A_i\mathbf{x} \leq -\mathbf{b}\} \sim \textit{Unif}(0,1)$$

with the specification $\boldsymbol{\eta} = [1, 0, \dots, 0]^N \in \mathbb{R}^d$, and the set boundaries

$$\begin{aligned}
\mathcal{V}_i^-(\mathbf{z}) &\equiv \max_{j:(A_i\mathbf{c})_j>0} \frac{b_j - (A_i\mathbf{z})_j}{(A_i\mathbf{c})_j},\\
\mathcal{V}_i^+(\mathbf{z}) &\equiv \min_{j:(A_i\mathbf{c})_j<0} \frac{b_j - (A_i\mathbf{z})_j}{(A_i\mathbf{c})_j},
\end{aligned}$$

where

$$\boldsymbol{c} \equiv \Sigma\boldsymbol{\eta}(\boldsymbol{\eta}^N\Sigma\boldsymbol{\eta})^{-1} = \Sigma\boldsymbol{\eta}\Sigma_{1,1}^{-1} = a\Sigma_{(:,1)},$$

$$\begin{aligned}
\mathbf{z} &\equiv (\boldsymbol{I}_d - \boldsymbol{c}\boldsymbol{\eta}^N)\mathbf{x} = \mathbf{x} - \boldsymbol{c}\boldsymbol{\eta}^N\mathbf{x} = \mathbf{x} - a\Sigma_{(:,1)}x_1,\\
a &= \frac{2}{k(k-1)\sigma_0^2}.
\end{aligned}$$

For $i \neq i'$,

$$\begin{aligned}
A_i\mathbf{c} &= AP_i\mathbf{c} = a\begin{pmatrix} \Sigma_{(i,1)} \\ \Sigma_{(i,1)} - \Sigma_{(i',1)} \end{pmatrix}, \\
A_i\mathbf{z} &= AP_i\mathbf{z} = \begin{pmatrix} x_i - a\Sigma_{(i,1)}x_i \\ \left(x_i - a\Sigma_{(i,1)}x_i\right) - \left(x_{i'} - a\Sigma_{(i',1)}x_i\right) \end{pmatrix}, \\
\mathbf{b} - A_i\mathbf{z} &= \begin{pmatrix} \gamma - \left(x_i - a\Sigma_{(i,1)}x_i\right) \\ \left(x_{i'} - a\Sigma_{(i',1)}x_i\right) - \left(x_i - a\Sigma_{(i,1)}x_i\right) \end{pmatrix}.
\end{aligned}$$

Therefore, we can estimate the false alarm rate using Lemma 2.2.1 and (2.4) and set the threshold γ accordingly. The performance of the estimation is presented in the next section.

2.2.2 Numerical Verification

In this section, we conduct a numerical experiment to verify the numerical accuracy of our estimation of the false alarm rate. Assuming standard normal distribution under the null hypothesis, we generate α according to Lemma 2.2.1 and compute the false alarm rate based on (2.4). The resulting false alarm rate curve by changing the threshold γ is plotted in Fig. 2.2. The result is based on 500 experiments, and the standard error, which is small, is shown as the shaded area in the plot.

We also compare the theoretically estimated γ from using formula (2.4) with the simulated γ in Table 2.1. In this experiment, $N = 50, m = 5$, and $d = 2, 118, 760$. The two γ's are quite close in this case, showing good approximation of the theoretical result.

Table 2.1: Theoretically proven γ and simulated γ threshold values under different false-alarm probabilities.

Probability	Theory γ	Simulated γ
0.2	13.34	14.43
0.15	14.77	14.68
0.1	16.00	15.93

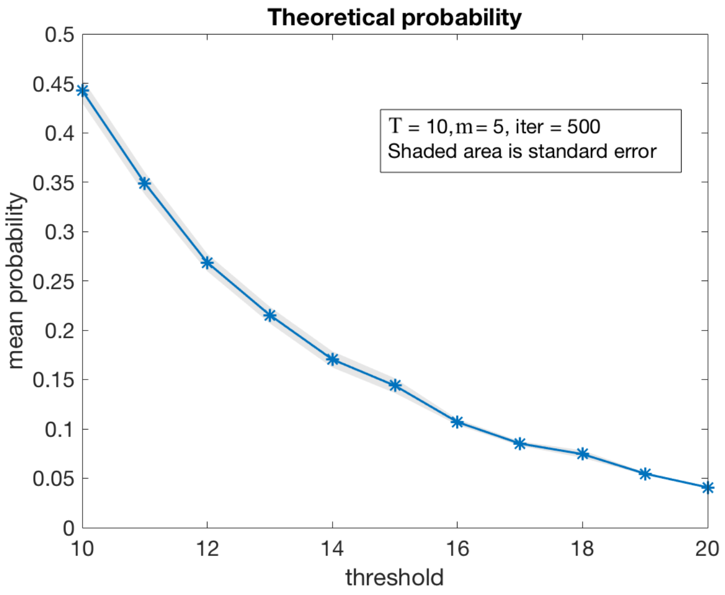


Figure 2.2: Theoretical false-alarm rates of the detection statistic by equation (2.4).

2.3 Non-parametric Approach based on Similarity

In this section, we describe a non-parametric detection statistic based on the similarity measure between subgraphs at different time. The idea is to compare the subgraphs formed with the same set of nodes occurring before and after time t to check for their graph structure similarity. If the graph structures are similar, they are likely from the same distribution, and if if the dissimilarity is large enough, we declare a change-point at t. For $i = 1, \ldots, d$, at time $t = 1, \ldots, N$, we check the similarity between $S^i(1), \ldots, S^i(t-1)$ and $S^i(t), \ldots, S^i(N)$. For simplicity, denote an arbitrary subgraph S^i as S in the rest of the analysis.

H_0 is rejected when $R(t, \tau, S)$ is significantly *smaller* than its expectation under the permutation null distribution. When $R(t, \tau, S)$ is small, it means that the number of edges connecting the two groups in the k-NN graph is small, and the two samples are likely from different distributions. If $R(t, \tau, S)$ is large, it implies that the samples are well-mixed and are likely to be from the same distribution.

It is shown in [11] and [12] that the standardized test statistic

$$\frac{R(t,\tau,S)-\mathbb{E}[R(t,\tau,S)]}{\sqrt{Var(R(t,\tau,S)}}$$

converges to the standard normal distribution under H_0 when $\frac{t}{N-t} \to \lambda \in (0,\infty)$ for multivariate data. The mean and variance for the statistics are

$$\begin{aligned}
&\mathbb{E}[R(t,\tau,S)] = \frac{4kt(N-t)}{N-1},\\
&Var(R(t,\tau,S)) = \frac{4kt(N-t)}{N-1}\Big(h\big(t,(N-t)\big)\\
&(\frac{1}{N}\sum_{n,n'=1}^{N} A^{+}_{n,n'}A^{+}_{n',n} + k - \frac{2k^2}{N-1})\\
&+\big(1-h(t,N-t)\big) + (\frac{1}{N}\sum_{n,n',n''=1}^{N} A^{+}_{n',n}A^{+}_{n'',n} - k^2)\Big),
\end{aligned}$$

where $h(t,N-t) = \frac{4(t-1)(N-t-1)}{(N-2)(N-3)}$.

Define the test statistic

$$R'(t,\tau,S) = -\frac{R(t,\tau,S)-\mathbb{E}[R(t,\tau,S)]}{\sqrt{Var(R(t,\tau,S)}}.$$

Suppose the change occurs at time τ, then $R'(t,\tau,S)$ will be large when t is close to τ (note the negative sign in the standardization).

The testing procedure can be written as

$$T(t,\tau,S) = \inf\{t : \max_{S^i\in\mathcal{S}} \max_{n_0\le t\le N-n_0} R'(t,\tau,S^i) > \gamma\}, \qquad (2.5)$$

where $1 < n_0 < N$.

2.4 Real-data Example

In this section, we demonstrate how the proposed detection statistics could be used in solving the local change-point detection problem in a seismic sensor networks using real data. We first check whether there is a change in the graphs, and then narrow down the change to a subgraph. For simplicity, we only apply the nonparametric approach.

2.4.1 The Seismic Data

The seismic sensor network that we study is illustrated in Fig. 2.3. It shows the physical location of the sensors measuring signals around the Old Faithful Geyser in the Yellowstone National Park. There are 18 sensors in the network, and edge information is contained in the pair-wise cross-correlation function between the sensors. The cross-correlation function is then transformed to a value called peak lag time, which is shown on the y-axis in Fig. 2.3. We observe a sequence of 101 graphs on this network over time, one at each "stage", ranging from stage -50 to 50 (shown as the x-axis). The nodes, or sensors, in the networks remain the same, while the edge value fluctuates as the peak lag time among the sensors changes. At stage 0, the geyser erupts, and the distribution of the peak lag time among the sensors affected by the eruption changes. Our goal is to detect the change in the sequence of the graphs at stage 0 and find the sensors responsible for the change. We have data on 11 stations: 001, 002, 003, 005, 006, 008, 009, 010, 014, 015, 016, and the peak lag time on 10 pairs of the stations. For the other 45 pairs without data, we assume that no edge forms between the sensors.

2.4.2 Change-point Detection

First we detect whether there is a change-point in the sequence of graphs. Two types of graphs are considered, the unweighted graph and the weighted graph.

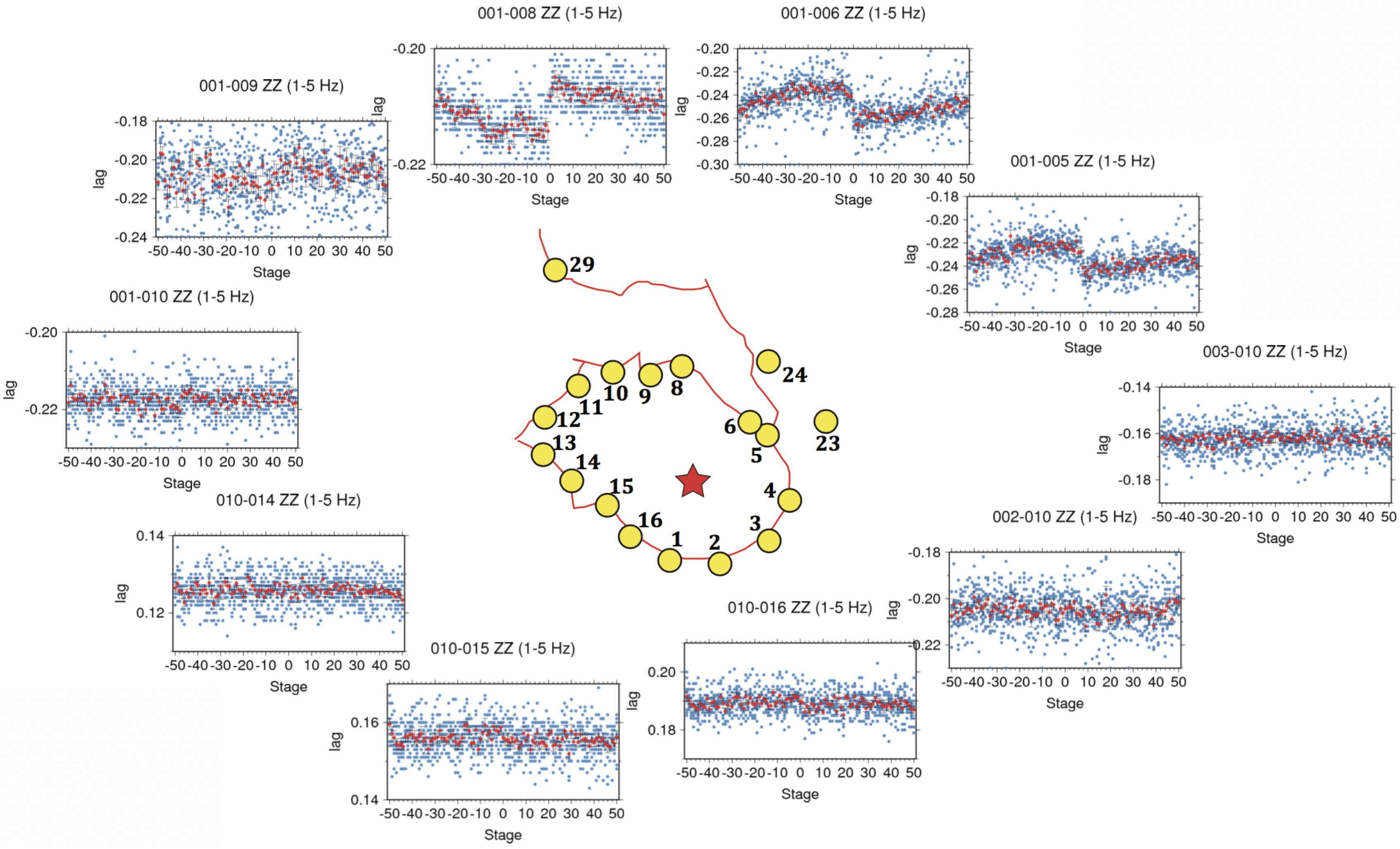


Figure 2.3: Peak lag time in the seismic sensor network which measures the geyser activity in the Yellowstone National Park.

2.4.2 *Unweighted Graph*

Denote the *mean* peak lag time (red points in Fig. 2.3) of a pair of sensors u, v at time t as $y_{u,v}(t)$. Assume that an edge forms between u, v at time t if $y_{u,v}(t)$ is greater than the average $\bar{y}_{u,v}$, that is, $y_{u,v}(t) > \bar{y}_{u,v}$, where $\bar{y}_{u,v} = \frac{1}{101}\sum_{t=1}^{101} y_{u,v}(t)$. We use the Weisfeiler-Lehman edge graph kernel [13] to measure the closeness of the graphs and find the k-NN as described in the non-parametric section. The test statistic $-R(t, \tau, S)$ is plotted in Fig. 2.4, and it peaks at stage 0, corresponding to the true change-point.

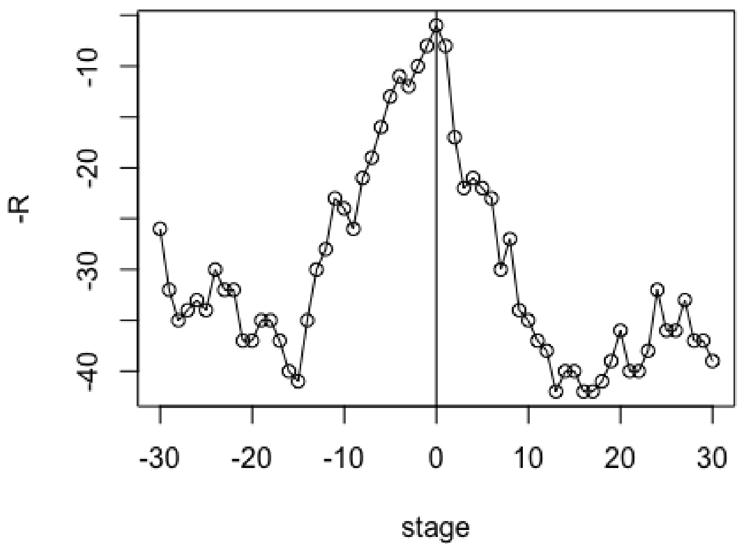
2.4.2 *Weighted Graph*

To construct weighted graphs, at each time t, we use the peak lag time between the two stations u, v as the "weight" on the edge between the nodes.

The test statistic $-R(t, \tau, S)$ is plotted in Fig. 2.4. Comparing with the previous experiment on unweighted graphs, we find that although both methods successfully identifies the change-point at stage 0, there are also two other local maxima for the weighted graph, which may interfere with the detection.

2.4.3 Change Location Detection

We are further interested in finding the location within the graph where the change happens. In other words, we identify a subset of m nodes that contribute to the overall change in the graphs. Ideally, the data on those nodes would be sufficient for the overall change detection. In this example, we assume $m = 3$ by observing Fig. 2.3. We have data for 11 nodes, and therefore $\binom{11}{3} = 165$ possible subsets of nodes. However, recall that only 10 edges are available. So in reality, only 56 subsets are considered. Given each subset of nodes, we preserve the edge information among the 3 nodes, and set the weight on other edges to 0. For each subset, we repeat the steps in the last example on weighted graphs as if the graphs only contain 3 nodes in the subset. Following the testing procedure in (2.5), we find that the subgraph maximizing the test statistic is formed by nodes 001, 008, and 009.



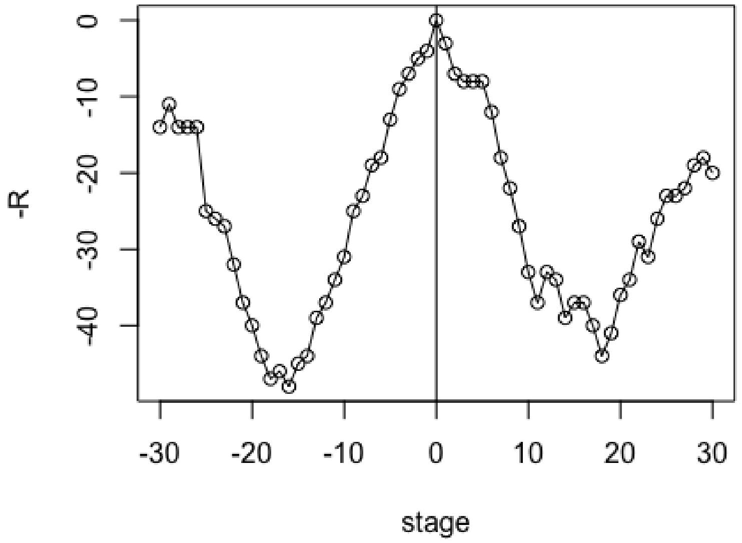


Figure 2.4: Test statistic for: Unweighted graph (top), weighted graph (bottom).

CHAPTER 3
PEDIATRIC KIDNEY TRANSPLANT SURVIVAL ANALYSIS USING STATISTICAL MACHINE LEARNING

3.1 Introduction

One of the greatest challenges of organ transplantation in the U.S. is the widening gap between the supply and demand for organs. On any given day, around 75,000 patients are on the waiting list for organs, but each year, only 34,000 organs are recovered [14]. The question of how to allocate the limited organs becomes critical and challenging. When matching an available organ to a potential recipient, the recipient's post-transplant survival estimate is one of the major considerations. In this paper, we analyze historical records of pediatric kidney transplantation in the U.S. to develop a statistical machine learning model that can (1) accurately predict the post-transplant survival curves for pediatric kidney transplant recipients and (2) identify the most important risk factors influencing the survival curves.

We focus on *pediatric* (age 0-17) kidney transplant recipients since most post-transplant survival prediction models are developed for adult kidney transplant recipients and do not always perform well on pediatric recipients. Pediatric and adult kidney transplant recipients have distinct physiological conditions, and their survival curves are different (Figure 3.1). Our survival prediction model for pediatric kidney transplant recipients shows better performance than a recent state-of-art model developed for transplant recipients of all ages [16].

To the best of our knowledge, there is no post-transplant survival prediction model for pediatric transplant recipients. Most existing studies on pediatric transplantation are retrospective and review the overall trends in the post-transplant survival probabilities for

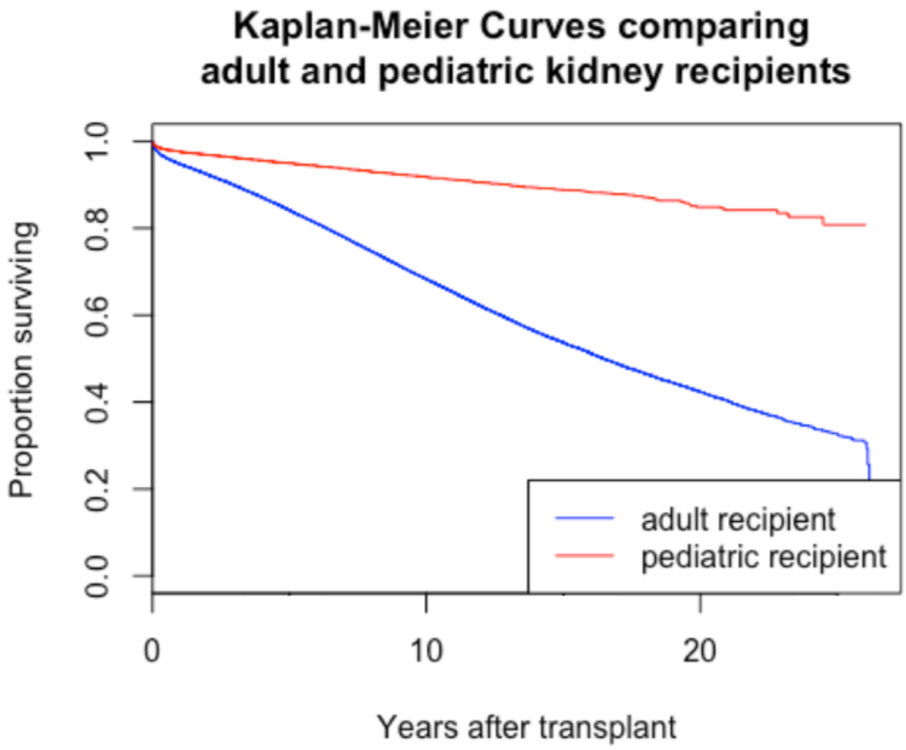


Figure 3.1: The Kaplan-Meier survival curves [15] for pediatric (age 0-17) and adult (age 18 and above) recipients of kidney transplantation are different.

the pediatric transplant recipients [17, 18, 19, 20, 21, 22, 23]. For example, it is shown that the survival probabilities for pediatric kidney transplant recipients improved from 1989 to 2014 [23]. To complement the existing literature, we develop a model to estimate the post-transplant survival curves for a specific recipient and donor pair. Similar models have been developed for transplant recipients of all ages [16] and for pediatric recipients of "increased risk" organs [24], but a model dedicated to the pediatric recipients of a *general* kidney, to be best of our knowledge, is unavailable yet.

In comparison to existing studies on pediatric transplantation, we adopt a systematic statistical variable selection method to identify the most significant risk factors influencing the pediatric post-transplant survival curves. In many existing studies, medical domain knowledge is used to select the risk factors [25, 26, 27, 28, 29]. There are also studies which evaluate the effect of a specific risk factor on the pediatric post-transplant survival rates, including but are not limited to donor age [30], HLA-DR match [31], polyomavirus nephropathy [32], and the obesity status of the transplant recipient [33]. In this paper, we combine systematic statistical variable selection techniques with medical domain knowledge to identify the most critical risk factors influencing the post-transplant survival curves for pediatric kidney transplant recipients.

While the current study focuses on kidney transplantation since kidney is one of the most commonly transplanted organs in the U.S., the methodology we develop is general and is applicable to other organs as well.

3.2 Methods

In this section, we describe the data, the data preparation process, and the survival analysis and variable selection methods used to build our survival prediction models for pediatric recipients of kidney transplantation.

In developing the methodology, we divide our data by donor type (deceased or living) and pre-process the data by removing transplant cases with heterogeneous survival distribution, imputing missing values, and feature engineering the variables. We then experiment with the classical Cox proportional hazard model [34] and the machine learning based random survival forest (RSF) model [35] and choose the one with higher prediction accuracy for each donor type.

3.2.1 The Pediatric Kidney Transplant Data

We use a dataset from UNOS (United Network for Organ Sharing), which contains 19,236 pediatric (age 0 - 17) kidney transplant cases in the U.S. from 1987 to 2014. For each transplant case, 487 features concerning the transplant recipient, the donor, and the procedure were recorded. One of the greatest challenges of survival prediction for pediatric kidney transplant recipients is the high censoring rate in pediatric datasets. In the UNOS dataset, 94.30% of the pediatric data are censored, compared with 79.17% for the general transplant data. Censoring means that when the data are collected, the event (in this case, death of the transplant recipient) does not happen during the observation time frame, and only the person's latest follow-up time is recorded. In statistics, when calculating the survival rate, the censored people are considered as not experiencing the event. Hence, the empirical survival rates of pediatric kidney transplant recipients are much higher than adult recipients (see Figure 3.1). With a limited number of death cases in the pediatric data, it is challenging to characterize the features of pediatric transplant cases with a high risk. It is equivalent to the unbalanced data issue in a machine learning classification problem. Therefore, it is more difficult to develop an accurate survival prediction model for pediatric kidney transplant recipients than for the general recipients. Nevertheless, our proposed model still achieves higher prediction accuracy than the existing models.

3.2.2 Data Preparation

We pre-process the UNOS dataset for the survival model by following three steps: (1) divide the dataset into two by donor type, and for each donor type, (2) apply a change-point detection method to ensure data homogeneity, and (3) conduct feature engineering and missing value imputation.

Divide the dataset into two by donor type, i.e. deceased or living

We study the two donor types separately and build a unique survival prediction model for each donor type for two reasons. Firstly, past studies [26, 36] found that the donor type is an impactful feature for the survival rates of transplant recipients. Receiving the organ from a living donor generally results in higher survival rates than from a deceased donor. To ensure that the prior finding holds in our pediatric kidney dataset, we use the log-rank test[37] to compare the 5-year survival curves for recipients of each donor type. The log-rank test is a statistical test used to compare the survival distributions of two samples. When the test gives a p-value lower than a significance level, often set to be 0.05, it indicates a significant difference between the two samples. The resulting p-value is 2.0×10^{-15}, which confirms a significant difference between the post-transplant survival curves for pediatric kidney recipients of the two donor types.

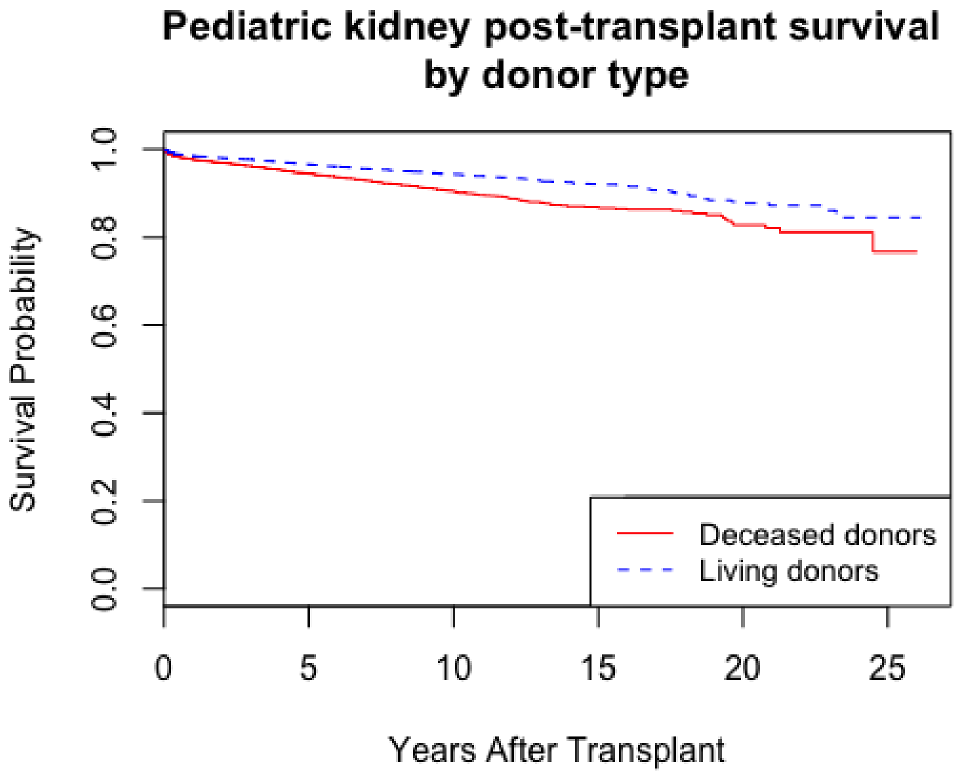


Figure 3.2: The post-transplant survival curves by the donor type are different, and we need a separate survival prediction model for each donor type.

Secondly, dividing the data by donor type and building a designated survival model for each donor type allow us to include donor type specific variables in the survival models. It also avoids imputing variable entries that are not missing but not applicable due to the donor type. For example, the transplant cases with deceased donors contain a variable called DON_MECH_DEATH (donor mechanism of death), which does not apply to the living donors. Similarly, transplant cases with living donors contain a variable called LIV_DON_TY (living donor type), which describes the living donor's relation to the recipient. The transplant cases with deceased donors also have this feature, but the entries are left blank.

After dividing the dataset by donor type, we have 9,927 transplant cases for the deceased donors, and 8,852 cases for the living donors. For each donor type, we conduct further data processing steps, which include using change-point detection to tackle data heterogeneity, conducting feature engineering, and imputing missing values.

3.3 Change-point Detection to tackle Data Heterogeneity

A significant feature of our dataset is heterogeneity due to the wide range of time (from 1987 to 2014) the data were collected and the distinct physiological complexities of transplant recipients and donors at different life stages (see Figure A.2 and Figure A.4). To ensure that we fit a survival model using "homogeneous" data, i.e. data with similar statistical property, we perform change-point detection to partition the data. In particular, we use the machine-learning-based decision tree [38] and the statistical log-rank test [37] for detecting change-points in the transplant year, the recipient age, and the donor age. The decision tree partitions the data into groups with similar survival time by finding the optimal splits in the range of the variables (i.e. transplant year, recipient age, and donor age). The data are split iteratively and organized in a tree structure until the data in each terminal tree node have similar survival time. We use the first three split values as change-points. To use the log-rank test for change-point detection, we scan through values in the range of the variables (i.e. transplant year, recipient age, and donor age) and apply the log-rank test repeatedly. The change-points found by the log-rank test are values at which the test result has a p-value lower than the significance level. In cases where the log-rank test generates significantly low p-values for a sequence of split values (as in Figure 3.4), we focus on the first and last split values in the sequence. If the terminal split value in the sequence has a neighbor whose p-value is higher than the significance level, it represents a change in the pattern of the test results, and we declare the terminal split value as a change-point.

Using the change-point detection methods, we find multiple change-points in the transplant year, the recipient age, and the donor age that impact the survival curves of the pediatric recipients of kidney transplantation. The change-points are summarized in Table

3.1. Figure 3.4 shows plots of the p-values from the log-rank tests, and Table 3.2 shows the specific p-values in Figure 3.4. Multiple change-points are detected by the decision tree and the log-rank test (Table 3.1). We choose the change-points by which the majority of the data are maintained after partitioning (see Figure 3.3), and round the change-points to the nearest integers in accordance with the conventions in the datasets. As a result, for the deceased donor dataset, we keep transplant cases with the transplant year between 1991 and 2010, recipient age between 2 and 17, and donor age between 17 and 35 (all inclusive). For the living donor dataset, we keep transplant cases with the transplant year between 2003 and 2008, recipient age between 11 and 17, and donor age between 23 and 45 (all inclusive). We assume the data for each donor type now follow the same survival distribution and can be characterized by a single survival prediction model. The data would also be randomly sampled for training and testing purposes in the model evaluation and variable selection process.

For the dataset of each donor type, we remove variables with more than 80% missing entries to ensure that the variables we consider are commonly recorded in practice. We also remove the categories of categorical variables that have fewer than 10 samples as there are insufficient data to model the impact of these categories on the survival outcome.

Missing values in the datasets are imputed using the kNNHDI algorithm [39]. The algorithm finds the closest resemblance to a transplant case with missing variable entries and imputes the missing values using its k Nearest Neighbors (kNN). The kNNHDI algorithm has the advantage of not assuming variable independence. Algorithms such as MICE [40] impute a variable with missing entries by regressing the variable on the other variables and assume independence among the regressors. In our dataset, however, the assumption of variable independence does not hold. We tune the parameters (the weight and the number of nearest neighbors) in the kNNHDI algorithm to minimize the validation error.

After data pre-processing, we have 3919 pediatric kidney transplant cases and 84 features in the deceased donor dataset, and 5444 pediatric kidney transplant cases and 40

Table 3.1: Change-points for the survival distribution of the dataset for each donor type using the decision tree and the log-rank test. The change-points from the decision tree method are the first three values used as tree node splits ordered in sequence. The log-rank test change-points are the split points (ordered ascendingly) where the p-value of the test result is at or below the 0.05 significance level.

Donor type	**Variable**	**Decision tree change-points**	**Log-rank test change-points**	**Change-points we use**
Deceased	TX_YR (transplant year)	2004.5, 2010.5, 1990.5	2012	1991, 2010
	AGE (recipient age)	1.5, 14.5, 7.5	8	2
	AGE_DON (donor age)	7.5, 35.5, 16.5	2, 9, 17, 35, 36, 41, 42, 54	17, 35
Living	TX_YR (transplant year)	2002.5, 2008.5, 1997.5	1988, 1990, 1992, 2010	2003, 2008
	AGE (recipient age)	1.5, 10.5, 4.5	1, 2	11
	AGE_DON (donor age)	22.5, 60.5, 45.5	17, 22, 60	23, 45

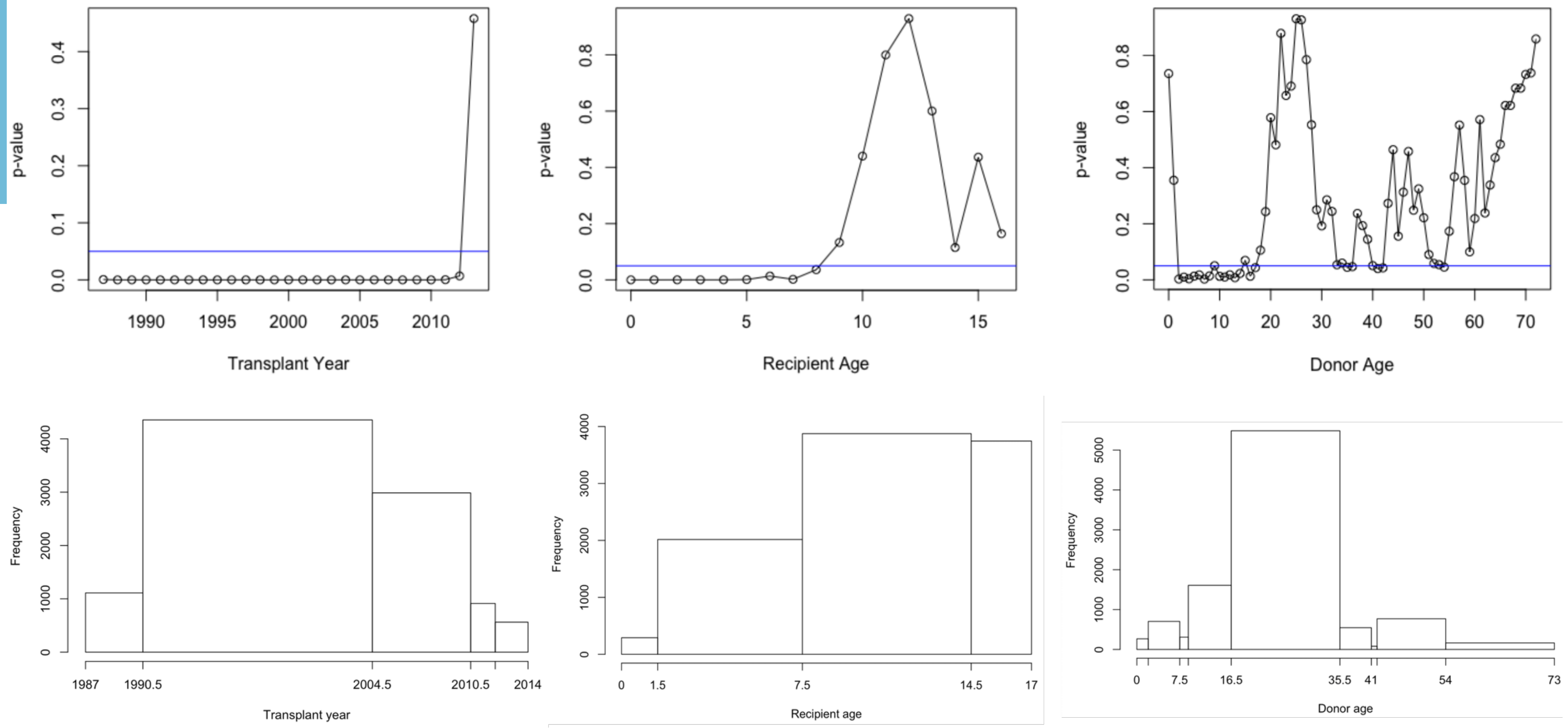


Figure 3.3: Deceased donors change detection. Top: p-values of the sequential log-rank test for change-point detection. The blue horizontal line is the reference for the 0.05 significant level. A p-value below the reference level indicates that the two samples split at the point have significantly different survival distributions. Bottom: number of transplant cases partitioned by the change-points. We choose the change-points that maximize the partitioned data to ensure data homogeneity and obtaining the maximum sample size.

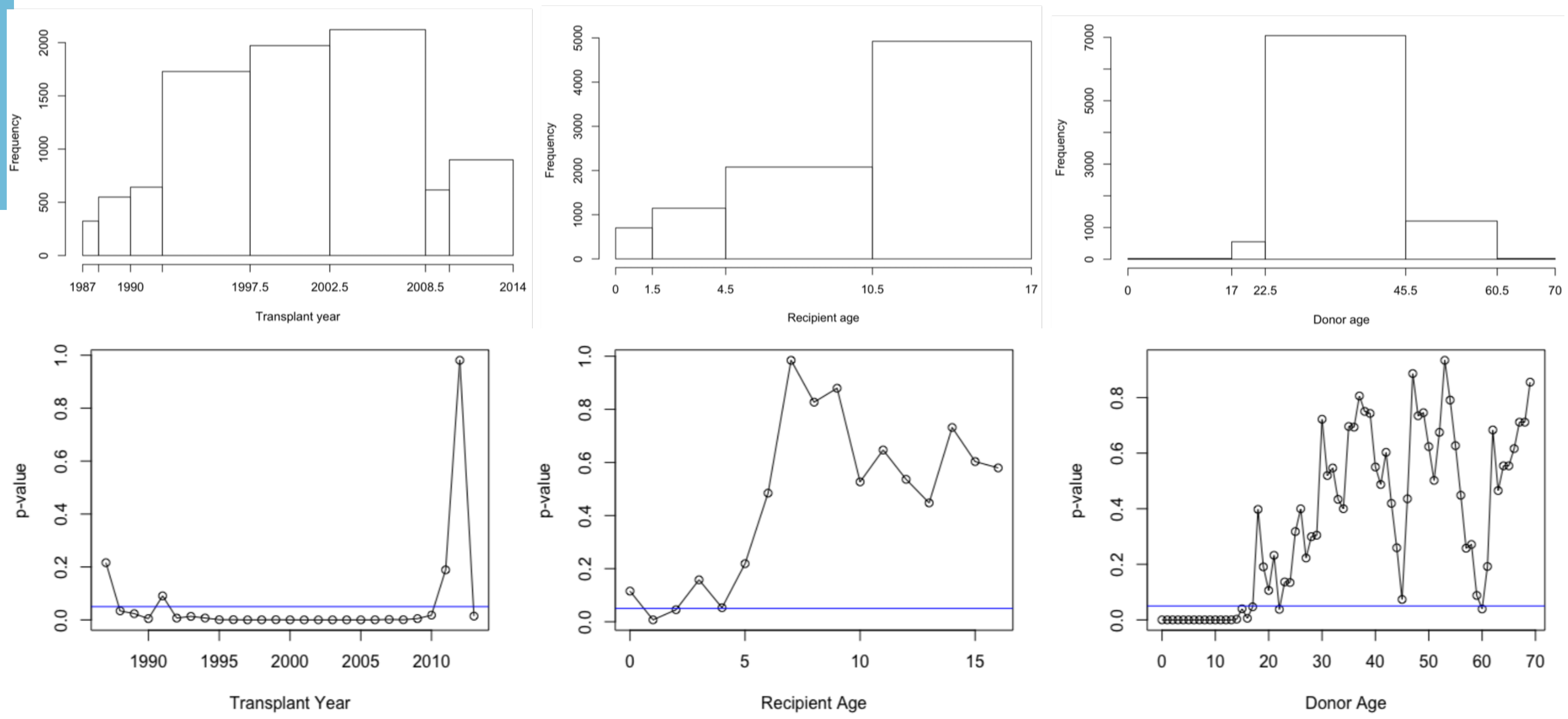


Figure 3.4: Living donors change detection. Top: p-values of the sequential log-rank test for change-point detection. The blue horizontal line is the reference for the 0.05 significant level. A p-value below the reference level indicates that the two samples split at the point have significantly different survival distributions. Bottom: number of transplant cases partitioned by the change-points. We choose the change-points that maximize the partitioned data to ensure data homogeneity and obtaining the maximum sample size.

Table 3.2: Specific p-values around the borderline of the 0.05 significance level in the log-rank test in Figure 3.4. **Bolded** are the split points where the corresponding p-values are at or lower than the 0.05 significant level, and their p-values. The **bold** split values are the change-points determined by the sequential log-rank test.

Donor type	Variable	Split value	p-value
Deceased donors	Recipient age	**8**	**0.036**
	Donor age	**9**	**0.050**
		17	**0.043**
		33	0.054
		34	0.060
		35	**0.045**
		36	0.047
		40	0.051
		41	**0.040**
		42	**0.043**
		52	0.059
		53	0.054
		54	**0.045**
Living donors	Recipient age	**2**	**0.045**
		4	0.053
	Donor age	**22**	**0.038**
		60	**0.039**

features in the living donor dataset. The specific features for the deceased and the living donors are in Table **??** and Table **??** in the Appendix respectively.

3.4 Survival Prediction and Variable Selection

We consider two types of survival prediction models for predicting the post-transplant survival curves for the pediatric kidney transplant recipients, and we select the one with better performance for each donor type. The two models are the statistical Cox proportional hazard model [34] and the machine learning based random survival forest (RSF) model [35]. The Cox model and its variants have been widely used in existing studies as a classical approach to solve survival prediction problems, while the RSF model is more recently developed and is not yet extensively used in practice. For each donor type (deceased or living), we compare the performance of the two models by their out-of-sample prediction accuracy metrics and select the model with better performance.

Specifically, in the RSF model, we use 1000 trees and restrict the average terminal node size to be 3 to balance the model performance and the computation speed.

We combine statistical methods and medical knowledge to select important features, or risk factors, in the Cox model and the RSF model. For each survival prediction model, we first use a statistical variable selection (also called feature selection or model selection) method suitable for the model to identify the most important risk factors. Hence, the variables selected in the Cox model and the RSF model are not necessarily identical. The variable selection procedure is performed using 5-fold cross-validation. For the dataset of each donor type, we randomly partition the data into five-folds, train the model on four folds of the data, and evaluate the model performance on the remaining one fold. The process is repeated until each fold of the data is being tested on once. The process ensures that the training and the testing data are 80% and 20% of the entire dataset in every training and testing run. The partition is valid based on the assumption that the data are from a homogenous survival distribution, which is ensured by the data pre-processing change-point detection step. We then use medical knowledge to determine the relevancy of the selected variables to our problem and finalize the features we include in the models.

In the Cox model, we use the group lasso penalty [41], since the majority of the variables (70% in the deceased donor dataset and 75% in the living donor dataset) are categorical. The group lasso penalty allows selecting all categories of a categorical variable at a time, while most other widely used variable selection methods, such as the regular lasso and elastic net, select certain categories of the categorical variables. In addition, methods such as stepwise variable selection are not efficient, given the large number of variables in the dataset. By adjusting the hyper-parameter in the penalty term, we determine the variables maximizing the out-of-sample c-index in each cross-validation trial. We take the union of the variables selected in each trial and use their medical interpretations to decide whether to include the variable in the final Cox model. We evaluate the performance of the proposed final Cox model by randomly sampling 80% of the data as training data and 20% of the data as testing data and calculating the average performance metrics of 10 such repetitions.

The variables in the RSF model are selected using the variable permutation importance (VIMP) score [35], which measures the contribution of a variable to the RSF model's out-of-sample prediction accuracy. We compute the VIMP score for every variable in each cross-validation trial, then rank the variables by their mean VIMP score of all trials (see Figure * for the deceased donors and Figure **??** for the living donors). We determine the number of variables to include in the final RSF model by experimenting with a different number of variables and checking the corresponding model performance. We start with the most important variable (the one with the highest VIMP score) and add variables one at a time in the order of their average VIMP score into the model until the model's cross-validated out-of-sample c-index start to decrease.

To evaluate the performance of our survival prediction models, we compare their performance with two other models from the literature. The first model is the EPTS (Estimated

Post Transplant Survival) model, which is the in-use model for determining the priority of adult transplant recipients on the kidney transplant waiting list. The model uses four features (recipient age, recipient diabetes status, recipient previous transplant yes/no, and recipient number of years on dialysis) as well as the transformation and interaction of these features to predict the post-transplant survival probabilities for adult kidney transplant recipients. An equivalence for the pediatric kidney transplant recipients is not established in practice, and pediatric recipients have a higher priority than adult recipients on the waiting list. The second model used for comparison is a recently developed state-of-art survival prediction model for kidney transplant recipients of all ages based on gradient-boosted trees [16].

We use two metrics for evaluating the performance of the survival prediction models: Harrell's concordance index [42] (c-index, or concordance index) and the 5-year integrated Brier score [43]. The two metrics are standard metrics commonly used in survival analysis. The c-index measures the concordance between the predicted and the observed survival time for all pairs of transplant recipients. It has a similar interpretation to the AUC (Area Under Curve) [44]: a value of 0.5 means randomly guessing which one of the pairs of the recipient has longer survival time, and a value of 1 means perfect prediction. Unlike the AUC, the c-index is considered satisfactory if between 0.6 and 0.7 [45]. The c-index is useful for comparing the survival time for recipients when an organ becomes available. The integrated Brier score is the squared error of probabilistic predictions integrated throughout the prediction horizon.

3.5 Results

The performance of our model and other models from the literature is summarized in Table 3.3 for the deceased donors and Table 3.6 for the living donors. The other models are the EPTS (Estimated Post Transplant Survival) model [46] (the in-use model for ranking the adult kidney transplant recipients on the waiting list) and a gradient-boosted tree model

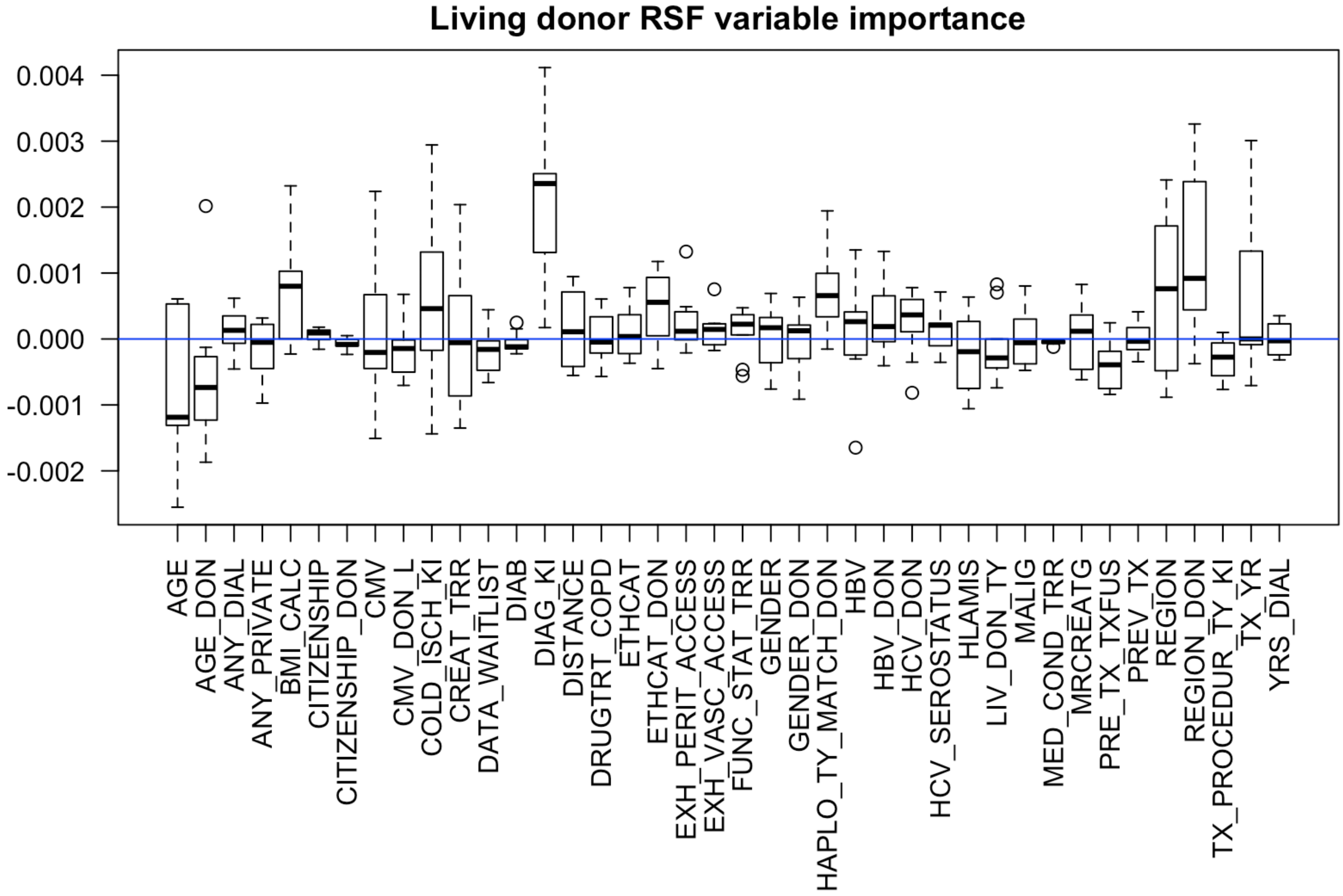


Figure 3.5: Variable importance values for variables using the RSF in the living donor dataset.

developed for kidney transplant recipients of all ages [16]. For both donor types, our model has improved out-of-sample concordance index and similar 5-year integrated Brier scores compared to the other models.

The features selected by each survival prediction model to be influential for the post-transplant survival results are shown in Table 3.5 for the deceased donors and Table 3.8 for the living donors. In the tables, we highlight in **bold** features that are, according to our models, particularly important for pediatric kidney transplant recipients and not for recipients of other age groups.

The Cox survival prediction model further identifies statistically significant features (with a p-value at or less than 0.05), which we include in Table 3.4 for the deceased donors and Table 3.7 for the living donors. The coefficients fitted by the Cox model are used to compute the hazard ratio of the feature, with a hazard ratio greater than 1 indicating a higher post-transplant risk as the feature value increases, if the feature is numeric. If the feature is categorical, a hazard ratio larger than 1 means that a pediatric kidney transplant recipient who falls under the feature category is predicted to have higher post-transplant risk than those under the feature's baseline category.

and the interpretation of selected variable coefficients in our Cox model are shown in Table 3.4. We find that in comparison with the EPTS and the general population model, the recipient current Panel Reactive Antibodies (PRA) level, recipient gender, donor age, donor ethnicity, transplant year, and recipient previous malignancy are specifically influencing the post-transplant survival curves for the pediatric kidney recipients. We observe that, contrary to intuition, insurance type (private vs. public) is not a significant risk factor for the post-transplant survival curves in either our Cox model or the RSF model for the pediatric kidney transplant recipients.

The Harrell's concordance index shows that both the Cox model and the RSF model we develop have a mean improvement of 0.09 from the EPTS model and 0.02 from the general population model. The average 3 year integrated Brier score for the models are on the same

Table 3.3: Deceased donor model performance comparison. The metrics reported are from 5 fold cross validation.

Performance measure	**Proposed Cox**	**Proposed RSF**	EPTS model (in-use)	All-age RSF[16]
c-index	**0.57**	**0.57**	0.51	0.53
5 year Brier score	0.036	0.036	0.036	0.036

Table 3.4: Deceased donor proposed Cox model significant (p ≤ 0.05) variable interpretation.

Variable: category	Baseline category	Hazard ratio	p-value
AGE_DON (donor age)	n/a	0.993	0.0913
CURRENT_PRA (PRA level)	n/a	1.0084	0.000101
DIAB (diabetes status): unknown	negative	1.729	0.00556
ETHCAT (ethnicity): hispanic	white	0.639	0.00300
GENDER (gender): male	female	0.823	0.0525
MECH_DEATH_DON (donor mechanism of death): cardiovascular and others	asphyxiation and anoxia	4.151	0.00724
YRS_DIAL (years on dialysis)	n/a	0.965	0.00257

level.

3.6 Discussion

We developed a survival prediction model for pediatric kidney transplant recipients. We showed that the donor type has a significant impact on the survival of the pediatric transplant recipients and developed separate survival prediction model for each donor type. Our proposed model is built using variables specifically selected for the pediatric kidney transplant recipients, and it has higher prediction accuracy than models based on the general kidney transplant recipients.

For further research, a possible direction is to incorporate the variable correlations in the survival model. In the dataset, variables such as the recipient's Body Mass Index (BMI), age, and time on dialysis can be highly correlated. Another example is the paired features of the donor and the recipient, such as the donor age and the recipient age. Incorporating variable interaction terms and consider variable correlations may improve the survival prediction model.

Table 3.5: Deceased donor variable selection comparison. The variables are ordered by their names in the UNOS dataset alphabetically. The **bold** variables are distinct ones appearing in the proposed models.

Variable name in dataset: meaning	**proposed Cox**	**proposed RSF**	EPTS	all-age RSF [16]
AGE (recipient age)	✓	✓	✓	✓
ANY_DIAL (any dialysis)		✓		
ANY_PRIVATE (insurance type)				✓
COLD_ISCH_KI (cold ischemic time)				✓
CREAT_TRR (creatine level at transplant)		✓		✓
CURRENT_PRA (current panel reactive antibodies)	✓	✓		
DIAB (diabetes status)	✓		✓	✓
DIAG_KI (kidney diagnosis)	✓			✓
ETHCAT (ethnicity)	✓			✓
ETHCAT_DON (donor ethnicity)	✓			
FUNC_STAT_TRR (functional status at transplant)	✓	✓		✓
GENDER (gender)	✓			
HCV_SEROSTAT (hepatitis C infection status)				✓
HIST_HYPERTENS_DON (donor history of hypertension)				✓
MALIG (any previous malignancy)		✓		
MECH_DEATH_DON (donor mechanism of death)	✓	✓		✓
MED_COND_TRR (medical condition at transplant)	✓			✓
PREV_TX (prior transplants)			✓	
REGION (region)				✓
TX_YR (transplant year)	✓	✓		
YRS_DIAL (years on dialysis)			✓	
Number of variables	11	8	4	13

* variable transformations and interaction terms as specified in the EPTS model are included.

† ANY_PRIVATE is used in place of PAYMENTSOURCE_AT_TRANSPLANT. HIST_DIABETES_DON: donor history of diabetes is missing or category unknown for most cases and is thus omitted.

Table 3.6: Living donor model performance comparison. The metrics reported are from 5-fold cross validation.

Performance measure	**Proposed Cox**	**Proposed RSF**	EPTS model (in-use)	All-age RSF[16]
c-index	**0.57**	**0.54**	0.48	0.49
5 year Brier score	0.018	0.018	0.018	0.018

Table 3.7: Living donor proposed Cox model significant (p ≤ 0.05) variable interpretation.

Variable: category	Baseline category	Hazard ratio	p-value
DIAG_KI: tubular and interstitial diseases	glomerular disease	2.399	3.63e-06
DIAG_KI: polycystic kidneys	glomerular disease	1.741	0.0924
HCV_DON: unknown	not infected	0.663	0.0591
HCV_SEROSTATUS: unknown	negative	2.330	0.0240

Our survival prediction model can serve as a powerful tool for making decisions in the organ allocation network. For the recipients and the physicians, the model provides a customized survival prediction curve that shows the expected survival probability over time if the patient accepts an offered organ. The model is also useful for prioritizing patients on the waiting list for organs.

At the meantime, we have received a new dataset with more recent transplant cases. The new dataset is a great opportunity to test whether the methodologies adopted in this chapter would easily transfer to another dataset. It is also useful for studying the evolution of transplant survival over time. The study of predicting the survival curves for pediatric kidney transplant recipients continues and the results would soon appear in another paper.

Table 3.8: Living donor variable selection comparison. The variables are ordered by their names in the UNOS dataset alphabetically. The **bold** variables are distinct ones appearing in the proposed Cox model and not in the EPTS model or the all-age RSF model.

Variable name in dataset: meaning	**proposed Cox**	**proposed RSF**	EPTS*	all-age RSF[16]†
AGE: age			✓	✓
ANY_PRIVATE: payment/insurance type				✓
BMI_CALC: body mass index		✓		
COLD_ISCH_KI: cold ischemic time		✓		✓
CREAT_TRR: creatine at transplant				✓
DIAB: diabetes status			✓	✓
DIAG_KI: kidney diagnosis	✓	✓		✓
DRUGTRT_COPD: drug treated COPD at registration	✓			
ETHCAT: ethnicity				✓
ETHCAT_DON: donor ethnicity		✓		
EXH_PERIT_ACCESS: exhausted vascular access at registration	✓	✓		
FUNC_STAT_TRR: functional status at transplant				✓
GENDER_DON: donor gender	✓			
HAPLO_TY_MATCH_DON: living donor and recipient haplo type match		✓		
HBV_DON: donor hepatitis B infection status		✓		
HCV_DON: donor hepatitis C infection status	✓	✓		
HCV_SEROSTAT: hepatitis C infection status	✓	✓		✓
HIST_HYPERTENS_DON: donor history of hypertension				✓
MALIG: any previous malignancy				
MED_COND_TRR: medical condition at transplant				✓
PREV_TX: prior transplants			✓	
REGION: region		✓		✓
REGION_DON: donor region		✓		
TX_PROCEDUR_TY_KI: transplant procedure type	✓			
TX_YR: transplant year		✓		
YRS_DIAL: years on dialysis	✓		✓	
Number of variables	8	12	4	12

*Variable transformations and interaction terms as specified in the EPTS model are included.

† ANY_PRIVATE is used in place of PAYMENTSOURCE_AT_TRANSPLANT. DEATH_MECH_DON: donor death mechanism is not used since it is not applicable to the living donors. HIST_DIABETES_DON: donor history of diabetes is missing or category unknown for most cases and is thus omitted.

CHAPTER 4

GRAPH BASED VARIABLE SELECTION FOR SURVIVAL ANALYSIS

4.1 Introduction

Variable selection is a fundamental problem in survival analysis. When developing an accurate survival predicting model, identifying the proper variables to include in the model is often essential. In many applications, there exists an underlying graphical structure for the predictors. For example, some predictors may have strong correlations or interactions. When predicting the survival probability of a transplant recipient, it is important to consider the compatibility of the recipient and the organ donor. In such cases, incorporating the graph structure into the penalty function for variable selection would allow more accurate inference.

In this study, we adopt the classical Cox proportional hazard model [34] as the baseline model for survival prediction. The goal is to obtain an accurate and consistent estimate of the unknown parameters, the coefficients of the predicting variables. Most current variable selection methods for the Cox proportional hazard model use the penalized likelihood function as the objective function. The most frequently used penalties include the classical lasso [47, 41], the ridge regression [48], the elastic net [49, 50], the smoothly clipped absolute deviation (SCAD) penalty [51, 52], the adaptive lasso [53, 54], the fused lasso [55, 56], and the minimax concave penalty [57]. In survival applications involving categorical variables, the group lasso penalty [58] is also often used.

We study a fused lasso type of penalty constraint to the Cox proportional hazard model and provide its performance guarantees. This, to the best of our knowledge, is a new addition to the study of variable selection in survival analysis. Following [59], a graph-based penalty function is applied to the Cox proportional hazard model. Graph based regression problems have been studied in [59], [60], [55].

We would like to mention that although the formulation and results in this paper is motivated by healthcare applications associated with organ transplantation, the proposed method can be useful in many different settings. For example, in social networks and seismology applications, the underlying graph structure between users and seismic sensors can also be utilized to design the penalty term for model regularization.

First we formulate the survival analysis model and the graph-based penalty function in Section 4.1. Next we generalize the method introduced by [59] to the survival analysis likelihood function. Theoretical analysis of the accuracy and consistency are provided in Section 4.2. In Section 4.3 we compare different regularization methods using simulation. In Section 4.4 we apply the proposed method to a real data example and illustrate the benefit. In Section 4.5 we make the conclusion. All proofs are delegated to the Appendix.

4.2 Proportional Hazards Model and Penalty

4.2.1 Problem Formulation

Denote T as the survival time, and T is a random variable with cumulative distribution function $F(t) = \mathbb{P}(T \leq t)$, and density function $f(t) = F'(t) = \frac{d}{dt}F(t)$. Define the survival function as the upper tail probability $S(t) = \mathbb{P}(T > t) = 1 - F(t)$, and similarly, the survival event density function is $s(t) = S'(t) = -f(t)$. The hazard function is

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}.$$

Denote cumulative hazard function as $H(t) = \int_0^t h(u)du$, then we have

$$S(t) = \exp(-H(t)).$$

Assume the usual survival data in the form $(y_1, \delta_1, \boldsymbol{x}_1), \ldots, (y_n, \delta_n, \boldsymbol{x}_n)$, where y_i is the time until an event of interest, $\delta_i = 1$ indicates a complete observation and $\delta_i = 0$ a right-censored observation, and $\boldsymbol{x}_i = \{x_{i1}, \dots, x_{ip}\}^T$ is the vector of predictorsc (covariates) for subject i. For simplicity, assume that there are no tied event times. Given $\{(y_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^n$, the likelihood function is

$$L = \prod_{i:\delta_i=1} f(y_i|\boldsymbol{x}_i) \prod_{i:\delta_i=0} S(y_i|\boldsymbol{x}_i) = \prod_{i:\delta_i=1} h(y_i|\boldsymbol{x}_i) \prod_{i=1}^{n} S(y_i|\boldsymbol{x}_i).$$

Throughout this study, we assume the Cox proportional hazard model [34], in which the hazard function at time t given $\boldsymbol{x}_i$ takes the form

$$h(t|\boldsymbol{x}_i) = h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i),$$

where h_0 is the baseline hazard function, and $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_p\}^T$ is the vector of parameters to be estimated. Let $H_0(t) = \int_0^t h_0(u)du$, then $H(t|\boldsymbol{x}_i) = H_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)$ and we have

$$S(y_i|\boldsymbol{x}_i) = \exp(-H(y_i|\boldsymbol{x}_i)) = \exp(-H_0(y_i) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)).$$

The full log-likelihood function is

$$l(\boldsymbol{\beta}) = \sum_{i:\delta_i=1} [\log h_0(y_i) + \boldsymbol{\beta}^T \boldsymbol{x}_i] - \sum_{i=1}^{n} H_0(y_i) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i). \tag{4.1}$$

Our goal is to infer the unknown parameters $\boldsymbol{\beta}$ given censored observations.

4.2.2 Partial Likelihood Function

The baseline hazard function $h_0(\cdot)$ is usually unknown and has not been parameterized. Therefore, we adopt the commonly used partial likelihood function [61] instead of the full log-likelihood shown in (4.1). To derive the partial likelihood function, we note that the probability of the event being observed for subject i at time y_i is the partial likelihood function

$$L_i(\boldsymbol{\beta}) = \frac{h(y_i|\boldsymbol{x}_i)}{\sum_{j:y_j \geq y_i} h(y_i|\boldsymbol{x}_j)} = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\sum_{j:y_j \geq y_i} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_j)}.$$

Assuming independence of the observations, the joint partial likelihood function becomes

$$L(\boldsymbol{\beta}) = \prod_{i:\delta_i=1} L_i(\boldsymbol{\beta}) = \prod_{i:\delta_i=1} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\sum_{j:y_j \geq y_i} \exp(\boldsymbol{\beta}^T \mathbf{x}_j)},$$

and the log likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \Big\{ \boldsymbol{\beta}^T \boldsymbol{x}_i - \log \Big(\sum_{j:y_j \geq y_i} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_j) \Big) \Big\}. \tag{4.2}$$

[52] also gives another interpretation of (4.2) as substituting the "least informative" nonparametric prior for $H_0(\cdot)$.

We will use the formulation (4.2) in the rest of this section.

4.2.3 Classical Lasso Based Penalties

The classical lasso based method for the Cox model solves the penalized optimization problem

$$\min_{\boldsymbol{\beta}} -\frac{1}{n} l(\boldsymbol{\beta}) + g(\boldsymbol{\beta}), \tag{4.3}$$

where $g(\boldsymbol{\beta})$ is some penalty term. For classical lasso [41], $g(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1$. For SCAD penalty [52], $g(\boldsymbol{\beta}) = \sum_{j=1}^{p} f_\lambda(|\beta_j|)$, where $f'_\lambda(\theta) = I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda), a > 2,\ \theta > 0$. For elastic net [50], $g(\boldsymbol{\beta}) = \frac{\gamma}{2} \sum_{j=1}^{p} \beta_j^2 + \lambda \sum_{j=1}^{p} |\beta_j|$. For fused lasso [56], $g(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j|$. For adaptive lasso [54], $g(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \tau_j |\beta_j|$ with positive weights τ_j. For group lasso [58], $g(\boldsymbol{\beta}) = \lambda \sum_{k=1}^{p} \|\boldsymbol{\beta}_{\mathcal{I}_k}\|_2$, where $\mathcal{I}_k$ is the set of variables belonging to the k^{th} group.

4.2.4 The Graph-based Penalty

We introduce a new penalty to the Cox model based on the predictor graph in order to select correlated variables in the model. Let $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T = (X_1, \ldots, X_p) \in \mathbb{R}^{n \times p}$. Assume a known inverse covariance structure among $X_1, \ldots, X_p$. For simplicity, we can construct an undirected and unweighted graph G that captures the correlations among the predictors. Let E be the adjacency matrix of the graph, where $E_{i,i'} = 1$ if there is an edge between X_i and $X_{i'}$, and 0 otherwise $(1 \leq i, i' \leq p)$. Let $\mathcal{N}_i$ be the set of indices of the neighboring predictors of X_i, i.e. $\mathcal{N}_i = \{k' : E_{i,i'} = 1\}$, and let $d_i = |\mathcal{N}_i|$. By incorporating the additional information on X, (4.3) can be rewritten as

$$\begin{aligned} \min_{\boldsymbol{\beta}, V^{(1)}, \ldots, V^{(p)}} \quad & -\frac{1}{n} \sum_{i=1}^{n} \delta_i \Big\{ \boldsymbol{\beta}^T \boldsymbol{x}_i - \log \Big(\sum_{j: y_j \geq y_i} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_j) \Big) \Big\} + \lambda \sum_{k=1}^{p} \tau_k \|V^{(k)}\|_2, \qquad (4.4) \\ \text{s.t.} \quad & \sum_{k=1}^{p} V^{(k)} = \boldsymbol{\beta}, \\ & \text{supp}(V^{(k)}) \subset \mathcal{N}_k, \end{aligned}$$

where τ_k is a positive weight for the k^{th} group.

We can further define a new norm of $\boldsymbol{\beta}$ as

$$\|\boldsymbol{\beta}\|_{G,\tau} = \min_{\sum_{i=1}^{p} V^{(i)} = \boldsymbol{\beta},\ \text{supp}(V^{(i)}) \subseteq \mathcal{N}_i} \sum_{i=1}^{p} \tau_i \|V^{(i)}\|_2. \qquad (4.5)$$

It can be verified that $\| \cdot \|_{G,\tau}$ satisfies the triangle inequality and is indeed a norm [62]. Using this norm, the formulation (4.4) is equivalent to

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} -\frac{1}{n} \sum_{i=1}^{n} \delta_i \Big\{ \boldsymbol{\beta}^T \boldsymbol{x}_i - \log \Big(\sum_{j: y_j \geq y_i} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_j) \Big) \Big\} + \lambda \|\boldsymbol{\beta}\|_{G,\tau}.$$

4.3 Computation

In this section, we show how the optimization problem (4.4) could be transformed to remove the constraint terms. The technique is developed based on the predictor duplication method in [59].

Let $\boldsymbol{x}^i_{\mathcal{N}_k}$ be the $|\mathcal{N}_k| \times 1$ subvector of $\boldsymbol{x}_i$, whose indices are in $\mathcal{N}_k$. Let $V^{(k)}_{\mathcal{N}_k}$ be the $|\mathcal{N}_k| \times 1$ subvector of $V^{(k)}$. Then $\sum_{i=1}^n \boldsymbol{\beta}^T \boldsymbol{x}_i = \sum_{i=1}^n \sum_{k=1}^p {V^{(k)}_{\mathcal{N}_k}}^T \mathbf{x}^i_{\mathcal{N}_k}$, and the log likelihood function can be rewritten as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^n \delta_i \Big\{ {V^{(i)}_{\mathcal{N}_i}}^T \mathbf{x}_{\mathcal{N}_i} - \log\Big(\sum_{j:y_j \geq y_i} \exp({V^{(j)}_{\mathcal{N}_j}}^T \mathbf{x}_{\mathcal{N}_j}) \Big) \Big\}. \tag{4.6}$$

Therefore, the optimization problem can be solved using existing solvers for the group lasso penalty, such as the `R grpreg` package.

After obtaining the coefficients $\hat{V}^{(i)}_{\mathcal{N}_i}$, let $\hat{V}^{(i)}_{\mathcal{N}_i^c} = 0$. Then $\boldsymbol{\beta} = \sum_{i=1}^p \hat{V}^{(i)}_{\mathcal{N}_i}$.

4.4 Theoretical Properties

4.4.1 Assumptions

Denote $\boldsymbol{\beta}_0 = \{\beta_{01}, \ldots, \beta_{0p}\}$ as the true parameters, $J_0 = \{i : \beta_{0i} \neq 0\}$ is the index of non-zeros parameters, $J_0^c = \{i : \beta_{0i} = 0\}$ is the index of zero parameters, and $s_0 = |J_0|$ denotes the number of non-zero parameters. Let $\mathcal{U}(\boldsymbol{\beta})$ denote the set of all optimal decompositions of $\boldsymbol{\beta}$ that minimizes $\|\boldsymbol{\beta}\|_{G,\tau}$. In other words, $\mathcal{U}(\boldsymbol{\beta})$ consists of all optimal solutions to the problem (4.5). Denote $K_{G,\tau}(\boldsymbol{\beta})$ as

$$K_{G,\tau}(\boldsymbol{\beta}) = \min_{(V^{(1)}, V^{(2)}, \ldots, V^{(p)}) \in \mathcal{U}(\boldsymbol{\beta})} |\{i : \|V^{(i)}\|_2 \neq 0\}|,$$

and $K_{G,\tau}$ as the supreme of $K_{G,\tau}(\boldsymbol{\beta})$ over all $\boldsymbol{\beta}$ satisfying $\mathrm{supp}(\boldsymbol{\beta}) \subseteq J_0$,

$$K_{G,\tau} = \sup_{\mathrm{supp}(\boldsymbol{\beta}) \subseteq J_0} K_{G,\tau}(\boldsymbol{\beta}).$$

We note that $K_{G,\tau} = s_0$ is the graph G has no edge; and $K_{G,\tau} = K_0$ if G consists of some disconnected complete subgraphs and J_0 is the union of K_0 node sets of those disconnected subgraphs.

Assumption 4.4.1 (Assumptions for the likelihood [52])**.** *We have the following assumptions for the partial likelihood function:*

1. $\int_0^1 h_0(t)dt < \infty$.

2. *The processes $\boldsymbol{x}(t)$ and $Y(t)$ are left-continuous with right hand limits, and*

$$\mathbb{P}\{Y(t) = 1, \forall t \in [0,1]\} > 0.$$

3. *There exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}_0$ such that*

$$\mathbb{E} \sup_{t\in[0,1],\ \boldsymbol{\beta}\in\mathcal{B}} Y(t)\boldsymbol{x}(t)^T\boldsymbol{x}(t)\exp(\boldsymbol{\beta}^T\boldsymbol{x}(t)) < \infty$$

4. *Define*

$$\begin{aligned} s^{(0)}(\boldsymbol{\beta}, t) &= \mathbb{E}Y(t)\exp(\boldsymbol{\beta}^T\boldsymbol{x}(t)) \\ s^{(1)}(\boldsymbol{\beta}, t) &= \mathbb{E}Y(t)\boldsymbol{x}(t)\exp(\boldsymbol{\beta}^T\boldsymbol{x}(t)) \\ s^{(2)}(\boldsymbol{\beta}, t) &= \mathbb{E}Y(t)\boldsymbol{x}(t)\boldsymbol{x}(t)^T\exp(\boldsymbol{\beta}^T\boldsymbol{x}(t)) \end{aligned}$$

where $s^{(0)}(\cdot, t)$, $s^{(2)}(\cdot, t)$, $s^{(2)}(\cdot, t)$ are continuous in $\boldsymbol{\beta} \in \mathcal{B}$, uniformly in $t \in [0,1]$. $s^{(0)}, s^{(1)}, s^{(2)}$ are bounded on $\mathcal{B} \times [0,1]$; $s^{(1)}$ is bounded away from zero on $\mathcal{B} \times [0,1]$.

The matrix

$$I(\boldsymbol{\beta}_0) = \int_0^1 v(\boldsymbol{\beta}_0, t) s^{(0)}(\boldsymbol{\beta}_0, t) h_0(t) dt$$

is finite positive definite, where

$$v(\boldsymbol{\beta}, t) = \frac{s^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \left(\frac{s^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)}\right)\left(\frac{s^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)}\right)^T.$$

The reason of imposing the above four assumptions is to obtain the local asymptotic quadratic property for the partial likelihood function $\ell(\boldsymbol{\beta})$, as well as the asymptotic normality of the maximum partial likelihood estimates [63, 64].

Assumption 4.4.2 (Assumptions for the predicted graph G)**.** *The following assumptions are required for the predicted graph G.*

1. *The neighboorhood $\mathcal{N}_i \subseteq J_0$, $\forall i \in J_0$.*

2. *There exists $\kappa > 0$ such that*

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^p\setminus\{0\}} \inf_{(V^{(1)},V^{(2)},\ldots,V^{(p)})\in\mathcal{U}(\boldsymbol{\beta})} \frac{\frac{1}{2}(\sum_{i=1}^p V^{(i)})^T I(\boldsymbol{\beta}_0)(\sum_{i=1}^p V^{(i)})}{\sum_j \tau_j^2 \|V^{(j)}\|_2^2} \geq \kappa.$$

Remark 1. *The Assumption 4.4.2 (1) assumes that the predicted graph G is consistent with the true parameter $\boldsymbol{\beta}_0$, the same as the assumption A2 in [59]. The Assumption 4.4.2 (2) is a restriction on the smallest eigenvalue of the Fisher information at true parameter β_0. Compared with the assumption A3 for the data matrix X in [59], here the assumption is for the Fisher information matrix, due to a different loss function $-l(\boldsymbol{\beta})$ here.*

4.4.2 Finite Sample Bounds

Theorem 4.4.3 (Oracle property)**.** *Under the Assumptions 4.4.1 and 4.4.2, let* $\tau^* = \min_{1\le i\le p} \tau_i$*. For any optimal solution* $\hat{\boldsymbol{\beta}}$ *of problem* (4.4)*, we have*

$$\frac{1}{n}\left\{l(\boldsymbol{\beta}_0) - l(\hat{\boldsymbol{\beta}})\right\} \le \frac{\lambda^2 K_{G,\tau}}{\kappa}, \quad \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \le \frac{\lambda\sqrt{pK_{G,\tau}}}{\kappa\tau^*}.$$

4.4.3 Asymptotic Normality

Theorem 4.4.4 (Asymptotic Normality)**.** *When dimension* p *is fixed, assume* $\sqrt{n}\lambda \to 0$ *and* $\tau_j = O(1)$ *for each* $j \in J_0$*,* $n^{\gamma+1/2}\lambda \to \infty$*,* $\boldsymbol{u}_{J_0^c} \neq 0$*, and* $\liminf_{n\to\infty} n^{-\gamma/2}\tau_j > 0$ *for each* $j \in J_0^c$*, under Assumptions 4.4.2 (1), we have as* $n \to \infty$*,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{J_0} - \boldsymbol{\beta}^0_{J_0}) \xrightarrow{d} N(0, I_{J_0}(\boldsymbol{\beta}_0)^{-1}), \quad \hat{\boldsymbol{\beta}}_{J_0^c} \xrightarrow{d} 0.$$

4.5 Simulation Study

To evaluate the performance of the graph regularizer for the Cox model, it is compared with some existing regularizers for the Cox model, including the classical lasso [47, 41], ridge regression [48], elastic net [49, 50], smoothly clipped absolute deviation (SCAD) [51, 52], and adaptive lasso (Alasso) [53, 54].

The regularized survival models are evaluated on the following performance measures:

- ℓ_2 error of the estimated coefficients: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2$;
- Relative prediction error (RPE):

  $$\text{RPE} = \frac{1}{N_{\text{test}}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T X_{\text{test}}^T X_{\text{test}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0),$$

  where N_{test} is the test data size and X_{test} are the test covariates;
- Harrell's concordance index (c-index)[42]. The c-index is a commonly used metric

for evaluating survival prediction models. It measures the ability of the model to correctly predict the ranking of the survival time given a pair of new observations and is equivalent to the AUC (Area Under Curve) [65]. A c-index of 0.5 is equivalent to random guessing and 1 is perfect prediction. In recent survival applications, a c-index between 0.6 and 0.7 is often considered satisfactory [45].

- Non-zero match ratio (NMR) and Zero match ratio (ZMR):

$$\text{NMR} = \frac{|\{i,j\} : \Omega_{ij} \neq 0, \hat{\beta}_i \neq 0, \hat{\beta}_j \neq 0\}|}{|\{i,j\} : \Omega_{ij} \neq 0, \beta_i^0 \neq 0, \beta_j^0 \neq 0\}|};$$

$$\text{ZMR} = \frac{|\{i,j\} : \Omega_{ij} \neq 0, \hat{\beta}_i = 0, \hat{\beta}_j = 0\}|}{|\{i,j\} : \Omega_{ij} \neq 0, \beta_i^0 = 0, \beta_j^0 = 0\}|}.$$

  NMR examines whether the estimated coefficients of a pair of connected ($\Omega_{ij} \neq 0$) variables useful ($\beta_i^0 \neq 0, \beta_j^0 \neq 0$) in simulating the survival outcomes are both nonzero ($\hat{\beta}_i \neq 0, \hat{\beta}_j \neq 0$), and ZMR examines whether the estimated coefficients of a pair of connected ($\Omega_{ij} \neq 0$) variables with no influence ($\hat{\beta}_i = 0, \hat{\beta}_j = 0$) on the simulated survival outcomes are both zero ($\hat{\beta}_i = 0, \hat{\beta}_j = 0$). Intuitively, the NMR measures the degree to which the model can identify useful variable relations, and the ZMR checks whether the model can discard non-informative variable relations.

Three types of predictor graph topologies are tested in the simulation study: (1) the sparse graph, (2) the ring graph, and (3) the graph with communities. Figure 4.1 shows illustrations that represent the three graph topologies.

The proposed graph regularizer shows overall the most promising performance among the regularizers for the Cox model that are tested in the simulation study and the real-data study.

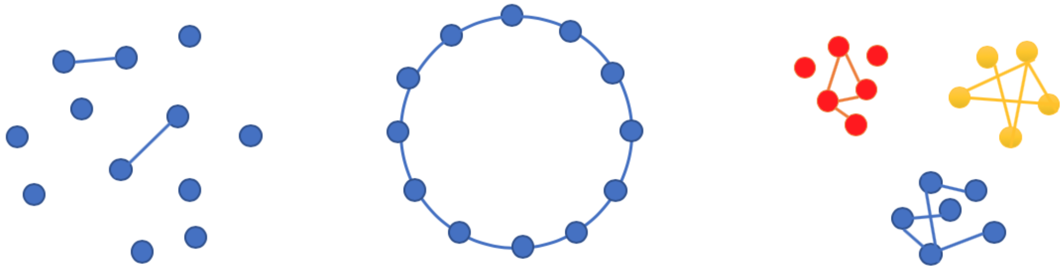


Figure 4.1: Illustration of the three predictor graph topologies in the simulation: the sparse graph, the ring graph, and the graph with three communities.

4.5.1 The Sparse Graph

Consider a sparse Erdos-Renyi predictor graph with a small edge formation probability p_0. Assume the predictors $(X_1, X_2, \ldots, X_p)^T \sim N(0, \Omega^{-1})$, where $p = 100$ and Ω is an inverse covariance matrix whose off-diagonal entries equal 0.5 with probability 0.01 and 0 otherwise. In practice, we compute $\Sigma = \Omega^{-1}$ using the *nearPD* transformation in the R *matrix* package [66] to ensure that Σ is positive definite. Let the true parameters be $\boldsymbol{\beta}^0 = \Omega\Sigma_{xy}$, where $\Sigma_{xy} = (c_1, c_2, \ldots, c_p)^T$. Let $c_i = 10$ for the top 4 predictors with maximum edges, and $c_i = 0$ otherwise. The experiment is similar to that in [59].

The survival time is simulated using the R *coxed* [67] package with a censor rate of 0.3. The train size is 100 and the test size is 400. The hyper-parameters in each model are tuned by cross validation using the training data. The experiment is repeated 50 times, and the results (mean and standard deviation) of the models are shown in Table 4.1, 4.2, and 4.3 for $p_0 = 0.01, 0.05, 0.1$ respectively.

Results

We observe that, when the predictor graph takes the form of a sparse Erdos-Renyi graph, the graph lasso has higher performance on the ℓ_2 error, the RPE, and the c-index, than other regularizers and baseline models, regardless of the edge formation probability. On the NMR, other than the ridge regression and the baseline Cox model, whose estimated coefficients are mostly non-zeros and are hence biased, the proposed graph lasso penalty has the highest performance. The graph lasso also has competent ZMR result. Ridge and the baseline Cox model sacrifice the ZMR for the NMR and fail to discard the unimportant variables.

As the edge formation probability p_0 increases, the performance of all models is reduced compared to their own when p_0 is smaller. Nevertheless, the graph lasso penalty consistently acquires better estimation and prediction than the other models regardless of the change in p_0.

Table 4.1: Performance on the sparse Erdos-Renyi predictor graph, $p_0 = 0.01$.

Model	ℓ_2 norm	RPE	c-index	NMR	ZMR
Graph lasso	**28.73**(0.39)	**415.98**(28.90)	**0.74** (0.042)	0.12(0.12)	0.92(0.093)
Lasso	30.29(0.12)	501.58(29.52)	0.66 (0.042)	0.056(0.096)	0.94(0.088)
Ridge regression	30.37(0.34)	504.85(29.59)	0.60 (0.023)	**1.00(0.00)**	0.71(0.28)
Elastic net	30.30(0.084)	502.24(29.71)	0.66(0.042)	0.086(0.12)	0.91(0.098)
SCAD	30.33(0.088)	503.29(29.75)	0.66(0.048)	0.028(0.040)	0.98(0.030)
Alasso	30.40(0.016)	505.69(29.64)	0.62 (0.074)	0.056(0.096)	**1.00** (0.0052)
$\hat{\boldsymbol{\beta}} = \mathbf{0}$	30.41(0.00)	506.34(29.70)	0.50 (0.00)	0.00(0.00)	**1.00**(0.00)
Cox without penalty	Inf (-)	$7.54 \times 10^7 (4.00 \times 10^8)$	0.53 (0.044)	**1.00**(0.00)	0.00(0.00)

Table 4.2: Performance on the sparse Erdos-Renyi predictor graph, $p_0 = 0.05$.

Model	ℓ_2 norm	RPE	c-index	NMR	ZMR
Graph lasso	**41.94**(0.46)	**576.57**(50.28)	**0.70**(0.034)	0.044(0.062)	0.93(0.063)
Lasso	42.28(0.15)	585.81 (49.80)	0.68(0.034)	0.068(0.083)	0.88(0.12)
Ridge regression	42.37(0.052)	589.49(50.06)	0.66(0.034)	**1.00**(0.00)	0.55(0.36)
Elastic net	42.27(0.15)	585.59(50.79)	0.67(0.033)	0.105(0.11)	0.807(0.17)
SCAD	42.37(0.092)	589.25(49.98)	0.68(0.048)	0.0056(0.014)	0.97(0.048)
Alasso	42.43(0.020)	590.93 (50.03)	0.62 (0.06)	0.068(0.083)	0.97(0.061)
$\hat{\boldsymbol{\beta}} = \mathbf{0}$	42.41(0.00)	591.73 (50.11)	0.50 (0.00)	0.00(0.00)	**1.00**(0.00)
Cox without penalty	Inf (-)	$5.04 \times 10^8 (1.59 \times 10^9)$	0.57(0.054)	0.97(0.088)	0.0053(0.019)

4.5.2 The Ring Graph

The second experiment is on a ring predictor graph where the variables are nodes on the ring and each node is connected to its immediate two neighbors. Let $(X_1, X_2, \ldots, X_p)^T \sim N(0, \Omega^{-1})$, where $p = 100$. Let $\Omega = B + \delta I$, where $B_{ij} = 0.5$ for $|i-j| < 2$ and $B_{ii} = 0$, δ

Table 4.3: Performance on the sparse Erdos-Renyi predictor graph, $p_0 = 0.1$.

Model	ℓ_2 norm	RPE	c-index	NMR	ZMR
Graph lasso	**59.99**(0.55)	**869.37**(58.42)	**0.70**(0.034)	0.020(0.027)	0.97(0.054)
Lasso	60.46(0.19)	885.81(58.62)	0.68(0.033)	0.038(0.055)	0.88(0.14)
Ridge regression	60.57(0.055)	890.19(58.15)	0.66(0.032)	**1.00**(0.00)	0.63(0.40)
Elastic net	60.47(0.15)	886.21(58.73)	0.68(0.033)	0.058(0.065)	0.84(0.16)
SCAD	60.57(0.058)	889.94(58.00)	0.67(0.039)	0.0042(0.0068)	0.98(0.031)
Alasso	60.60(0.025)	891.45 (58.00)	0.61 (0.064)	0.038(0.055)	0.96(0.058)
$\hat{\boldsymbol{\beta}} = \mathbf{0}$	60.62(0.00)	892.38(57.97)	0.50 (0.00)	0.00(0.00)	**1.00**(0.00)
Cox without penalty	Inf (-)	$5.54 \times 10^9(2.70 \times 10^{10})$	0.54(0.049)	0.82(0.21)	0.024(0.029)

is chosen to make the condition number of Ω equal to p. Let the true parameter $\beta^0 = \Omega\Sigma_{xy}$, where $\Sigma_{xy} = \mathbf{1}$. The ZMR is not calculated as it is not applicable in the ring graph setting.

Results

We observe that the graph lasso has the best performance on the ℓ_2 norm, the RPE, and the c-index when the predictor graph is a ring graph. The NMRs for the ridge regression and the baseline Cox model are high since their estimated coefficients are mostly non-zeros. The competing models have close performance as the graph lasso since the relations among the variables in the ring graph are relatively simple.

Table 4.4: Performance on the ring predictor graph.

Model	ℓ_2 norm	RPE	c-index	NMR
Graph lasso	**23.82**(0.23)	**232.78**(16.62)	**0.68**(0.032)	0.020(0.026)
Lasso	23.98(0.045)	235.41(16.03)	0.66(0.035)	0.0060(0.016)
Ridge regression	23.97(0.037)	235.24(16.14)	0.66(0.038)	**1.00**(0.00)
Elastic net	23.96(0.083)	235.17(16.35)	0.66(0.035)	0.020(0.053)
SCAD	24.00(0.0092)	235.65(16.00)	0.61(0.034)	0.00022(0.0015)
Alasso	24.00(0.0039)	235.68 (16.03)	0.55 (0.040)	0.006(0.016)
$\hat{\boldsymbol{\beta}} = \mathbf{0}$	24.00(0.00)	235.70(16.03)	0.50 (0.00)	0.00(0.00)
Cox without penalty	Inf (-)	$2.07 \times 10^5(3.60 \times 10^5)$	0.55(0.044)	0.99(0.0095)

4.5.3 The Graph with Communities

Suppose some of the predictors have community identities, and for predictors in the same community, an edge forms with probability p_{inner}. For predictors in different communities or those not in any communities, let the probability of edge formation among them be

p_{outer}. Let $p_{inner} = 0.5, 0.7, 0.9$, and $p_{outer} = 0.01$. For dimension $p = 100$, we assume there exist three communities, each with size 30. The performance comparison is in Table 4.5, 4.6, and 4.7 for $p_{inner} = 0.5, 0.7$, and 0.9 respectively.

Results

We observe that the graph lasso penalty has the best ℓ_2 norm and c-index regardless of the value of p_{inner}. As p_{inner} increases, the communities become more dense, and the relations among the variables become more complex. Therefore, it is more difficult for the models to acquire accurate estimation and prediction. The competing models have close performance as the graph lasso.

Table 4.5: Performance on the 3-community predictor graph, $p_{inner} = 0.5$.

Model	ℓ_2 norm	RPE	c-index	NMR	ZMR
Graph lasso	**59.80**(0.75)	**746.56**(78.63)	**0.70**(0.036)	0.011(0.010)	0.98(0.040)
Lasso	60.56(0.058)	770.61(77.78)	0.66(0.031)	0.0074(0.012)	0.92(0.11)
Ridge regression	60.60(0.029)	772.13(76.88)	0.64(0.034)	**1.00**(0.00)	0.70(0.34)
Elastic net	60.55(0.081)	769.99(77.67)	0.66(0.032)	0.027(0.047)	0.87(0.17)
SCAD	60.60(0.028)	772.26(77.00)	0.64(0.036)	0.0014(0.0021)	0.98(0.055)
Alasso	60.62(0.0057)	773.20 (76.98)	0.57(0.043)	0.0074(0.012)	0.98(0.049)
$\hat{\boldsymbol{\beta}} = \mathbf{0}$	60.62(0.00)	773.48(76.91)	0.50 (0.00)	0.00(0.00)	**1.00**(0.00)
Cox without penalty	Inf (-)	$1.40 \times 10^6(5.16 \times 10^6)$	0.54(0.055)	0.82(0.21)	0.024(0.029)

Table 4.6: Performance on the 3-community predictor graph, $p_{inner} = 0.7$.

Model	ℓ_2 norm	RPE	c-index	NMR	ZMR
Graph lasso	**77.74**(0.53)	737.92(43.31)	**0.70**(0.044)	0.0042(0.0046)	0.98(0.032)
Lasso	78.40(0.035)	**734.13**(41.24)	0.64(0.044)	0.0055(0.0099)	0.95(0.087)
Ridge regression	78.40(0.024)	734.50(41.13)	0.62(0.038)	**1.00**(0.00)	0.79(0.31)
Elastic net	78.40(0.028)	734.25(41.16)	0.64(0.043)	0.010(0.020)	0.94(0.11)
SCAD	78.41(0.026)	734.72(41.09)	0.61(0.040)	0.0010(0.0023)	0.98(0.040)
Alasso	78.42(0.0041)	735.14(40.98)	0.54 (0.038)	0.0055(0.0099)	0.99(0.020)
$\hat{\boldsymbol{\beta}} = \mathbf{0}$	78.42(0.00)	735.27(40.96)	0.50 (0.00)	0.00(0.00)	**1.00**(0.00)
Cox without penalty	Inf (-)	$1.81 \times 10^6(1.03 \times 10^7)$	0.55(0.051)	**1.00**(0.00)	0.024(0.029)

Table 4.7: Performance on the 3-community predictor graph, $p_{inner} = 0.9$.

Model	ℓ_2 norm	RPE	c-index	NMR	ZMR
Graph lasso	**90.06**(0.43)	875.48(88.090)	**0.68**(0.041)	0.0203(0.027)	0.98(0.032)
Lasso	90.55(0.017)	834.60(60.21)	0.611(0.037)	0.0018(0.0056)	0.97(0.0483)
Ridge regression	90.55(0.0094)	834.54(60.27)	0.55(0.027)	**1.00**(0.00)	0.90(0.18)
Elastic net	90.55(0.020)	**834.48**(60.10)	0.61(0.038)	0.0029(0.0083)	0.96(0.070)
SCAD	90.55(0.0076)	834.71(60.25)	0.55(0.029)	5.13×10^{-5}(0.00025)	**1.00**(0.020)
Alasso	90.55(0.0016)	834.84 (60.19)	0.53 (0.038)	0.0018(0.0056)	**1.00**(0.012)
$\hat{\boldsymbol{\beta}} = \mathbf{0}$	90.55(0.00)	834.87(60.19)	0.50 (0.00)	0.00(0.00)	**1.00**(0.00)
Cox without penalty	Inf (-)	$2.80 \times 10^8(1.61 \times 10^9)$	0.54(0.045)	0.99(0.051)	0.00(0.00)

4.6 Real Data Examples

4.6.1 The Pediatric Kidney Transplant Data

Predicting the survival time for transplant recipients is a crucial task for the transplant community. Accurate survival prediction can provide useful information for organ allocation decisions. A challenge with transplant survival prediction is that the data recorded for each transplant case are usually high dimensional and highly dependent. Therefore, building a predictor graph and and using the graph regularizer can be especially beneficial for solving the variable selection problems when building survival prediction models.

We use the proposed graph regularized Cox model to predict the survival time of pediatric recipients of kidney transplants. The dataset we use contains 19,236 pediatric kidney transplant cases in the U.S. from 1987 to 2014, and for each transplant case, 487 predictors are recorded. The dataset is provided by the UNOS (United Network for Organ Sharing).

Depending on the donor type, which is shown to be a significant variable influencing the post-transplant survival time for kidney transplant recipients [26, 36], the dataset is divided into two datasets marked with different donor types. For the data of each donor type, we develop the proposed graph penalized Cox model and compare it with some existing penalized Cox models. We process the data as described in the previous chapter, using steps including feature engineering, data heterogeneity detection, and missing data imputation. The prepared data dimension is 3905×66 for deceased donors and 5444×42 for living donors. Counting the different levels of the categorical variables, we have in total 145 variables for transplant cases with deceased donors and 98 variables for living donors. The variables and their definitions are found in Appendix A.

To use the graph penalty, we construct a predictor graph from the data where each numerical variable and each categorical variable level is a node, and their connectivities are represented by the formation of edges. One way to create a predictor graph is the following. We form edges for variables as follows.

1. Numerical predictors with significantly high inverse covariance (see Figure 4.2, 4.3). We connect variables pairs whose Pearson's test p-value is 0.01 or lower.

2. Levels of categorical variables that measure similar traits of the transplant recipient and donor. As an example, we connect the variable "HBV: positive" (Recipient HBV infection status: Positive) and "HBV_DON: positive" (Donor HBV infection status: Positive). This connection is based on our assumption that being in similar conditions as the donor is beneficial for the survival of an organ transplant recipient.

3. The different levels under the same categorical variable, similarly as in group lasso.

As a result, we derive the graphical structure for the predictors in the pediatric kidney transplant data in Table 4.8, 4.9.

We use 5-fold cross validation to test the models. The performance of the graph regularized Cox model is compared with some other current regularizers in Table 4.10 and Figure 4.4, 4.5. The blue line in each figure is the median of the graph regularized model, and the red line is the 0.5 reference line for making valid prediction. Since the true parameters are unknown in the real data, we only compute the c-index. We observe that the graph regularizer has the highest mean and median c-index for both donor types. The improvement of using the graph lasso is more prominent on the living donor dataset. This result is possible due to the fact that the living donor is more often related to the recipient and is likely to have closer biological and environmental characteristics as the recipient. More variables are also recorded from the living donors than from the deceased donors in

Table 4.8: Variable relations in the predictor graph based on the pediatric kidney transplant data with living donors. The different levels of the same categorical variables (groups as defined in the group lasso) are omitted here and can be found in Appendix A.

Variable	Neighbors
AGE	AGE_DON, BMI_CALC, COLD_ISCH_KI, CREAT_TRR, FUNC_STAT_TRR
AGE_DON	AGE, TX_YR, YRS_DIAL
BMI_CALC	AGE
COLD_ISCH_KI	AGE, DISTANCE, TX_YR, YRS_DIAL
CREAT_TRR	AGE, FUNC_STAT_TRR, TX_YR, YRS_DIAL
DISTANCE	COLD_ISCH_KI
FUNC_STAT_TRR	AGE_DON, COLD_ISCH_KI, CREAT_TRR, YRS_DIAL
HLAMIS	TX_YR, YRS_DIAL
TX_YR	AGE_DON, COLD_ISCH_KI, CREAT_TRR, HLAMIS
YRS_DIAL	AGE_DON, COLD_ISCH_KI, CREAT_TRR, FUNC_STAT_TRR, HLAMIS
HBVsusceptible.not.infected.wo.antibody	HBV_DONsusceptible.not.infected.wo.antibody
HBVvery.likely.infected	HBV_DONvery.likely.infected
HBVunknown	HBV_DONunknown
PRE_TX_TXFUSY	PREV_TXY
CITIZENSHIPRESIDENT.ALIEN	CITIZENSHIP_DONRESIDENT.ALIEN
CITIZENSHIPNON.RESIDENT.ALIEN	CITIZENSHIP_DONNON.RESIDENT.ALIEN
CMVP	CMV_DON_LP
CMVU	CMVND, CMV_DON_LND.or.U
ETHCATBlack	ETHCAT_DONBlack
ETHCATHispanic	ETHCAT_DONHispanic
ETHCATAsian	ETHCAT_DONAsian
ETHCATAmer.Ind.Alaska.Native	ETHCAT_DONAmer.Ind.Alaska.Native
ETHCATNative.Hawaiian.other.Pacific.Islander	ETHCAT_DONNative.Hawaiian.other.Pacific.Islander
ETHCATMultiracial	ETHCAT_DONMultiracial
GENDERM	GENDER_DONM
HCV_SEROSTATUSU	HCV_DONunknown

Table 4.9: Variable relations in the predictor graph based on the pediatric kidney transplant data with deceased donors. The different levels of the same categorical variables (groups as defined in the group lasso) are omitted here and can be found in Appendix A.

Variable	Neighbors
AGE	BMI_CALC, BMI_DON_CALC, BUN_DON, FUNC_STAT_TRR, YRS_DIAL
BMI_CALC	AGE, TX_YR
BMI_DON_CALC	AGE, CREAT_DON, TX_YR
BUN_DON	AGE, CREAT_DON, HLAMIS, SGPT_DON, TBILI_DON, TX_YR
CREAT_DON	BMI_DON_CALC, BUN_DON, TX_YR
FUNC_STAT_TRR	AGE, YRS_DIAL
HLAMIS	BUN_DON, DISTANCE, TX_YR
SGPT_DON	BUN_DON, TBILI_DON
TBILI_DON	BUN_DON, SGPT_DON
TX_YR	BMI_CALC, BMI_DON_CALC, BUN_DON, CREAT_DON, HLAMIS, YRS_DIAL
YRS_DIAL	AGE, FUNC_STAT_TRR, TX_YR
CMVP	CMV_DONP
CMVU	CMVND, CMV_DONN, CMV_DONND, CMV_DONU
DIABU	DIABY, DIABETES_DONU, DIABETES_DONY
ETHCATBlack	ETHCAT_DONBlack
ETHCATHispanic	ETHCAT_DONHispanic
ETHCATAsian	ETHCAT_DONAsian
ETHCATAmer.Ind.Alaska.Native	ETHCAT_DONAmer.Ind.Alaska.Native
ETHCATNative.Hawaiian.other.Pacific.Islander	ETHCAT_DONNative.Hawaiian.other.Pacific.Islander
ETHCATMultiracial	ETHCAT_DONMultiracial
GENDERM	GENDER_DONM

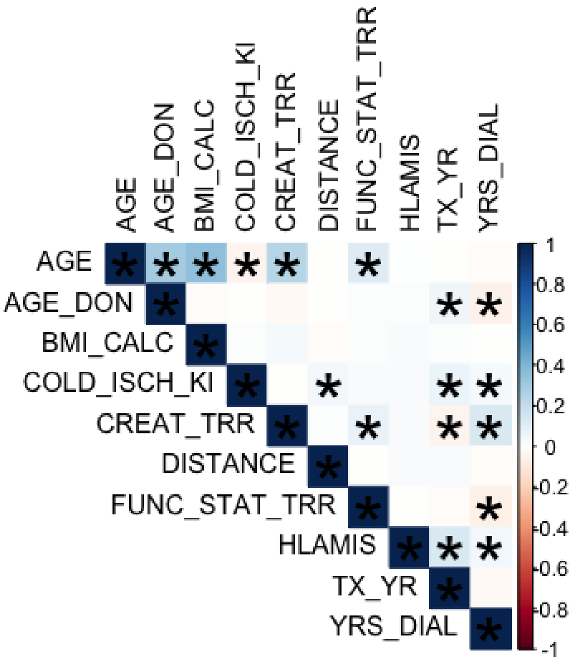


Figure 4.2: Inverse covariance of the numerical variables in the living donor dataset. The asterisked pairs have Pearson's test p-value 0.01 or lower.

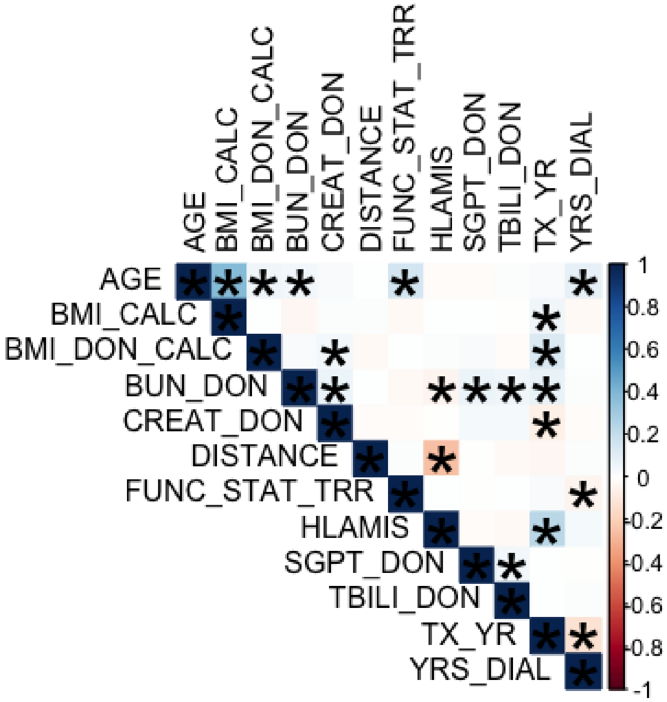


Figure 4.3: Inverse covariance of the numerical variables in the deceased donor dataset. The asterisked pairs have Pearson's test p-value 0.01 or lower.

Table 4.10: The performance of different penalties on the pediatric kidney transplant data.

Model	Living donors c-index	Deceased donors c-index
Graph lasso	**0.59**(0.045)	**0.58**(0.055)
Lasso	0.57(0.039)	0.57(0.055)
Ridge regression	0.49(0.039)	0.56(0.060)
Elastic net	0.57(0.038)	0.58(0.045)
SCAD	0.57(0.028)	0.57(0.056)
Alasso	0.57 (0.040)	0.57 (0.049)
Group lasso	0.57(0.051)	0.57(0.038)
Cox without penalty	0.49(0.039)	0.55(0.058)
$\hat{\boldsymbol{\beta}} = \mathbf{0}$	0.50(0.00)	0.50(0.00)

the dataset. Therefore, the living donor predictor graph we can create is more complicated than the deceased donor's, which gives the graph lasso regularizer more advantage over other penalties in predicting the survival outcome for pediatric recipients of living donor kidneys.

The variables selected by the models with different penalties also differ. Specifically, we compare the variables selected by the graph lasso and the group lasso. The comparison is in Table 4.11 for the deceased donors and Table 4.12 for the living donors. The complete variable definition is in Appendix A.

We first discuss the variables selected in the survival prediction models for pediatric recipients of deceased donor kidneys. For the graph lasso, we show variables whose average coefficients in the cross validation are less than 0.1, and there are 8 such variables. If we apply the same threshold for the group lasso model, we find 60 out of the total 145 variables. In this sense, the graph lasso is much more effective in variable selection than the group lasso. To make further comparison of the specific variables selected by the two models, for the group lasso, we raise the average coefficient threshold to 0.5 to narrow down to 13 variables. When we compare these variables with the ones selected by the graph lasso, we see 4 common ones: DEATH_MC_DON (deceased donor mechanism of death), DRUGTRT_COPD (recipient drug treated COPD (chronic obstructive pulmonary disease) at registration yes/no), EDUCATION (recipient educational level), IN-

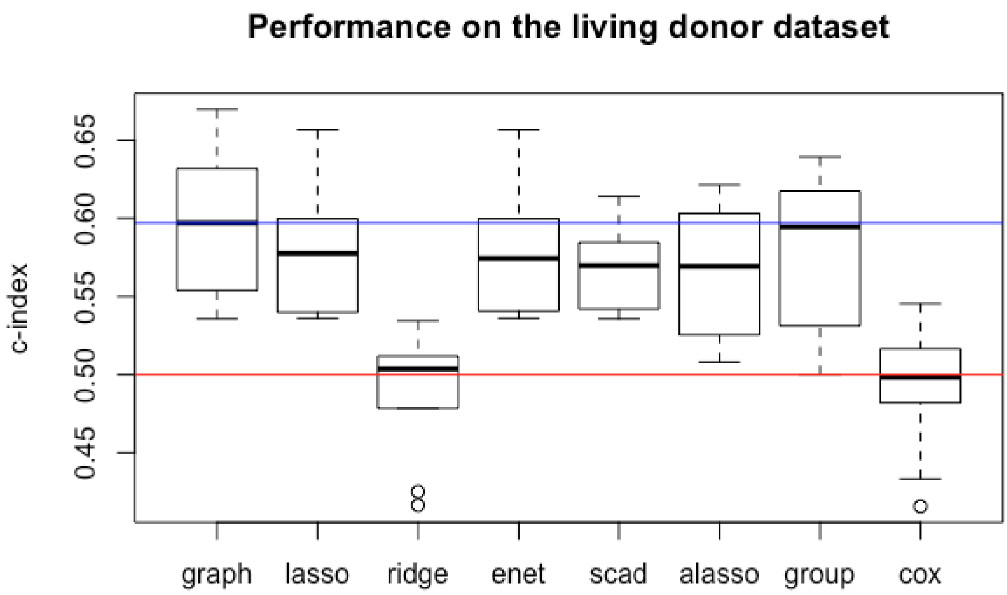


Figure 4.4: The boxplot of the model c-indices on the living donor dataset.

OTROP_SUPPORT_DON (deceased donor inotropic medication at procurement yes/no), although the specific levels selected are different (see Table 4.11).

For living donors, the graph lasso selects 31 out of the total 98 variables, and the group lasso selects 62. The variables with coefficients larger than 0.05 in both models are in Table 4.12. The commonly selected variables include CITIZENSHIP_DON (donor citizenship), DIAG_KI (recipient kidney diagnosis): TUBULAR AND INTERSTITIAL DISEASES and HCV_DON (donor Hepatitis C infection status): Unknown, and REGION (recipient UNOS region).

Notice that the variables selected by the group lasso penalty here are somewhat different from the ones in the previous chapter. This is due to the reason that, in the previous chapter, in addition to statistical properties, we also considered medical knowledge when selecting variables. Furthermore, we chose different hyper-parameter thresholds. The threshold in this chapter is chosen for the variable comparison of the graph and the group lasso, while the threshold in the last chapter was chosen so that the group lasso model could be compared with the other models in the same context. Lastly, due to the correlation among the

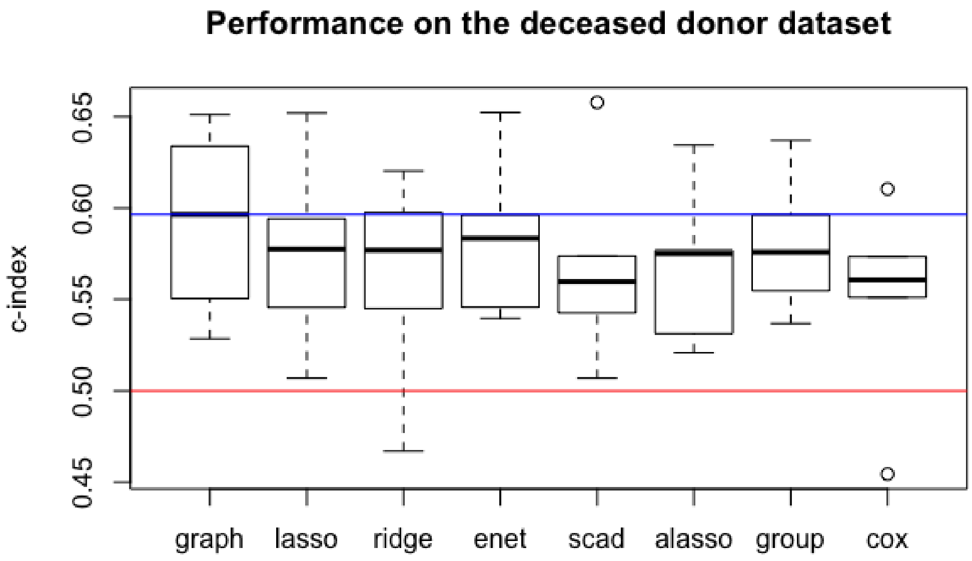


Figure 4.5: The boxplot of the model c-indices on the deceased donor dataset.

variables, it is possible that one of several variables with the same hidden cause is selected as a representative.

4.6.2 The Primary Biliary Cirrhosis Sequential (*pbcseq*) Data

The *pbcseq* data [68, 69] in the R *survival* package [70] are recorded by the Mayo Clinic to study the primary biliary cirrhosis (PBC) of the liver from 1974 to 1984. It contains the information of 1945 patients and 17 predicting variables. Definitions of the variables can be found in Table B.1 in Appendix B.

To create a predictor graph, we analyze the relations of the variables in the *pbcseq* dataset. For the numerical variables, we compute their inverse covariance (shown in Figure 4.6). We connect pairs of variables if their Pearson's test p-value is less than 0.05 [71]. The connected variable pairs are asterisked in Figure 4.6. The predictor graph of the numerical variables is illustrated in Figure 4.7. For the categorical variables, we connect variables representing different levels under the same categorical variable. As a result, we can obtain

Table 4.11: Variables selected for the pediatric kidney transplant data, deceased donors.

Variable	Baseline level	Selected level	Graph lasso	Group lasso
ANY_DIAL	No	Unknown	✓	
CARDARREST_NEURO	No	Unknown	✓	
DEATH_MC_DON	ASPHYXIATION to ANOXIA	BLUNT.INJURY.to.HEAD.TRAUMA	✓	
		BLUNT.INJURY.to.other.causes.of.death		✓
		CARDIOVASCULAR.to.other.causes.of.death		✓
DIABETES_DON	No	Unknown		✓
DRUGTRT_COPD	No	Yes		✓
		Unknown	✓	
EDUCATION	None	ATTENDED.COLLEGE.TECHNICAL.SCHOOL		✓
		POST.COLLEGE.GRADUATE.DEGREE		✓
		GRADE.SCHOOL 0.8	✓	
END_STAT_KI	KI: Active (1)	KI: Active critical status (6)		✓
		KI: Old temporarily inactive (7)		✓
ETHCAT	White	American Indian/Alaska Native		✓
		Native Hawaiian/otherPacificIslander		✓
		Unknown		✓
EXH_PERIT_ACCESS	No	Yes		✓
HCV_SEROSTATUS	No	Not Done	✓	
INOTROP_SUPPORT_DON	No	Unknown	✓	✓
PT_DIURETICS_DON	No	Unknown	✓	

Table 4.12: Variables selected for the pediatric kidney transplant data, living donors.

Variable	Baseline level	Selected level	Graph lasso	Group lasso
CITIZENSHIP_DON	US Citizen	Resident Alien	✓	
		Non Resident Alien		✓
CMV	No	Not Done		✓
DIAG_KI	GLOMERULAR DISEASES	TUBULAR AND INTERSTITIAL DISEASES	✓	✓
		POLYCYSTIC KIDNEYS		✓
		RENOVASCULAR AND OTHER VASCULAR DISEASES		✓
		NEOPLASMS		✓
EXH_PERIT_ACCESS	No	Yes		✓
EXH_VASC_ACCESS	No	Yes		✓
HCV_DON	Not Infected	Unknown	✓	✓
REGION	1	2	✓	
		5		✓
		6		✓
		7		✓
		10		✓

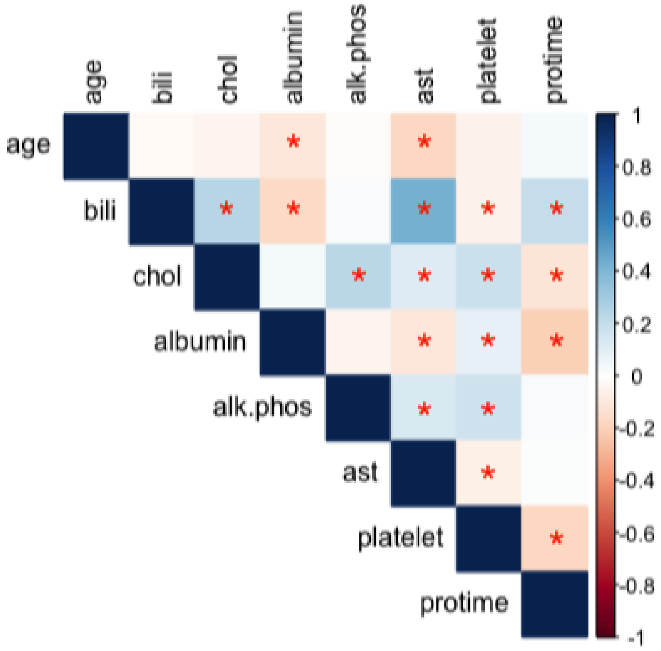


Figure 4.6: The inverse covariance of the numerical variables in the *pbcseq* dataset. The asterisked (*) variable pairs are significantly dependent.

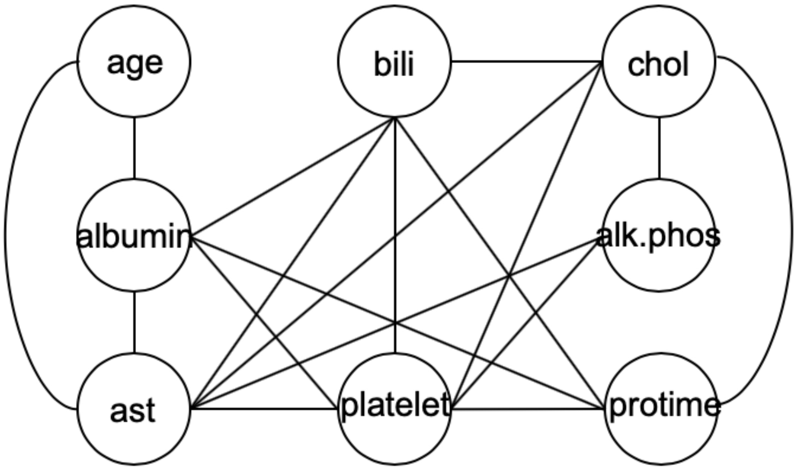


Figure 4.7: The predictor graph of the numerical variables in the *pbcseq* dataset.

the variable neighborhood relations in Table 4.13.

We compare the performance of the graph penalty to the other penalties using 10-fold cross validation on the *pbcseq* dataset. Since this is a real data problem and the true parameters are unknown to us, only the c-index can be computed. The results are shown in Table

Table 4.13: Numerical variables and their neighbors in the *pbcseq* dataset.

Variable	Neighbors
age	albumin, ast
bili	chol, albumin, ast, platelet, protime
chol	bili, alk.phos, ast, platelet, protime
albumin	age, bili, ast, platelet, protime
alk.phos	chol, ast, platelet
ast	age, bili, chol, albumin, alk.phos, platelet
platelet	bili, chol, albumin, alk.phos, ast, protime
protime	bili, chol, albumin, platelet

4.14 and Figure 4.8, where the blue reference line in the figure is the median of the graph lasso c-index.

As shown in Table 4.14, the graph lasso has the highest c-index on the *pbcseq* dataset. The ridge regression, the elastic net, and the SCAD penalties also have good performance. The boxplot shows that, the graph lasso penalty has the highest median c-index. The ridge regression and the elastic net have about the same median c-index as the graph lasso, but their distributions of the c-index are lower than the graph lasso.

Therefore, we can conclude that the graph lasso penalty has satisfactory performance on the *pbcseq* dataset, although its performance improvement is limited by the fact that the problem is not high-dimensional ($p = 17$) and the graphical structure among the variables is relatively simple.

Table 4.14: The performance of different penalties on the *pbcseq* dataset.

Model	c-index
Graph lasso	**0.88** (0.086)
Lasso	0.86 (0.082)
Ridge regression	0.87 (0.092)
Elastic net	0.87 (0.085)
SCAD	0.87 (0.079)
Alasso	0.86 (0.088)
Group lasso	0.86 (0.076)
Cox without penalty	0.83 (0.098)
$\hat{\boldsymbol{\beta}} = \mathbf{0}$	0.50 (0.00)

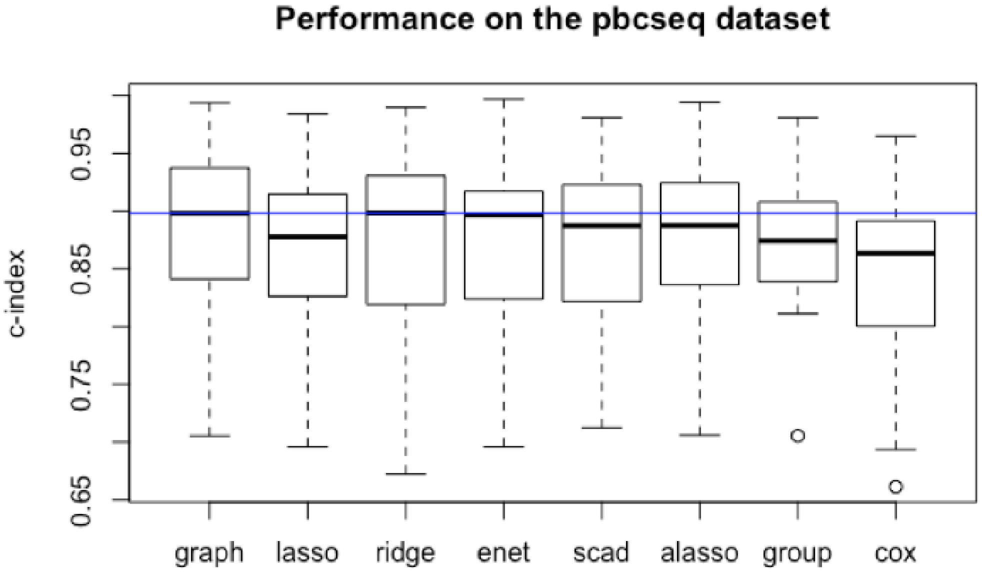


Figure 4.8: The boxplot of the model c-indices on the *pbcseq* dataset.

4.7 Discussion

In this chapter, we developed a graph-based penalty for the Cox proportional hazard model, called the graph lasso. It takes the advantage of the structure of predictors and makes effective variable selection by selecting correlated predictors together. The predictor relations can be determined numerically or by domain knowledge, and the relations are summarized in a graph, where correlated predictors are connected by edges. Hence the name graph lasso. We formulated the graph penalized problem and decomposed the penalty term to transform the problem into one that could be solved using existing solvers of the group lasso problem. Essentially, by using the predictor graph, we re-defined the "groups" in the group lasso problem. Furthermore, one predictor can belong to multiple "groups" at a time, and the overall predictor structure is much more complex than in group lasso. We can flexibly experiment with the predictor relations by making changes to the predictor graph, and find out the effect on the model's estimation and prediction accuracy.

We demonstrated the theoretical performance guarantee of the proposed graph lasso penalty, and showed its efficacy using simulation as well as real data examples. In both simulation and real data studies, the graph lasso showed overall the most promising performance when compared to the other prevalently used penalties for the Cox model.

A direction worthy of further study is developing a computationally more efficient algorithm for solving the graph lasso problem, which can be costly to solve when the problem dimension becomes extremely high.

Variable selection has been an especially critical task in survival analysis, where the variable dimension is often high and the variable relations are complicated. The problems of correlated variables and paired variables often arise in survival studies. The introduction of new regularization methods such as the graph lasso could contribute to solving these problems in a more elegant way.

CHAPTER 5
CONCLUSION

The thesis focuses on the statistical detection and survival analysis for some complicated data types, including network data, censored data, and graphical data. In Chapter 2, we propose to use the graph-scan statistic to detect the local change in a sequence of network data. We develop a parametric statistic and a non-parametric statistic, derive the theoretical false-alarm rates, and apply the statistic to detect a subgraph change in a sequence of seismic sensor networks. In Chapter 3, we study survival analysis in an applied healthcare problem of pediatric kidney transplant, and the goal is to make prediction for the survival time until an event happens. Survival analysis can be seen as a type of change-detection, where the change or abnormality happens when the survival event of interest takes place. The data in survival analysis are usually recorded as censored data, and dedicated models, such as the Cox model and the random survival forest model, are used for analyzing censored data. When analyzing the pediatric kidney transplant data, we notice that the data are high-dimensional, and many variables are correlated or paired. The predictor structure inspires us to develop a new variable selection method for the Cox model, which we call the graph lasso. In Chapter 4, we develop the graph-based lasso penalty formulation, derive its performance guarantees, and compare it with some existing penalties in simulation and real data examples.

Complicated data types are not only challenges in statistical problems, but also new opportunities for theory and methodology development. In a world where the data could be noisy, missing, censored, high-dimensional, correlated, paired, or graphical, and where all models the greatest statisticians ever proposed could be *wrong*, we strive to make some *useful* statistical estimation, inference, and prediction and contribute a tiny bit to homo sapiens' understanding of the world.

Appendices

APPENDIX A

CHAPTER 3: PEDIATRIC KIDNEY TRANSPLANT SURVIVAL ANALYSIS USING STATISTICAL MACHINE LEARNING

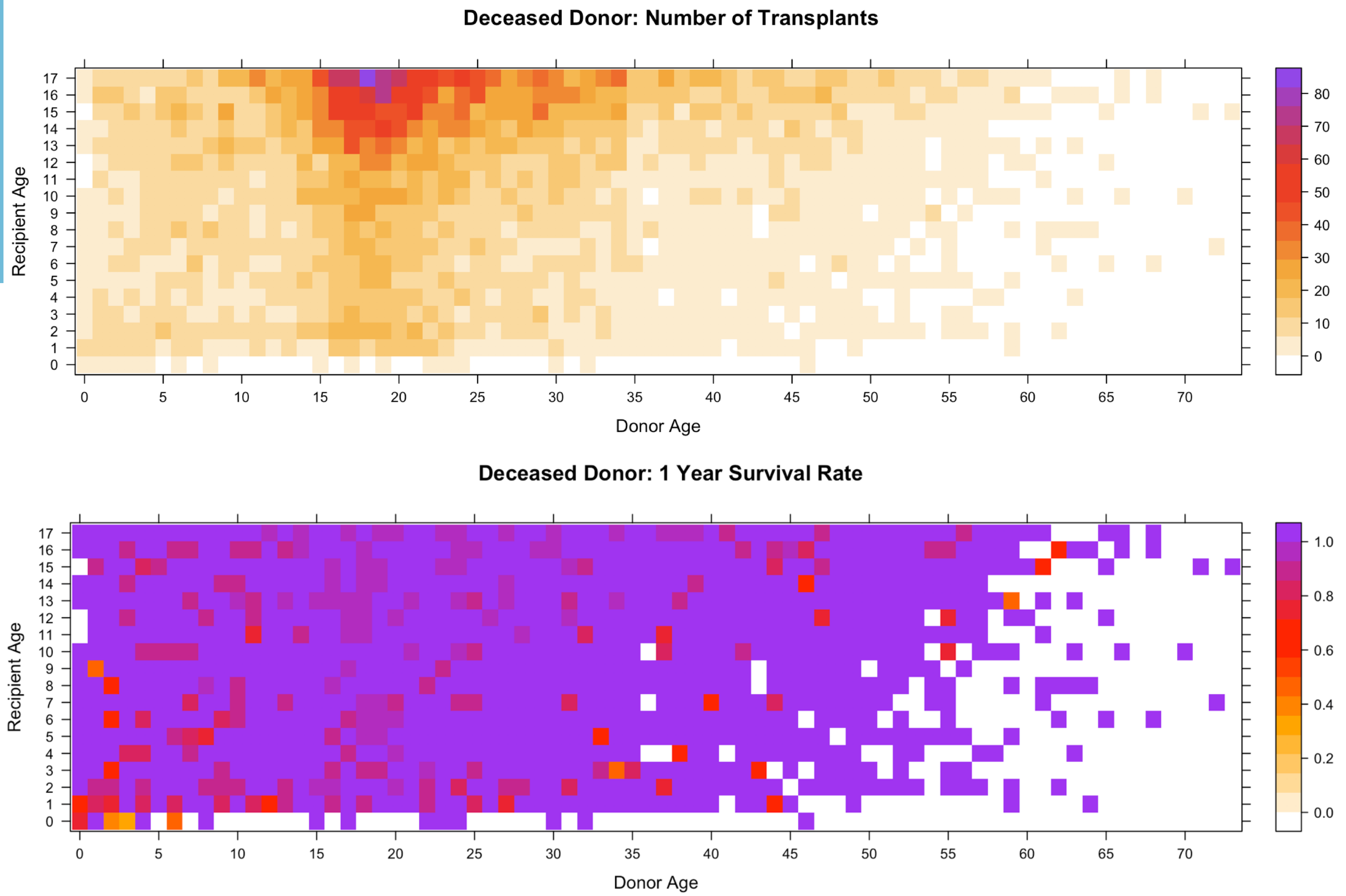


Figure A.1: Heat maps for deceased donors by the recipient age and the donor age. Top: transplant frequencies. Bottom: 1-year Kaplan-Meier (KM) empirical survival probabilities.

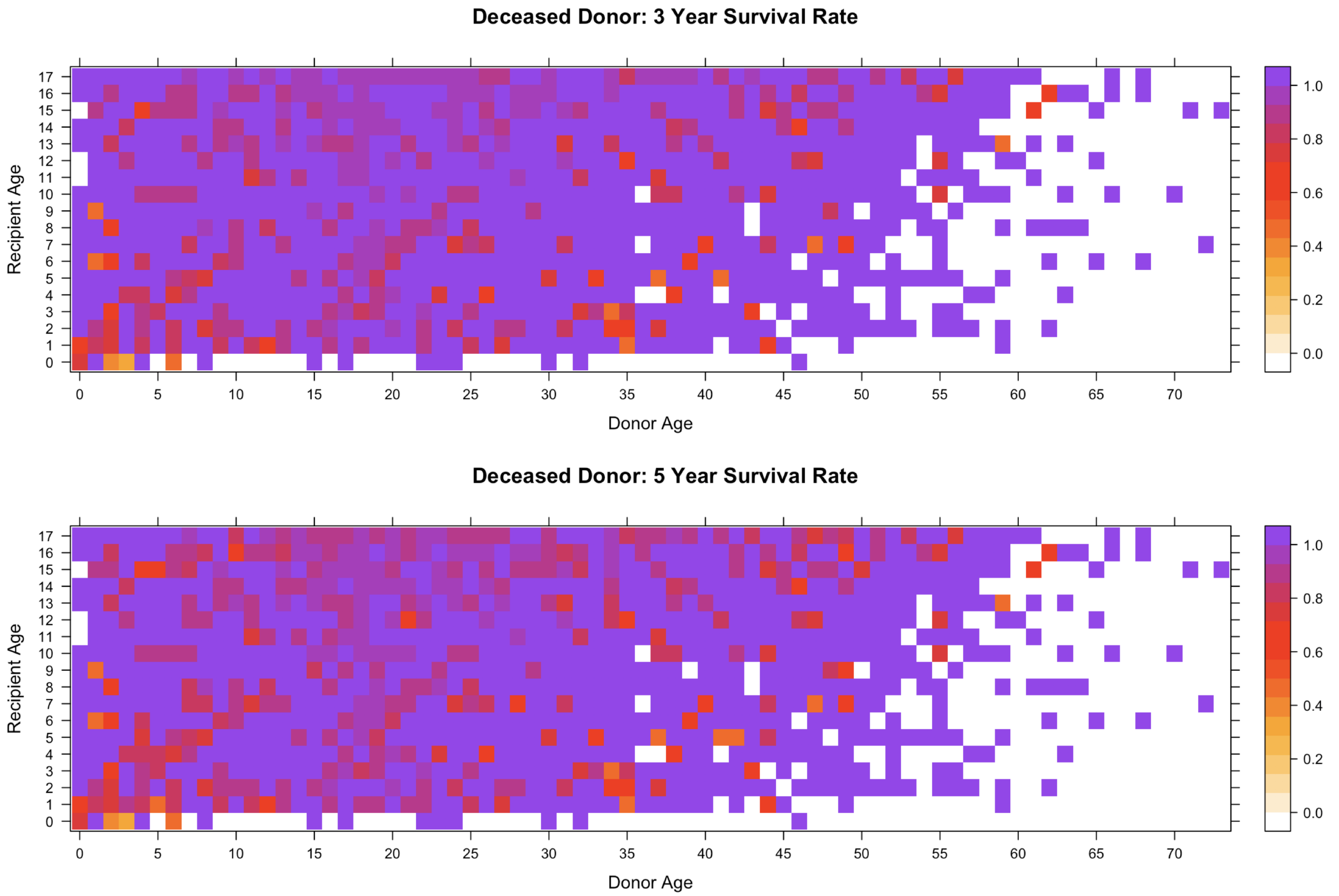


Figure A.2: Heat maps for deceased donors by the recipient age and the donor age (cont.). Top: 3-year Kaplan-Meier (KM) empirical survival probabilities. Bottom: 5-year Kaplan-Meier (KM) empirical survival probabilities.

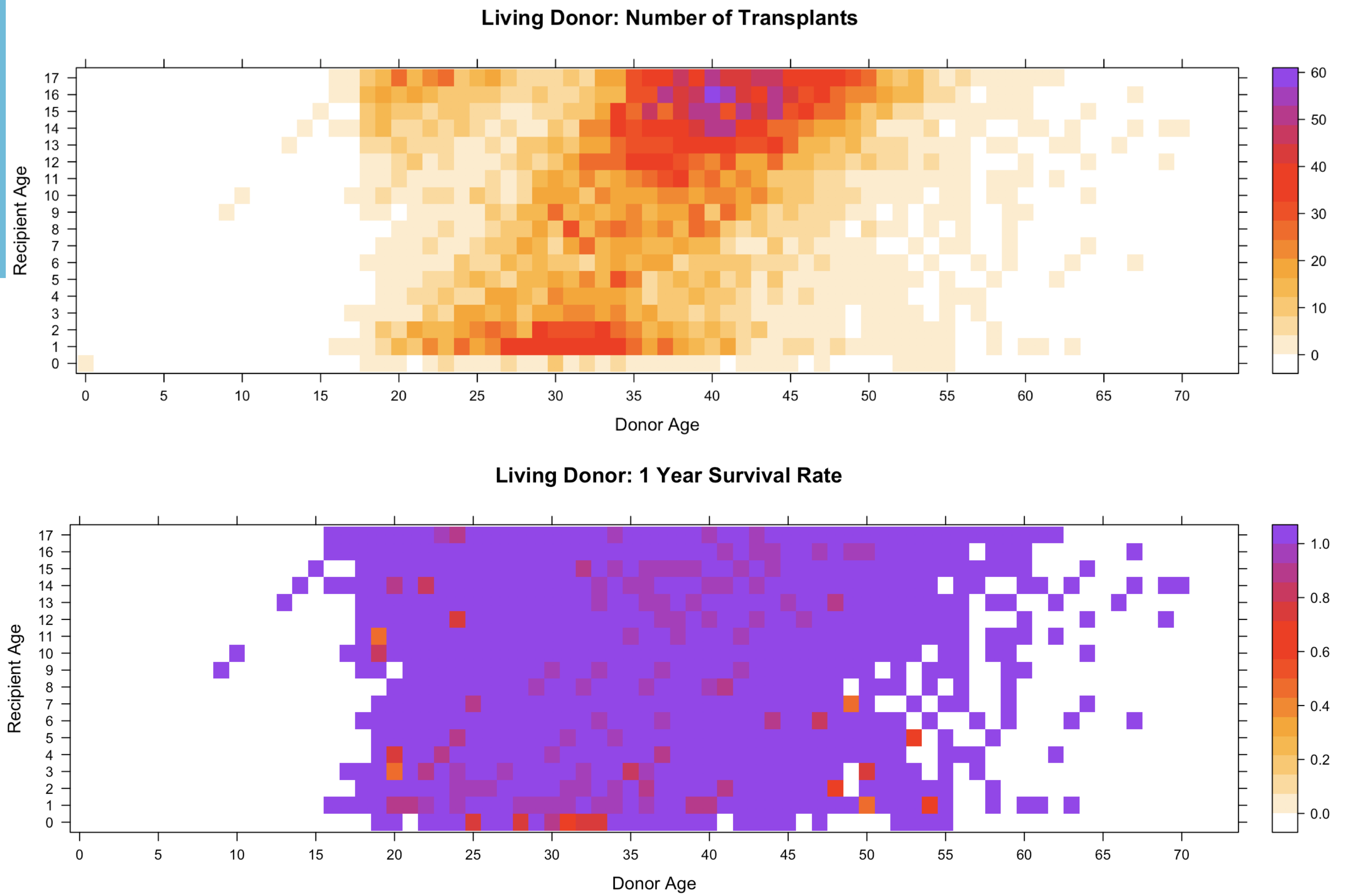


Figure A.3: Heat maps for living donors by the recipient age and the donor age. Top: transplant frequencies. Bottom: 1-year Kaplan-Meier (KM) empirical survival probabilities.

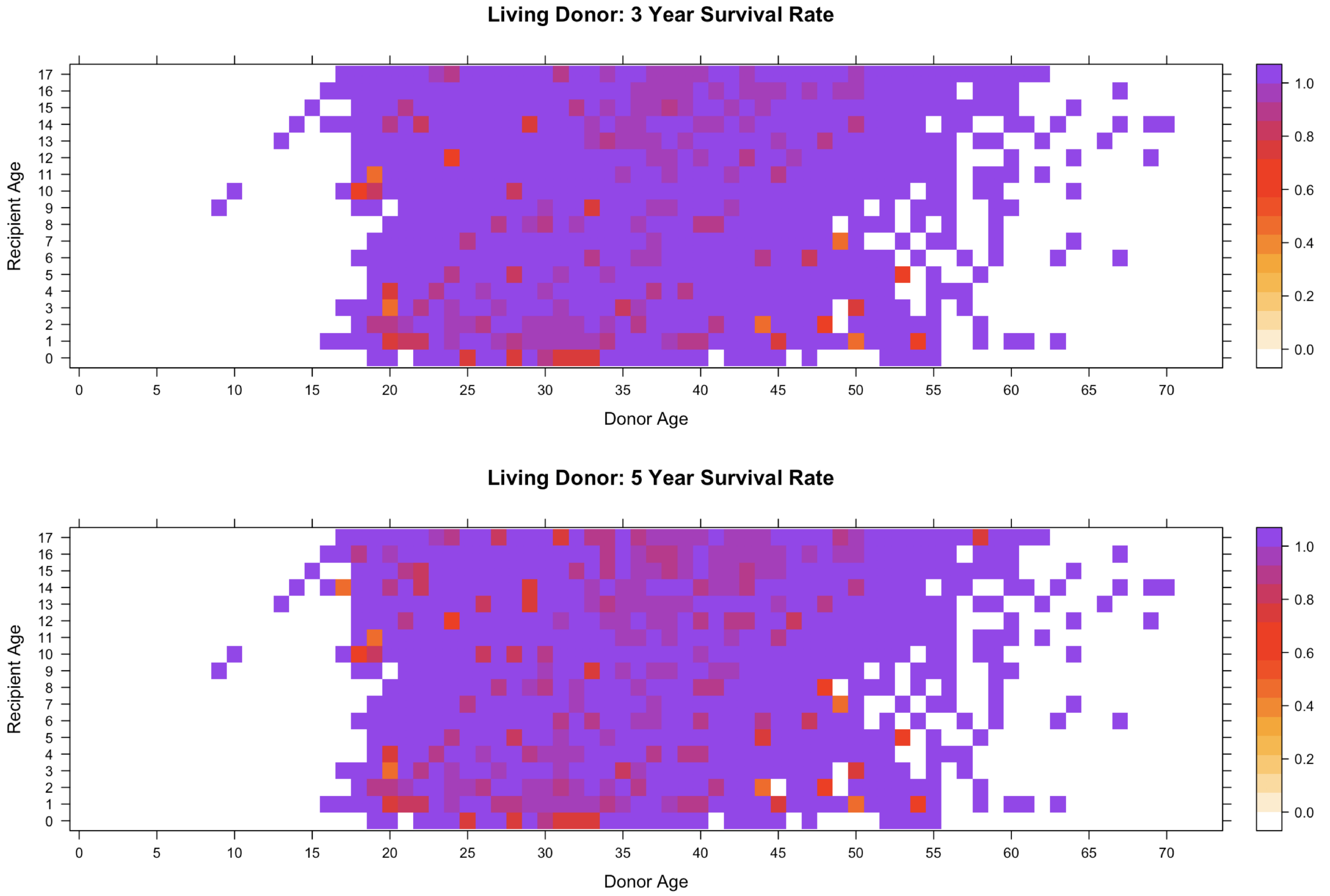


Figure A.4: Heat maps for living donors by the recipient age and the donor age (cont.). Top: 3-year Kaplan-Meier (KM) empirical survival probabilities. Bottom: 5-year Kaplan-Meier (KM) empirical survival probabilities.

Table A.1: Feature engineering details for the pediatric kidney transplant dataset.

Create new variables:

1. TX_YR (transplant year). It captures information in TX_DATE, transplant date.
2. YRS_DIAL (years on dialysis). It is determined by subtracting DIAL_DATE (dialysis date) from TX_DATE (transplant date) and rounding it to the nearest year.

Combine variables measuring the same condition of a transplant recipient or donor:

1. ANY_DIAL (any dialysis prior to the transplant). It combines DIAL_TRR (dialysis at transplant) and DIAL_TCR (dialysis at registration), and is “Y” if either of the two is “Y”, and “N” otherwise.
2. ANY_PRIVATE (any private insurance utilized for the transplant). It combines PRI_PAYMENT_TRR_KI, SECONDARY_PAY_TRR_KI, PRI_PAYMENT_TCR_KI, and SECONDARY_PAY_TCR_KI, and is “Y” if any of the them is “Y”, and “N” otherwise.
3. CMV (cytomegalovirus infection status). It combines two lab test results: CMV_IGG and CMV_IGM, and is “Y” if either of the two is “Y”, and “N” otherwise.
4. HBV (recipient hepatitis B virus infection status). It combines the lab test results of HBV_CORE and HBV_SUR_ANTIGEN following the guidelines from the CDC [72].
5. HBV_DON (donor hepatitis B virus infection status). It combines HBV_CORE_DON, HBSAB_DON, and HBV_SUR_ANTIGEN_DON following the guidelines from the CDC [72].

Living donor specific variables:

1. CMV_DON_L (living donor CMV status). It combines the test results of CMV_IGG_DON, CMV_IGM_DON, CMV_OLD_LIV_DON, and CMV_NUCLEIC_DON. CMV_DON_L is positive if any of the variables is positive. Otherwise the CMV_IGG_DON test result is used. (The deceased donor dataset has the variable: CMV_DON, which is missing for the majority of samples in the living donor dataset).
2. HCV_DON (living donor hepatitis C virus infection status). It combines HCV_ANTIBODY_DON, HCV_RIBA_DON, and HCV_RNA_DON following the guidelines from the CDC [73].

...continued table

Group the categories of some categorical variables

1. DIAG_KI (recipient kidney diagnosis). It originally contains 69 categories. We group the variables according to OPTN specifications [74], and remove categories with sample size smaller than 10. The number of categories in DIAG_KI are reduced to 10.

2. FUNC_STAT_TRR (recipient functional status at transplant). We group the functional status categories and convert them to numerical values that reflect the wellness of the transplant recipient. The conversion from categorical functional status to numerical values are in Table A.2.

Living donor specific variables:

1. LIV_DON_TY (living donor type). It originally has 15 categories, and we group the categories into 4. Some of the categories have limited number of samples and we group them with other similar categories. The new categories are Bio-Other, Bio-Parent, Bio-Sibling, and Non-Biological.

2. REGION_DON (living donor geographical region). It maps HOME_STATE_DON to the corresponding UNOS region.

Table A.2: Combining the categories of FUNC_STAT_TRR (functional status at transplant).

Grouped Values	**Original Values**
10	10% - Moribund, fatal processes progressing rapidly
10	10% - No play; does not get out of bed
20	20% - Very sick, hospitalization necessary: active treatment necessary
20	20% - Often sleeping; play entirely limited to very passive activities
30	30% - Severely disabled: hospitalization is indicated, death not imminent
30	30% - In bed; needs assistance even for quiet play
40	40% - Disabled: requires special care and assistance
40	40% - Mostly in bed; participates in quiet activities
50	Performs activities of daily living with TOTAL assistance.
50	50% - Requires considerable assistance and frequent medical care
50	50% - Can dress but lies around much of day; no active play; can take part in quiet play/activities
60	Performs activities of daily living with SOME assistance.
60	60% - Requires occasional assistance but is able to care for needs
60	60% - Up and around, but minimal active play; keeps busy with quieter activities
70	70% - Cares for self: unable to carry on normal activity or active work
70	70% - Both greater restriction of and less time spent in play activity
80	80% - Normal activity with effort: some symptoms of disease
80	80% - Active, but tires more quickly
90	90% - Able to carry on normal activity: minor symptoms of disease
90	90% - Minor restrictions in physically strenuous activity
100	Performs activities of daily living with NO assistance.
100	100% - Normal, no complaints, no evidence of disease
100	100% - Fully active, normal

Table A.3: Features in the **deceased** donor dataset after data pre-processing. These 84 features are the candidate features for statistical variable selection.

Features for both the recipient and the donor (27 features):

1. AGE (recipient age), AGE_DON (donor age)
2. AMIS (recipient and donor HLA (Human Leukocyte Antigen) -A locus mismatch status), BMIS (HLA-B locus mismatch status), HLAMIS (HLA mismatch status), DRMIS (HLA-DR mismatch status)
3. BMI_CALC (recipient calculated BMI (Body Mass Index)), BMI_DON_CALC (donor calculated Body Mass Index)
4. CITIZENSHIP (recipient citizenship), CITIZENSHIP_DON (donor citizenship).
5. CMV (recipient CMV (cytomegalovirus) infection status), CMV_DON_L (donor CMV infection status).
6. CREAT_TRR (recipient creatinine level at transplant), CREAT_DON (donor creatinine level)
7. DIAB (recipient diabetes status), DIABETES_DON (donor diabetes status)
8. ETHCAT (recipient ethnicity), ETHCAT_DON (donor ethnicity)
9. GENDER (recipient gender), GENDER_DON (donor gender)
10. HCV_SEROSTATUS (recipient hepatitis C infection status), HCV_C_ANTI_DON (donor hepatitis C infection status)
11. PERM_STATE (recipient state of residency), HOME_STATE_DON (donor home state)
12. REGION (recipient UNOS region), REGION_DON (donor UNOS region)
13. SHARE_TY (organ share type, i.e. local, regional, or national)

...continued table

Recipient features (21 features):

1. ANY_DIAL (dialysis yes/no)
2. ANY_PRIVATE (any private insurance available yes/no)
3. CURRENT_PRA (current PRA (Panel Reactive Antibodies) level)
4. DAYSWAIT_CHRON_KI (total days on the kidney waiting list)
5. DIAG_KI (kidney diagnosis)
6. DRUGTRT_COPD (drug treated COPD (chronic obstructive pulmonary disease) at registration yes/no)
7. EDUCATION (education level)
8. END_STAT_KI (kidney status at the time of the transplant)
9. EXH_PERIT_ACCESS (exhausted vascular access at registration yes/no)
10. EXH_VASC_ACCESS (exhausted peritoneal access at registration yes/no)
11. FUNC_STAT_TRR (functional status at transplant)
12. HBV (hepatitis B infection status)
13. MALIG (any previous malignancy yes/no)
14. MED_COND_TRR (medical condition at transplant)
15. NPKID (number of previous kidney transplants)
16. PAYBACK (transplant as the result of a payback yes/no)
17. PERIP_VASC (peripheral vascular disease at registration yes/no)
18. PREV_TX (previous kidney transplant yes/no)
19. TX_PROCEDUR_TY_KI (transplant procedure type, i.e. left or right kidney)
20. TX_YR (transplant year)
21. YRS_DIAL (years on dialysis)

...continued table

Donor features (36 features):

1. ANTICONV_DON (deceased donor - anticonvulsants within 24 hours pre-cross clamp yes/no)
2. ANTIHYPE_DON (deceased donor - antihypertensives within 24 hours pre-cross clamp yes/no)
3. BLOOD_INF_DON (deceased donor - blood as infection status yes/no)
4. BUN_DON (deceased donor - terminal blood urea nitrogen level)
5. CANCER_SITE_DON (deceased donor - cancer site yes/no)
6. CARDARREST_NEURO (deceased donor - cardiac arrest post brain death yes/no)
7. CLIN_INFECT_DON (deceased donor - clinical infection yes/no)
8. COLD_ISCH_KI (kidney cold ischemic time)
9. DDAVP_DON (deceased donor - synthetic anti diuretic hormone (DDAVP) yes/no)
10. DEATH_CIRCUM_DON (deceased donor - circumstance of death)
11. DEATH_MC_DON (deceased donor - mechanism of death)
12. DISTANCE (miles from donor hospital to transplant center)
13. DON_RETYP (deceased donor - retyped at transplant center yes/no)
14. HIST_CANCER_DON (deceased donor - history of cancer yes/no)
15. HIST_CIG_DON (deceased donor - history of cigarettes in past $>$ 20 pack years yes/no)
16. HIST_COCAINE_DON (deceased donor - history of cocaine use in the past yes/no)
17. HIST_HYPERTENS_DON (deceased donor - history of hypertension yes/no)
18. HIST_OTH_DRUG_DON (deceased donor - history of other drug use in the past yes/no)
19. INOTROP_AGENTS (deceased donor - inotropic agent support yes/no)

...continued table

20. INOTROP_SUPPORT_DON (deceased donor - inotropic medication at procurement yes/no)
21. INTRACRANIAL_CANCER_DON (deceased donor - intracanial cancer at procurement yes/no)
22. LT_KI_BIOPSY (deceased donor - left kidney biopsy at recovery yes/no)
23. NON_HRT_DON (deceased donor - non heart beating donor yes/no)
24. PROTEIN_URINE (deceased donor - protein in urine yes/no)
25. PT_DIURETICS_DON (deceased donor - diuretics B/N brain death within 24 hours of procurement yes/no)
26. PT_STEROIDS_DON (deceased donor - steroids B/N brain death within 24 hours of procurement yes/no)
27. PT_T3_DON (deceased donor - triiodothyronine-t3 B/N brain death within 24 hours of procurement yes/no)
28. PT_T4_DON (deceased donor - thyroxine-t4 B/N brain death within 24 hours of procurement yes/no)
29. PULM_INF_DON (deceased donor - infection pulmonary source yes/no)
30. RT_KI_BIOPSY (deceased donor - right kidney biopsy at recovery yes/no)
31. SGOT_DON (deceased donor - terminal SGOT/AST level)
32. SGPT_DON (deceased donor - terminal SGPT/ALT level)
33. TATTOOS (deceased donor - tattoos yes/no)
34. TBILI_DON (deceased donor - terminal total bilirubin level)
35. URINE_INF_DON (deceased donor - infection urine source yes/no)
36. VASODIL_DON (deceased donor - vasodilators within 24 hours pre-cross clamp yes/no)

Table A.4: Features in the **living** donor dataset after data pre-processing. These 40 features are the candidate features for statistical variable selection.

Features for both the recipient and the donor (18 features):

1. AGE (age), AGE_DON (donor age)
2. HLAMIS (HLA (Human Leukocyte Antigen) mismatch level)
3. CITIZENSHIP (recipient citizenship), CITIZENSHIP_DON (donor citizenship)
4. CMV (recipient CMV (cytomegalovirus) infection status), CMV_DON_L (CMV (cytomegalovirus) infection status)
5. ETHCAT (recipient ethnicity), ETHCAT_DON (donor ethnicity)
6. GENDER (recipient gender), GENDER_DON (donor gender)
7. HAPLO_TY_MATCH_DON (living donor-recipient haplo type match)
8. HBV (recipient hepatitis B infection status), HBV_DON (donor hepatitis B infection status)
9. HCV_SEROSTATUS (recipient hepatitis C infection status), HCV_DON (donor hepatitis C infection status)
10. REGION (recipient UNOS region), REGION_DON (donor UNOS region)

... continued table

Recipient features (19 features):

1. ANY_DIAL (dialysis yes/no)
2. ANY_PRIVATE (any private insurance available)
3. BMI_CALC (calculated BMI (Body Mass Index))
4. CREAT_TRR (creatinine level at transplant)
5. DATA_WAITLIST (waitlist data reported yes/no)
6. DIAB (diabetes status)
7. DIAG_KI (kidney diagnosis)
8. DRUGTRT_COPD (drug treated COPD (chronic obstructive pulmonary disease) at registration)
9. EXH_PERIT_ACCESS (exhausted vascular access at registration yes/no)
10. EXH_VASC_ACCESS (exhausted peritoneal access at registration yes/no)
11. FUNC_STAT_TRR (functional status at transplant)
12. MALIG (any previous malignancy)
13. MED_COND_TRR (medical condition at transplant)
14. MRCREATG (recipient most recent creatinine greater than 2mg/dl at registration)
15. PREV_TX (previous transplant of kidney yes/no)
16. PRE_TX_TXFUS (number of pre-transplant transfusions at transplant)
17. TX_PROCEDUR_TY_KI (transplant procedure type, left or right kidney)
18. TX_YR (transplant year)
19. YRS_DIAL (years on dialysis)

Donor features (3 features):

1. COLD_ISCH_KI (kidney cold ischemic time)
2. DISTANCE (miles from donor hospital to transplant center)
3. LIV_DON_TY (living donor relation to recipient)

APPENDIX B

CHAPTER 4: GRAPH BASED VARIABLE SELECTION FOR SURVIVAL ANALYSIS

B.1 Useful Lemmas

Lemma B.1.1 (Lemma A.1 in [**lounici2009taking**].)**.** *Let χ^2_d to a chi-squared random variable with d degrees of freedom, we have*

$$\mathbb{P}(\chi^2_d > d + t) \leq \exp\left(-\frac{1}{8}\min\{t, \frac{t^2}{d}\}\right).$$

Can be replaced by other forms

Lemma B.1.2 (Subgradient conditions)**.** *A vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is a solution to the optimization problem* (4.4) *if and only if $\boldsymbol{\beta}$ can be decomposed as $\boldsymbol{\beta} = \sum_{i=1}^{p} V^{(i)}$ where $V^{(i)}$ satisfy that: $\forall i$, (a) $V^{(i)}_{\mathcal{N}_i^c} = 0$; (b) either $V^{(i)}_{\mathcal{N}_i} \neq 0$ and $\frac{\partial}{\partial \boldsymbol{\beta}_{\mathcal{N}_i}}\ell(\boldsymbol{\beta}) = n\lambda\tau_i \frac{V^{(i)}_{\mathcal{N}_i}}{\|V^{(i)}_{\mathcal{N}_i}\|_2}$, or $V^{(i)}_{\mathcal{N}_i} = 0$ and $\|\frac{\partial}{\partial \boldsymbol{\beta}_{\mathcal{N}_i}}\ell(\boldsymbol{\beta})\|_2 \leq n\lambda\tau_i$.*

Proof. This is a direct result from Lemma 11 in [62]. □

Lemma B.1.3 ([59])**.** *For any predictor graph G and positive weights τ_i, suppose $V^{(1)}$, $V^{(2)}$, ..., $V^{(p)}$ is an optimal decomposition of $\boldsymbol{\beta} \in \mathbb{R}^p$, then for any $S \subseteq \{1, 2, \ldots, p\}$, $\{V^{(j)}, j \in S\}$ is also an optimal decomposition of $\sum_{j \in S} V^{(j)}$.*

B.2 Local approximation for partial likelihood function

Let $T, C, \boldsymbol{x}$ denote the survival time, censoring time and the associated covariates. Consider the general setting that the covariate may vary with time $\boldsymbol{x}(t)$. The theory of counting process can be used to express the log-likelihood function. More specifically, define $N_i(t) = 1\{T_i \leq t, T_i \leq C_i\}$ and $Y_i(t) = 1\{T_i \geq t, C_i \geq t\}$, where $1\{\cdot\}$ is the indicator function. Without loss of generality, we only consider the time horizon $[0, 1]$. The results can be extended to interval $[0, \infty)$ [63]. For completeness, here we provide the method used in [63] to express the log-likelihood function as a quadratic function in a $n^{-1/2}$ neighborhood of the true parameter $\boldsymbol{\beta}_0$. [63] has proved the consistency and asymptotic normality of the maximum likelihood estimates of the parameter $\boldsymbol{\beta}_0$ when there is no penalty term.

It can be verified that the partial likelihood equals to

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^1 \boldsymbol{\beta}^T \boldsymbol{x}_i(s) dN_i(s) - \int_0^1 \log \left\{ \sum_{i=1}^{n} Y_i(s) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(s)) \right\} d\bar{N}(s),$$

where $\bar{N} = \sum_{i=1}^{n} N_i$.

Under assumptions 4.4.1, for each $\boldsymbol{\beta}$ in a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}_0$, we have [63]:

$$\frac{1}{n}\{l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}_0)\} =$$
$$\int_0^1 \left[(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathcal{T}} s^{(1)}(\boldsymbol{\beta}_0, t) - \log \left\{ \frac{s^{(0)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}_0, t)} \right\} s^{(0)}(\boldsymbol{\beta}_0, t) \right] h_0(t) dt + O_P\left(\frac{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|}{\sqrt{n}}\right).$$

Note that the first order derivative of the right hand side equals to 0 at $\boldsymbol{\beta}_0$. By Taylor's expansion, we have

$$\frac{1}{n}\{l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}_0)\} = -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \{I(\boldsymbol{\beta}_0) + o_P(1)\}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + O_P(n^{-1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|).$$

Define the scaled objective function with penalty as $\mathcal{L}(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + n\lambda\|\boldsymbol{\beta}\|_{G,\mathcal{T}}$. For α_n, if we can show that for any given $\epsilon > 0$, there exists a large constant C such that

$$\mathbb{P}\left\{ \sup_{\|\boldsymbol{u}\|=C} \mathcal{L}(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) > \mathcal{L}(\boldsymbol{\beta}_0) \right\} \geq 1 - \epsilon.$$

Then this will implie that with probability at least $1 - \epsilon$ there exists a local minima in the ball $\{\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u} : \|\boldsymbol{u}\| \leq C\}$. Hence, there exists a local minimizer such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| =$

$O_P(\alpha_n)$. To show this, we first note that

$$\frac{1}{n}\left\{\mathcal{L}(\boldsymbol{\beta}_0) - \mathcal{L}(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u})\right\} = \frac{1}{n}\left\{l(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) - l(\boldsymbol{\beta}_0) - \lambda\left[\|\boldsymbol{\beta_0} + \alpha_n \boldsymbol{u}\|_{G,\tau} - \|\boldsymbol{\beta_0}\|_{G,\tau}\right]\right\}. \tag{B.1}$$

$$\frac{1}{n}\left\{l(\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{u}) - l(\boldsymbol{\beta}_0)\right\} = -\frac{1}{2}\alpha_n^2 \boldsymbol{u}^T\{I(\boldsymbol{\beta}_0) + o_P(1)\}\boldsymbol{u} + O_P(n^{-1/2}\alpha_n\|\boldsymbol{u}\|).$$

The first term is of the order $O(\alpha_n^2 C^2)$, and the second term is of the order $O(n^{-1/2}\alpha_n C)$.

Now look at the penalty term, since $\|\boldsymbol{\beta_0} + \alpha_n \boldsymbol{u}\|_{G,\tau} \geq \|\boldsymbol{\beta_0}\|_{G,\tau} - \alpha_n\|\boldsymbol{u}\|_{G,\tau}$, we have

$$-\lambda\left[\|\boldsymbol{\beta_0} + \alpha_n \boldsymbol{u}\|_{G,\tau} - \|\boldsymbol{\beta_0}\|_{G,\tau}\right] \leq \lambda\alpha_n\|\boldsymbol{u}\|_{G,\tau}.$$

If the term $O(\alpha_n^2 C^2)$ dominates the whole expression (B.1), then we have that there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $\mathcal{L}(\boldsymbol{\beta})$ that is close to true parameter $\boldsymbol{\beta}_0$.

B.3 Proofs

Proof to Theorem 4.4.3. Suppose $\hat{\boldsymbol{\beta}}$ is the optimal solution to the regularization problem, then for any $\boldsymbol{\beta} \in \mathbb{R}^p$, we have

$$-\frac{1}{n}l(\hat{\boldsymbol{\beta}}) + \lambda\|\hat{\boldsymbol{\beta}}\|_{G,\tau} \leq -\frac{1}{n}l(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_{G,\tau}.$$

Let $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, we have

$$\frac{1}{n}\left\{l(\boldsymbol{\beta}_0) - l(\hat{\boldsymbol{\beta}})\right\} \leq \lambda\left(\|\boldsymbol{\beta}_0\|_{G,\tau} - \|\hat{\boldsymbol{\beta}}\|_{G,\tau}\right).$$

Denote $\{S^{(1)}, S^{(2)}, \ldots, S^{(p)}\}$ as an arbitrary optimal decomposition of $\boldsymbol{\beta}_0$ and $\{T^{(1)}, T^{(2)}, \ldots, T^{(p)}\}$ as an arbitrary optimal decomposition of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$. We have $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \sum_{i=1}^{p} T^{(i)}$. Thus

$$\begin{aligned}\frac{1}{n}\left\{l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}_0)\right\} &= -\frac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\{I(\boldsymbol{\beta}_0) + o_P(1)\}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + O_P(n^{-1/2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|)\\ &= -\frac{1}{2}(\sum_{i=1}^{p} T^{(i)})^T\{I(\boldsymbol{\beta}_0) + o_P(1)\}(\sum_{i=1}^{p} T^{(i)}) + O_P(n^{-1/2}\|\sum_{i=1}^{p} T^{(i)}\|).\end{aligned}$$

By Assumption 4.4.2 (1), we have $S^{(j)} = 0, \forall j \in J_0^c$, thus

$$\|\boldsymbol{\beta}_0\|_{G,\tau} = \|\sum_{j\in J_0} S^{(j)}\|_{G,\tau},$$

and

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}\|_{G,\tau} &= \|\sum_{j\in J_0} T^{(j)} + \sum_{j\notin J_0} T^{(j)} + \sum_{j\in J_0} S^{(j)}\|_{G,\tau}\\ &\geq \|\sum_{j\notin J_0} T^{(j)} + \sum_{j\in J_0} S^{(j)}\|_{G,\tau} - \|\sum_{j\in J_0} T^{(j)}\|_{G,\tau}\\ &= \|\sum_{j\notin J_0} T^{(j)}\|_{G,\tau} + \|\sum_{j\in J_0} S^{(j)}\|_{G,\tau} - \|\sum_{j\in J_0} T^{(j)}\|_{G,\tau}.\end{aligned}$$

Therefore

$$\|\boldsymbol{\beta}_0\|_{G,\tau} - \|\hat{\boldsymbol{\beta}}\|_{G,\tau} \leq \|\sum_{j\in J_0} T^{(j)}\|_{G,\tau} - \|\sum_{j\notin J_0} T^{(j)}\|_{G,\tau} \leq \|\sum_{j\in J_0} T^{(j)}\|_{G,\tau} = \sum_{j\in J_0} \tau_j\|T^{(j)}\|_2, \tag{B.2}$$

where the final step is by Lemma 2, and

$$\|\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\|_{G,\tau} = \|\sum_{j\in J_0} T^{(j)}\|_{G,\tau} + \|\sum_{j\notin J_0} T^{(j)}\|_{G,\tau}. \tag{B.3}$$

By the definition of $K_{G,\tau}$, we have

$$\frac{1}{n}\left\{l(\boldsymbol{\beta}_0) - l(\hat{\boldsymbol{\beta}})\right\} \leq \lambda \sum_{j\in J_0} \tau_j\|T^{(j)}\|_2 \leq \lambda K_{G,\tau}^{1/2}\sqrt{\sum_{j\in J_0} \tau_j^2\|T^{(j)}\|_2^2}.$$

On the other hand, from the Assumption 4.4.2 (2), we have

$$\frac{1}{n}\left\{l(\boldsymbol{\beta}_0) - l(\hat{\boldsymbol{\beta}})\right\} \geq \kappa \sum_{j=1}^{p} \tau_j^2 \|T^{(j)}\|_2^2 \geq \kappa \sum_{j\in J_0} \tau_j^2 \|T^{(j)}\|_2^2.$$

Combine with the aforementioned equation, we have

$$\frac{1}{n}\left\{l(\boldsymbol{\beta}_0) - l(\hat{\boldsymbol{\beta}})\right\} \leq \lambda K_{G,\tau}^{1/2} \sqrt{\sum_{j\in J_0} \tau_j^2 \|T^{(j)}\|_2^2} \leq \frac{\lambda K_{G,\tau}^{1/2}}{\sqrt{\kappa}} \sqrt{\frac{1}{n}\left\{l(\boldsymbol{\beta}_0) - l(\hat{\boldsymbol{\beta}})\right\}},$$

which translates into the following:

$$\frac{1}{n}\left\{l(\boldsymbol{\beta}_0) - l(\hat{\boldsymbol{\beta}})\right\} \leq \frac{\lambda^2 K_{G,\tau}}{\kappa}.$$

Furthermore,

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = \|\sum_{j=1}^{p} T^{(j)}\|_2 \leq \frac{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_{G,\tau}}{\tau^*} &= \frac{\sum_{j=1}^{p} \tau_j \|T^{(j)}\|_2}{\tau^*} \\
&\leq \frac{\sqrt{p} \cdot \sqrt{\sum_{j=1}^{p} \tau_j^2 \|T^{(j)}\|_2^2}}{\tau^*} \leq \frac{\sqrt{p}}{\sqrt{\kappa}\tau^*} \cdot \sqrt{\frac{1}{n}\left\{l(\boldsymbol{\beta}_0) - l(\hat{\boldsymbol{\beta}})\right\}} \\
&\leq \frac{\lambda\sqrt{pK_{G,\tau}}}{\kappa\tau^*}.
\end{aligned}$$

□

Proof to Theorem 4.4.4. For each $\boldsymbol{u} \in \mathbb{R}^p$, define $Q_n(\boldsymbol{u}) = -l(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}) + n\lambda\|\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}\|_{G,\tau}$. Then we have

$$\hat{\boldsymbol{u}} = \sqrt(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \arg\min_{\boldsymbol{u}\in\mathbb{R}^p} Q_n(\boldsymbol{u}).$$

We also have

$$\begin{aligned} Q_n(\boldsymbol{u}) - Q_n(0) &= l(\boldsymbol{\beta}_0) - l(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}) + n\lambda(\|\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}\|_{G,\tau} - \|\boldsymbol{\beta}_0\|_{G,\tau}) \\ &= \frac{1}{2}\boldsymbol{u}^T I(\boldsymbol{\beta}_0)\boldsymbol{u} + o_P(1) + n\lambda(\|\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}\|_{G,\tau} - \|\boldsymbol{\beta}_0\|_{G,\tau}) \end{aligned}$$

For the second term, we have

$$n\lambda(\|\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}\|_{G,\tau} - \|\boldsymbol{\beta}_0\|_{G,\tau}) = n\lambda(\|(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}})_{J_0}\|_{G,\tau} - \|\boldsymbol{\beta}_0\|_{G,\tau} + \|(\frac{\boldsymbol{u}}{\sqrt{n}})_{J_0^c}\|_{G,\tau})$$

Suppose $V^{(1)}, \dots, V^{(p)}$ is an optimal decomposition of $\boldsymbol{u}$, then we have

$$n\lambda(\|(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}})_{J_0}\|_{G,\tau} - \|\boldsymbol{\beta}_0\|_{G,\tau}) \leq \sqrt{n}\lambda \sum_{j \in J_0} \tau_j \|V^{(j)}\|_2.$$

If $\sqrt{n}\lambda \to 0$ and $\tau_j = O(1)$ for each $j \in J_0$, then for each fixed $\boldsymbol{u}$, we have

$$n\lambda(\|(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}})_{J_0}\|_{G,\tau} - \|\boldsymbol{\beta}_0\|_{G,\tau}) \to 0, \text{ as } n \to \infty.$$

If $n^{\gamma+1/2}\lambda \to \infty$, $\boldsymbol{u}_{J_0^c} \neq 0$, and $\liminf_{n\to\infty} n^{-\gamma/2}\tau_j > 0$ for each $j \in J_0^c$, then

$$n\lambda\|(\frac{\boldsymbol{u}}{\sqrt{n}})_{J_0^c}\|_{G,\tau} = \sqrt{n}\lambda \sum_{j \in J_0^c} \tau_j \|V^{(j)}\|_2 = n^{\gamma+1/2}\lambda \cdot n^{-\gamma/2} \sum_{j \in J_0^c} \tau_j \|V^{(j)}\|_2 \to \infty.$$

Hence, we have

$$Q_n(\boldsymbol{u}) - Q_n(0) \xrightarrow{d} \begin{cases} l(\boldsymbol{\beta}_0) - l(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}) & \text{supp}(\boldsymbol{u}) \subset J_0 \\ \infty & \text{o.w.} \end{cases}$$

This implies that $\hat{\boldsymbol{\beta}}_{J_0^c} \xrightarrow{d} 0$. We note that $\hat{\boldsymbol{u}} = \arg\min Q_n(\boldsymbol{u}) = \arg\min Q_n(\boldsymbol{u}) - Q_n(0)$, thus it suffices to show that the $\hat{\boldsymbol{u}} = \arg\max_{\text{supp}(\boldsymbol{u}) \subset J_0} l(\boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}})$ is asymptotically normal distributed. To prove this, denote the derivative of the partial log-likelihood with respect to $\boldsymbol{\beta}$ as

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^1 \boldsymbol{x}_i(s) dN_i(s) - \int_0^1 \frac{\sum_{i=1}^n Y_i(s)\boldsymbol{x}_i(s) \exp\{\boldsymbol{\beta}^T \boldsymbol{x}_i(s)\}}{\sum_{i=1}^n Y_i(s) \exp\{\boldsymbol{\beta}^T \boldsymbol{x}_i(s)\}} d\bar{N}(s).$$

And the second order derivative as

$$S(\boldsymbol{\beta}) = -\int_0^1 \Bigg(\frac{\sum_{i=1}^n Y_i(s)\boldsymbol{x}_i(s)\boldsymbol{x}_i(s)^T \exp\{\boldsymbol{\beta}^T \boldsymbol{x}_i(s)\}}{\sum_{i=1}^n Y_i(s) \exp\{\boldsymbol{\beta}^T \boldsymbol{x}_i(s)\}} - \left(\frac{\sum_{i=1}^n Y_i(s)\boldsymbol{x}_i(s)\boldsymbol{x}_i(s)^T \exp\{\boldsymbol{\beta}^T \boldsymbol{x}_i(s)\}}{\sum_{i=1}^n Y_i(s) \exp\{\boldsymbol{\beta}^T \boldsymbol{x}_i(s)\}} \right)^2 \Bigg) d\bar{N}(s).$$

Using taylor expansion, we have

$$U(\hat{\boldsymbol{\beta}}) - U(\boldsymbol{\beta}_0) = S(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

where $\boldsymbol{\beta}^*$ is on the line segment between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, and $-S(\boldsymbol{\beta})$ is a positive semidefinite matrix. By Theorem 3.2 in [63], we have

$$\frac{1}{\sqrt{n}} U_{J_0}(\boldsymbol{\beta}_0) \xrightarrow{d} N(0, I_{J_0}(\boldsymbol{\beta}_0)), \quad -\frac{1}{n} S(\boldsymbol{\beta}^*) \xrightarrow{p} N(0, I(\boldsymbol{\beta}_0)) \text{ as } n \to \infty.$$

Since $U(\hat{\boldsymbol{\beta}}) = 0$, we have $-S(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = U(\boldsymbol{\beta}_0)$, thus by Slutsky's Theorem, we have

$$\sqrt{n} I_{J_0}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, I_{J_0}(\boldsymbol{\beta}_0)),$$

which translates into

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{J_0} - \boldsymbol{\beta}^0_{J_0}) \xrightarrow{d} N(0, I_{J_0}(\boldsymbol{\beta}_0)^{-1}).$$

□

B.4 Variable definitions in the real data examples.

Table B.1: Definitions of variables in the *pbcseq* dataset

id	case number
age	in years
sex	m/f
trt	1/2/NA for D-penicillmain, placebo, not randomised
time	number of days between registration and the earlier of death, transplantion, or study analysis in July, 1986
status	status at endpoint, 0/1/2 for censored, transplant, dead
day	number of days between enrollment and this visit date all measurements below refer to this date
albumin	serum albumin (mg/dl)
alk.phos	alkaline phosphotase (U/liter)
ascites	presence of ascites
ast	aspartate aminotransferase, once called SGOT (U/ml)
bili	serum bilirunbin (mg/dl)
chol	serum cholesterol (mg/dl)
copper	urine copper (ug/day)
edema	0 no edema, 0.5 untreated or successfully treated 1 edema despite diuretic therapy
hepato	presence of hepatomegaly or enlarged liver
platelet	platelet count
protime	standardised blood clotting time
spiders	blood vessel malformations in the skin
stage	histologic stage of disease (needs biopsy)
trig	triglycerides (mg/dl)

REFERENCES

[1] Y. Wang, A. Chakrabarti, D. Sivakoff, and S. Parthasarathy, “Fast change point detection on dynamic social networks,” *arXiv preprint arXiv:1705.07325*, 2017.

[2] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park, “Scan statistics on enron graphs,” *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 229–247, 2005.

[3] J. Sharpnack, A. Rinaldo, and A. Singh, “Detecting anomalous activity on networks with the graph fourier scan statistic,” *IEEE Transactions on Signal Processing*, vol. 64, no. 2, pp. 364–379, 2016.

[4] J. D. Lee, D. L. Sun, Y. Sun, J. E. Taylor, *et al.*, “Exact post-selection inference, with application to the lasso,” *The Annals of Statistics*, vol. 44, no. 3, pp. 907–927, 2016.

[5] S.-M. Wu, K. M. Ward, J. Farrell, F.-C. Lin, M. Karplus, and R. B. Smith, “Anatomy of old faithful from subsurface seismic imaging of the yellowstone upper geyser basin,” *Geophysical Research Letters*, vol. 44, no. 20, 2017.

[6] N. Verzelen, E. Arias-Castro, *et al.*, “Community detection in sparse random networks,” *The Annals of Applied Probability*, vol. 25, no. 6, pp. 3465–3510, 2015.

[7] D. Marangoni-Simonsen and Y. Xie, “Sequential changepoint approach for online community detection.,” *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1035–1039, 2015.

[8] H. Chen, N. Zhang, *et al.*, “Graph-based change-point detection,” *The Annals of Statistics*, vol. 43, no. 1, pp. 139–176, 2015.

[9] H. Chen, “Sequential change-point detection based on nearest neighbors,” *arXiv preprint arXiv:1604.03611*, 2016.

[10] Z. I. Botev, M. Mandjes, and A. Ridder, “Tail distribution of the maximum of correlated gaussian random variables,” in *Proceedings of the 2015 Winter Simulation Conference*, IEEE Press, 2015, pp. 633–642.

[11] M. F. Schilling, “Multivariate two-sample tests based on nearest neighbors,” *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 799–806, 1986.

[12] N. Henze, “A multivariate two-sample test based on the number of nearest neighbor type coincidences,” *The Annals of Statistics*, pp. 772–783, 1988.

[13] N. Shervashidze, P. Schweitzer, E. J. v. Leeuwen, K. Mehlhorn, and K. M. Borgwardt, “Weisfeiler-lehman graph kernels,” *Journal of Machine Learning Research*, vol. 12, no. Sep, pp. 2539–2561, 2011.

[14] CDC, *Transplant safety overview: Key facts*, `https://www.cdc.gov/transplantsafety/overview/key-facts.html`, Accessed: 2019-02-01, 2019.

[15] E. L. Kaplan and P. Meier, “Nonparametric estimation from incomplete observations,” *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[16] E. Mark, D. Goldsman, B. Gurbaxani, P. Keskinocak, and J. Sokol, “Using machine learning and an ensemble of methods to predict kidney transplant survival,” *PloS one*, vol. 14, no. 1, e0209068, 2019.

[17] W. E. Harmon, R. A. McDonald, J. D. Reyes, N. D. Bridges, S. C. Sweet, C. M. Sommers, and M. K. Guidinger, “Pediatric transplantation, 1994–2003,” *American journal of transplantation*, vol. 5, no. 4p2, pp. 887–903, 2005.

[18] J. J. Kim and S. D. Marks, “Long-term outcomes of children after solid organ transplantation,” *Clinics*, vol. 69, pp. 28–38, 2014.

[19] J. Magee, S. Krishnan, M. Benfield, D. Hsu, and B. Shneider, “Pediatric transplantation in the united states, 1997–2006,” *American Journal of Transplantation*, vol. 8, no. 4p2, pp. 935–945, 2008.

[20] J. C. Magee, J. C. Bucuvalas, D. G. Farmer, W. E. Harmon, T. E. Hulbert-Shearon, and E. N. Mendeloff, “Pediatric transplantation,” *American Journal of Transplantation*, vol. 4, no. s9, pp. 54–71, 2004.

[21] L. Rees, “Long-term outcome after renal transplantation in childhood,” *Pediatric Nephrology*, vol. 24, no. 3, pp. 475–484, 2009.

[22] R. Shapiro and M. M. Sarwal, “Pediatric kidney transplantation,” *Pediatric Clinics*, vol. 57, no. 2, pp. 393–400, 2010.

[23] K. J. Van Arendonk, B. J. Boyarsky, B. J. Orandi, N. T. James, J. M. Smith, P. M. Colombani, and D. L. Segev, “National trends over 25 years in pediatric kidney transplant outcomes,” *Pediatrics*, vol. 133, no. 4, pp. 594–601, 2014.

[24] S. M. Wrenn, P. W. Callas, T. Kapoor, A. F. Aunchman, A. N. Paine, J. A. Pineda, and C. E. Marroquin, “Increased risk organ transplantation in the pediatric population,” *Pediatric transplantation*, vol. 21, no. 8, 2017.

[25] E. N. Ellis, K. Martz, L. Talley, M. Ilyas, K. L. Pennington, and R. T. Blaszak, “Factors related to long-term renal transplant function in children,” *Pediatric Nephrology*, vol. 23, no. 7, pp. 1149–1155, 2008.

[26] D. W. Gjertson and J. M. Cecka, “Determinants of long-term survival of pediatric kidney grafts reported to the united network for organ sharing kidney transplant registry,” *Pediatric transplantation*, vol. 5, no. 1, pp. 5–15, 2001.

[27] A. H. Hwang, Y. W. Cho, J. Cicciarelli, M. Mentser, Y. Iwaki, and B. E. Hardy, “Risk factors for short-and long-term survival of primary cadaveric renal allografts in pediatric recipients: A unos analysis,” *Transplantation*, vol. 80, no. 4, pp. 466–470, 2005.

[28] P. Rianthavorn, S. J. Kerr, A. Lumpaopong, A. Jiravuttipong, A. Pattaragarn, K. Tangnararatchakit, Y. Avihingsanon, P. Thirakupt, and V. Sumethkul, “Outcomes and predictive factors of pediatric kidney transplants: An analysis of the thai transplant registry,” *Pediatric transplantation*, vol. 17, no. 2, pp. 112–118, 2013.

[29] A. Vats, K. Gillingham, A. Matas, and B. Chavers, “Improved late graft survival and half-lives in pediatric kidney transplantation: A single center experience,” *American Journal of Transplantation*, vol. 2, no. 10, pp. 939–945, 2002.

[30] W. E. Harmon, S. R. Alexander, A. Tejani, and D. Stablein, “The effect of donor age on graft survival in pediatric cadaver renal transplant recipients–a report of the north american pediatric renal transplant cooperative study.,” *Transplantation*, vol. 54, no. 2, pp. 232–237, 1992.

[31] H. Gritsch, J. Veale, A. Leichtman, M. Guidinger, J. Magee, R. McDonald, W. Harmon, F. Delmonico, R. Ettenger, and J. Cecka, “Should pediatric patients wait for hla-dr-matched renal transplants?” *American Journal of Transplantation*, vol. 8, no. 10, pp. 2056–2061, 2008.

[32] J. M. Smith, R. A. McDonald, L. S. Finn, P. J. Healey, C. L. Davis, and A. P. Limaye, “Polyomavirus nephropathy in pediatric kidney transplant recipients,” *American Journal of Transplantation*, vol. 4, no. 12, pp. 2109–2117, 2004.

[33] J. D. Terrace and G. C. Oniscu, “Paediatric obesity and renal transplantation: Current challenges and solutions,” *Pediatric Nephrology*, vol. 31, no. 4, pp. 555–562, 2016.

[34] D. Cox, “Regression models and li-tables (with discussion),” *J Royal Stat Soc, Series B*, vol. 34, pp. 187–220, 1972.

[35] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, “Random survival forests,” *The annals of applied statistics*, pp. 841–860, 2008.

[36] P. I. Terasaki, J. M. Cecka, D. W. Gjertson, and S. Takemoto, "High survival rates of kidney transplants from spousal and living unrelated donors," *New England Journal of Medicine*, vol. 333, no. 6, pp. 333–336, 1995.

[37] N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration," *Cancer Chemother Rep*, vol. 50, pp. 163–170, 1966.

[38] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[39] W. Guo, J. Jin, K. Paynabar, B. Miller, and J. Carpenter, "A decision support system on surgical treatments for rotator cuff tears," *IIE Transactions on Healthcare Systems Engineering*, vol. 5, no. 3, pp. 197–210, 2015.

[40] S. v. Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, pp. 1–68, 2010.

[41] R. Tibshirani, "The lasso method for variable selection in the cox model," *Statistics in medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[42] F. E. Harrell Jr, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in medicine*, vol. 15, no. 4, pp. 361–387, 1996.

[43] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.

[44] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[45] M. Laimighofer, J. Krumsiek, F. Buettner, and F. J. Theis, "Unbiased prediction and feature selection in high-dimensional survival regression," *Journal of Computational Biology*, vol. 23, no. 4, pp. 279–290, 2016.

[46] OPTN, *A guide to calculating and interpreting the estimated post-transplant survival (epts) score used in the kidney allocation system (kas)*, `https://optn.transplant.hrsa.gov/media/1511/guide_to_calculating_interpreting_epts.pdf`, Accessed: 2019-12-09, 2019.

[47] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[48] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[49] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[50] Y. Wu, "Elastic net for cox's proportional hazards model with a solution path algorithm," *Statistica Sinica*, vol. 22, p. 27, 2012.

[51] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[52] J. Fan, R. Li, *et al.*, "Variable selection for cox's proportional hazards model and frailty model," *The Annals of Statistics*, vol. 30, no. 1, pp. 74–99, 2002.

[53] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[54] H. H. Zhang and W. Lu, "Adaptive lasso for cox's proportional hazards model," *Biometrika*, vol. 94, no. 3, pp. 691–703, 2007.

[55] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.

[56] N. Chaturvedi, R. X. de Menezes, and J. J. Goeman, "Fused lasso algorithm for cox proportional hazards and binomial logit models with application to copy number profiles," *Biometrical Journal*, vol. 56, no. 3, pp. 477–492, 2014.

[57] C. H. Zhang, "Penalized linear unbiased selection," *Department of Statistics and Bioinformatics, Rutgers University*, vol. 3, 2007.

[58] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[59] G. Yu and Y. Liu, "Sparse regression incorporating graphical structure among predictors," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 707–720, 2016.

[60] Y. Li, B. Mark, G. Raskutti, and R. Willett, "Graph-based regularization for regression problems with highly-correlated designs," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, IEEE, 2018, pp. 740–742.

[61] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.

[62] G. Obozinski, L. Jacob, and J.-P. Vert, “Group lasso with overlaps: The latent group lasso approach,” *arXiv preprint arXiv:1110.0413*, 2011.

[63] P. K. Andersen and R. D. Gill, “Cox’s regression model for counting processes: A large sample study,” *The annals of statistics*, pp. 1100–1120, 1982.

[64] S. A. Murphy and A. W. Van der Vaart, “On profile likelihood,” *Journal of the American Statistical Association*, vol. 95, no. 450, pp. 449–465, 2000.

[65] J. Huang and C. X. Ling, “Using auc and accuracy in evaluating learning algorithms,” *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[66] M. Maechler, *Matrix*, 2019.

[67] J. Kropko and J. J. Harden, *Coxed: Duration-based quantities of interest for the cox proportional hazards model*, 2020.

[68] P. A. Murtaugh, E. R. Dickson, G. M. Van Dam, M. Malinchoc, P. M. Grambsch, A. L. Langworthy, and C. H. Gips, “Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits,” *Hepatology*, vol. 20, no. 1, pp. 126–134, 1994.

[69] T. R. Fleming and D. P. Harrington, *Counting processes and survival analysis*. John Wiley & Sons, 2011, vol. 169.

[70] T. M. Therneau, T. Lumley, A. Elizabeth, and C. Cynthia, *Survival: Survival analysis*, 2020.

[71] S. Kim, “Ppcor: An r package for a fast calculation to semi-partial correlation coefficients,” *Communications for statistical applications and methods*, vol. 22, no. 6, p. 665, 2015.

[72] *What are the common blood tests available to diagnose hepatitis b?* `https://www.cdc.gov/hepatitis/hbv/bfaq.htm`, Accessed: 2019-02-01.

[73] *Reference for interpretation of hepatitis c virus (hcv) test results*, `https://www.cdc.gov/hepatitis/Resources/OrderPubs/HealthProf/Ref-IntHCVTestResults_Eng.pdf`, Accessed: 2019-02-01.

[74] *Reasons for kidney transplants*, `https://optn.transplant.hrsa.gov/data/organ-datasource/kidney/`, Accessed: 2019-02-01.